Preproducibility: What we may not, with advantage, omit

UCLA California Center for Population Research

Philip B. Stark

Department of Statistics, University of California, Berkeley

19 April 2023

Computational Reproducibility

Starting with the same data, can you produce the same tables, figures, and quantitative conclusions?

Experimental Replicability

Does repeating “the same” experiment and analyzing the resulting data the same way give “substantially the same” result?

Experimental Replicability

Does repeating “the same” experiment and analyzing the resulting data the same way give “substantially the same” result?

Variations: same lab, same reagents? different lab, different reagents?

Frontispiece of R.A. Fisher's (1935) *The Design of Experiments*:

I AM very sorry, Pyrophilus, that to the many (elsewhere enumerated) difficulties which you may meet with, and must therefore surmount, in the serious and effectual prosecution of experimental philosophy I must add one discouragement more, which will perhaps as much surprise as dishearten you; and it is, that besides that you will find (as we elsewhere mention) many of the experiments published by authors, or related to you by the persons you converse with, false and unsuccessful (besides this, I say), you will meet with several observations and experiments which, though communicated for true by candid authors or undistrusted eye-witnesses, or perhaps recommended by your own experience may, upon further trial, disappoint your expectation, either not at all succeeding constantly or at least varying much from what you expected.

–Robert Boyle, 1673, Concerning the Unsuccessfulness of Experiments.

Fisher on experimental "proof"

> ... [N]o isolated experiment, however significant in itself, can suffice for the experimental demonstration of any natural phenomenon; for the "one chance in a million" will undoubtedly occur, with no less and no more than, its appropriate frequency, however surprised we may be that it should occur to *us*. **In order to assert that a natural phenomenon is experimentally demonstrable we need, not an isolated record, but a reliable method of procedure.** In relation to the test of significance, we may say that a phenomenon is experimentally demonstrable when we know how to conduct an experiment which will rarely fail to give us a statistically significant result.

–Fisher, 1935, *The Design of Experiments*

WORLD VIEW *A personal take on events*

NAOMI A. P. STARK

No reproducibility without preproducibility

Instead of arguing about whether results hold up, let's push to provide enough information for others to repeat the experiments, says **Philip Stark**.

From time to time over the past few years, I've politely refused requests to referee an article on the grounds that it lacks enough information for me to check the work. This can be a hard thing to explain.

Our lack of a precise vocabulary — in particular the fact that we don't have a word for 'you didn't tell me what you did in sufficient detail for me to check it' — contributes to the crisis of scientific reproducibility. In computational science, 'reproducible' often means that enough information is provided to allow a dedicated reader to repeat the calculations in the paper for herself. In biomedical disciplines, 'reproducible' often means that a different lab, starting the experiment from scratch, would get roughly the same experimental result.

In 1992, philosopher Karl Popper wrote: "Science may be described as the art of systematic oversimplification — the art of discerning what we may with advantage omit." What may be omitted depends on the discipline. Results that generalize to all universes (or perhaps do not even require a universe) are part of mathematics. Results that generalize to our Universe belong to physics. Results that generalize to all life on Earth underpin molecular biology. Results that generalize to all mice are murine biology. And results that hold only for a particular mouse in a particular lab in a particular experiment are arguably not science.

Communicating a scientific result requires enumerating, recording and reporting those

SCIENCE SHOULD BE 'SHOW ME', NOT 'TRUST ME'.

or analysis is preproducible if it has been described in adequate detail for others to undertake it. Preproducibility is a prerequisite for reproducibility, and the idea makes sense across disciplines.

The distinction between a preproducible scientific report and current common practice is like the difference between a partial list of ingredients and a recipe. To bake a good loaf of bread, it isn't enough to know that it contains flour. It isn't even enough to know that it contains flour, water, salt and yeast. The brand of flour might be omitted from the recipe with advantage, as might the day of the week on which the loaf was baked. But the ratio of ingredients, the operations, their timing and the temperature of the oven cannot.

Given preproducibility — a 'scientific recipe' — we can attempt to make a similar loaf of scientific bread. If we follow the recipe but do not get the same result, either the result is sensitive to small details that cannot be controlled, the result is incorrect or the recipe was not precise enough (things were omitted to disadvantage).

Depending on the discipline, preproducibility might require information about materials (including organisms and their care), instruments and procedures; experimental design; raw data at the instrument level; algorithms used to process the raw data; computational tools used in analyses, including any parameter settings or ad hoc choices; code, processed data and software build environments; or analyses that were tried and abandoned.

Peer review is hamstrung by lack of pre-

Preproducibility

An experiment or analysis is ***preproducible*** *if it has been described in adequate detail for others to undertake it.*

Preproducibility

> *An experiment or analysis is* ***preproducible*** *if it has been described in adequate detail for others to undertake it.*

In a nutshell: Show Your Work!

Preproducibility

> *An experiment or analysis is **preproducible** if it has been described in adequate detail for others to undertake it.*

In a nutshell: Show Your Work!

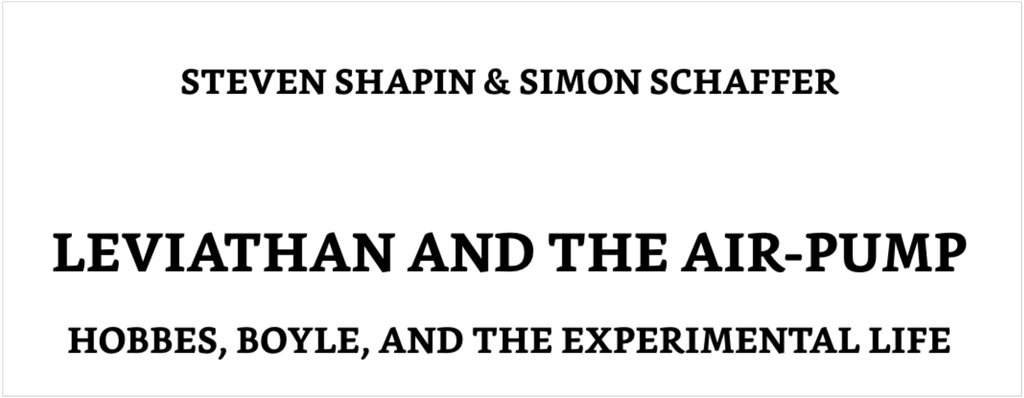
Provide *evidence* that your claims are correct, and a way to check them

What is the purpose of scientific publishing?

- Establish priority / get credit?
- Communicate claims?

What is the purpose of scientific publishing?

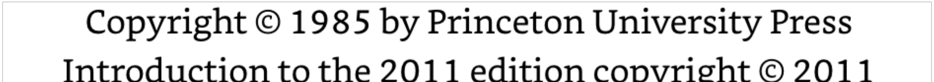
- Establish priority / get credit?
- Communicate claims?
- Provide evidence that claims are correct?
- Provide enough information that others can re-undertake and verify?
- Provide methods to others, to contribute to science as a societal undertaking?



STEVEN SHAPIN & SIMON SCHAFFER

LEVIATHAN AND THE AIR-PUMP

HOBBES, BOYLE, AND THE EXPERIMENTAL LIFE



ological Essays of 1661 were written to another nephew, Richard Jones; the *History of Colours* of 1664 was originally written to an unspecified friend.[74] The purpose of this form of communication was explicitly to proselytize. The *New Experiments* was published so "that the person I addressed them to might, without mistake, and with as little trouble as possible, be able to repeat such unusual experiments. . . ."[75] The *History of Colours* was designed "not barely to relate [the experiments], but . . . to teach a young gentleman to make them."[76] Boyle wished to encourage young gentlemen to "addict" themselves to experimental pursuits and thereby to multiply both experimental philosophers and experimental facts.

In Boyle's view, replication was rarely accomplished. When he came to publish the *Continuation of New Experiments* more than eight years after the original air-pump trials, Boyle admitted that, despite his care in communicating details of the engine and his procedures, there had been few successful replications.[77] This situation had not materially changed by the mid-1670s. In the seven or eight years after the *Continuation*, Boyle said that he had heard "of very few experiments made, either in the engine

I used, or in any other made after the model thereof." Boyle now expressed despair that these experiments would ever be replicated. He said that he was now even more willing "to set down divers things with their minute circumstances" because "probably many of these experiments would be never either re-examined by others, or reiterated by myself." Anyone who set about trying to replicate such experiments, Boyle said, "will find it no easy task."[78]

Prolixity and Iconography

The third way by which witnesses could be multiplied is far more important than the performance of experiments before direct witnesses or the facilitating of their replication: it is what we shall call *virtual witnessing*. The technology of virtual witnessing involves the production in a *reader's* mind of such an image of an experimental scene as obviates the necessity for either direct witness or replication.[79] Through virtual witnessing the multiplication of witnesses could be, in principle, unlimited. It was therefore the most powerful technology for constituting intellectual collective had mutually to assure themselves and others that belief in an empirical experience was warranted. Matters of fact were the outcome of the process of having an empirical experience, warranting it to oneself, and assuring others that grounds for their belief were adequate. In that process a multiplication of the witnessing experience was fundamental. An experience, even of a rigidly controlled experimental performance, that one man alone witnessed was not adequate to make a matter of fact. If that experience could be extended to many, and in principle to all men, then the result could be constituted as a matter of fact. In this way, the matter of fact is to be seen as both an epistemological and a social category. The foundational item of experimental knowledge, and of what counted as properly grounded knowledge generally, was an artifact of communication and whatever social forms were deemed necessary to sustain and enhance communication.

of trust and assurance that the things had been done and done in the way claimed.

The technology of virtual witnessing was not different in kind to that used to facilitate actual replication. One could deploy the same linguistic resources in order to encourage the physical replication of experiments or to trigger in the reader's mind a naturalistic image of the experimental scene. Of course, actual replication was to be preferred, for this eliminated reliance upon testimony altogether. Yet, because of natural and legitimate suspicion among those who were neither direct witnesses nor replicators, a greater degree of assurance was required to produce assent in virtual witnesses. Boyle's literary technology was crafted to secure this assent.

Buckheit and Donoho, 1995

An article about computational result is advertising, not scholarship. The actual scholarship is the full software environment, code and data, that produced the result.

By working preproducibly, you . . .

- allow others “without mistake, and with as little trouble as possible, to be able to repeat such unusual experiments”
- make “multiplication” and “virtual witnessing” possible
- provide evidence that your claim is a fact

If you do not work preproducibly, you . . .

- merely advertise the result
- ask others to take the result on faith
- withhold crucial evidence needed to check, repeat, or use your work
- make actual replication/reproduction even less likely

Science should be *show me*, not *trust me*.

Nullius in verba

Many concepts, many labels, used inconsistently

- replicable
- reproducible
- repeatable
- confirmable
- stable
- generalizable
- reviewable
- auditable
- verifiable
- validatable

Many concepts, many labels, used inconsistently

- replicable
- reproducible
- repeatable
- confirmable
- stable
- generalizable
- reviewable
- auditable
- verifiable
- validatable

Generally about whether something happens again.

Many concepts, many labels, used inconsistently

- replicable
- reproducible
- repeatable
- confirmable
- stable
- generalizable
- reviewable
- auditable
- verifiable
- validatable

Generally about whether something happens again.

No term for “not enough information to try.”

Preproducibility versus Reproducibility and Replicability

- A failure of *preproducibility* is often a failure of scientific *communication*.
- A failure of *reproducibility* or *replicability* could be a false discovery, a failure of practice, or a sign of something scientifically interesting

Some *ceteris* assumed *paribus* . . . approximately.

- Similar result if experiment is repeated in same lab?
- Similar result if procedure repeated elsewhere, by others?
- Similar result under similar circumstances?
- Same numbers/graphs if data analysis is repeated by others?

Some *ceteris* assumed *paribus* . . . approximately.

- Similar result if experiment is repeated in same lab?
- Similar result if procedure repeated elsewhere, by others?
- Similar result under similar circumstances?
- Same numbers/graphs if data analysis is repeated by others?

With respect to what changes is the result stable?
Changes of what size?
How stable?

What *ceteris* need not be *paribus*?

Science may be described as the art of systematic over-simplification—the art of discerning what we may with advantage omit. —Karl Popper

What *ceteris* need not be *paribus*?

> *Science may be described as the art of systematic over-simplification—the art of discerning what we may with advantage omit. —Karl Popper*

Preproducibility means identifying, specifying, recording, and communicating those things that we may *not* with advantage omit.

Level of generalization *defines* scientific disciplines**

- If you want to generalize to all time and all universes: math
- If you want to generalize to our universe: physics
- If you want to generalize to all life on Earth: molecular and cell biology
- If you want to generalize to all fish: ichthyology
- If you want to generalize to TL strain of *Danio rerio*: I don't know
- This animal in this lab in this experiment today: maybe not science?

Level of generalization *defines* scientific disciplines**

- If you want to generalize to all time and all universes: math
- If you want to generalize to our universe: physics
- If you want to generalize to all life on Earth: molecular and cell biology
- If you want to generalize to all fish: ichthyology
- If you want to generalize to TL strain of *Danio rerio*: I don't know
- This animal in this lab in this experiment today: maybe not science?

Tolerable variation in conditions depends on the desired inference.

Level of generalization *defines* scientific disciplines**

- If you want to generalize to all time and all universes: math
- If you want to generalize to our universe: physics
- If you want to generalize to all life on Earth: molecular and cell biology
- If you want to generalize to all fish: ichthyology
- If you want to generalize to TL strain of *Danio rerio*: I don't know
- This animal in this lab in this experiment today: maybe not science?

Tolerable variation in conditions depends on the desired inference.

If variations in conditions that are irrelevant to the discipline cause the results to vary, there's a replicability problem: the *outcome* doesn't have the right level of abstraction.

Level of generalization *defines* scientific disciplines**

- If you want to generalize to all time and all universes: math
- If you want to generalize to our universe: physics
- If you want to generalize to all life on Earth: molecular and cell biology
- If you want to generalize to all fish: ichthyology
- If you want to generalize to TL strain of *Danio rerio*: I don't know
- This animal in this lab in this experiment today: maybe not science?

Tolerable variation in conditions depends on the desired inference.

If variations in conditions that are irrelevant to the discipline cause the results to vary, there's a replicability problem: the *outcome* doesn't have the right level of abstraction.

** "All science is either physics or stamp collecting." —Lord Rutherford

JBS Haldane, 1926. “On Being the Right Size,” *Harper’s Magazine*

> *You can drop a mouse down a thousand-yard mine shaft; and, on arriving at the bottom, it gets a slight shock and walks away, provided that the ground is fairly soft. A rat is killed, a man is broken, a horse splashes. For the resistance presented to movement by the air is proportional to the surface of the moving object. . . .*

Abstraction and Replicability

- If something only happens under *exactly* the same circumstances, unlikely to be useful.
- What factors may we, with advantage, omit?
- If attempt to replicate/reproduce fails, *why* did it fail? (cf Newton)
    - effect is intrinsically variable or intermittent
    - result is a statistical fluke or “false discovery”
    - something that mattered was different

Abstraction and Replicability

- If something only happens under *exactly* the same circumstances, unlikely to be useful.
- What factors may we, with advantage, omit?
- If attempt to replicate/reproduce fails, *why* did it fail? (cf Newton)
  - effect is intrinsically variable or intermittent
  - result is a statistical fluke or "false discovery"
  - something that mattered was different

If the necessary qualification is too restrictive, the result might change disciplines.

Questions

- materials, instruments, procedures, & conditions specified adequately to allow repeating data collection?

Questions

- materials, instruments, procedures, & conditions specified adequately to allow repeating data collection?
- data analysis described adequately to check/repeat?

Questions

- materials, instruments, procedures, & conditions specified adequately to allow repeating data collection?
- data analysis described adequately to check/repeat?
- code & data available to re-generate figures and tables?

Questions

- materials, instruments, procedures, & conditions specified adequately to allow repeating data collection?
- data analysis described adequately to check/repeat?
- code & data available to re-generate figures and tables?
- code readable and checkable?

Questions

- materials, instruments, procedures, & conditions specified adequately to allow repeating data collection?
- data analysis described adequately to check/repeat?
- code & data available to re-generate figures and tables?
- code readable and checkable?
- software versions and build environment specified adequately?

Questions

- materials, instruments, procedures, & conditions specified adequately to allow repeating data collection?
- data analysis described adequately to check/repeat?
- code & data available to re-generate figures and tables?
- code readable and checkable?
- software versions and build environment specified adequately?
- what is the evidence that the result is correct?

Questions

- materials, instruments, procedures, & conditions specified adequately to allow repeating data collection?
- data analysis described adequately to check/repeat?
- code & data available to re-generate figures and tables?
- code readable and checkable?
- software versions and build environment specified adequately?
- what is the evidence that the result is correct?
- how generally do the results hold? how stable are the results to perturbations of the experiment?

Questions, questions

- What's the underlying experiment?

Questions, questions

- What’s the underlying experiment?
- What are the raw data? How were they collected/selected?

Questions, questions

- What's the underlying experiment?
- What are the raw data? How were they collected/selected?
- How were raw data processed to get "data"?

Questions, questions

- What's the underlying experiment?
- What are the raw data? How were they collected/selected?
- How were raw data processed to get "data"?
- How were processed data analyzed?

Questions, questions

- What's the underlying experiment?
- What are the raw data? How were they collected/selected?
- How were raw data processed to get "data"?
- How were processed data analyzed?
- Was that the right analysis?

Questions, questions

- What's the underlying experiment?
- What are the raw data? How were they collected/selected?
- How were raw data processed to get "data"?
- How were processed data analyzed?
- Was that the right analysis?
- Was it done correctly?

Questions, questions

- What's the underlying experiment?
- What are the raw data? How were they collected/selected?
- How were raw data processed to get "data"?
- How were processed data analyzed?
- Was that the right analysis?
- Was it done correctly?
- Were the results reported correctly?

Questions, questions

- What's the underlying experiment?
- What are the raw data? How were they collected/selected?
- How were raw data processed to get "data"?
- How were processed data analyzed?
- Was that the right analysis?
- Was it done correctly?
- Were the results reported correctly?
- Were there *ad hoc* choices? Did they matter?

Questions, questions

- What's the underlying experiment?
- What are the raw data? How were they collected/selected?
- How were raw data processed to get "data"?
- How were processed data analyzed?
- Was that the right analysis?
- Was it done correctly?
- Were the results reported correctly?
- Were there *ad hoc* choices? Did they matter?
- What other analyses were tried? How was multiplicity treated?

Questions, questions

- What's the underlying experiment?
- What are the raw data? How were they collected/selected?
- How were raw data processed to get "data"?
- How were processed data analyzed?
- Was that the right analysis?
- Was it done correctly?
- Were the results reported correctly?
- Were there *ad hoc* choices? Did they matter?
- What other analyses were tried? How was multiplicity treated?
- Can someone else use the procedures and tools?

Variation: wanted and unwanted

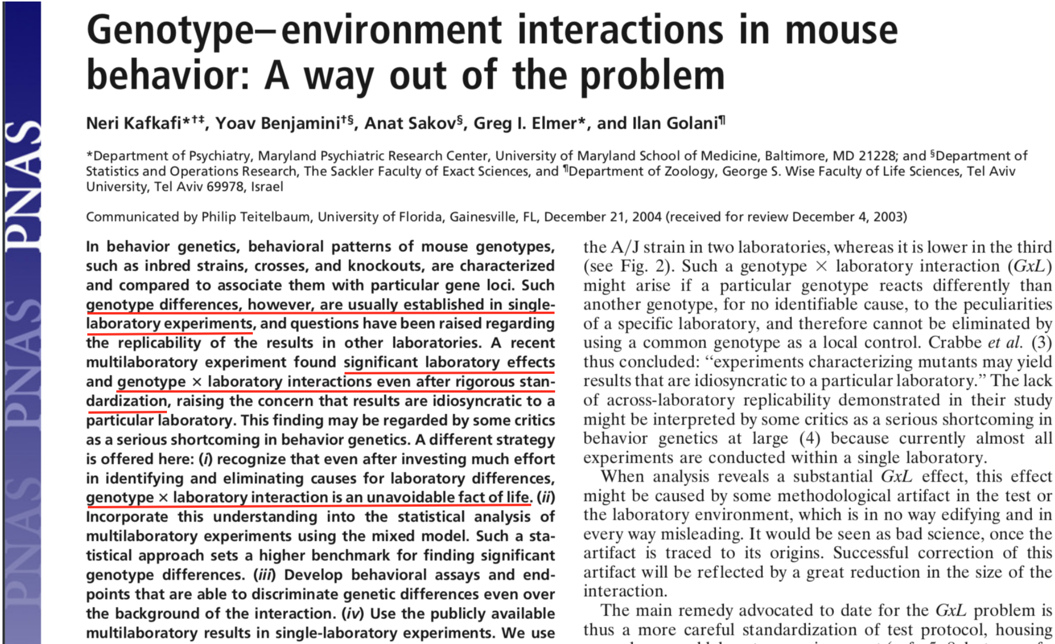
- study population, survey wording, genotype, biology, lab, procedures, handlers, reagents, feed/diet, water circulation, water quality, temperature, pH, conductivity, noise, visual background, size of cross-breeding cohorts, subclinical infections . . .

Variation: wanted and unwanted

- study population, survey wording, genotype, biology, lab, procedures, handlers, reagents, feed/diet, water circulation, water quality, temperature, pH, conductivity, noise, visual background, size of cross-breeding cohorts, subclinical infections . . .
- Want results stable wrt *some* kinds of variability

Variation: wanted and unwanted

- study population, survey wording, genotype, biology, lab, procedures, handlers, reagents, feed/diet, water circulation, water quality, temperature, pH, conductivity, noise, visual background, size of cross-breeding cohorts, subclinical infections ...
- Want results stable wrt *some* kinds of variability
- OTOH, variability *itself* can be scientifically interesting



Genotype–environment interactions in mouse behavior: A way out of the problem

Neri Kafkafi*†‡, Yoav Benjamini†§, Anat Sakov§, Greg I. Elmer*, and Ilan Golani¶

*Department of Psychiatry, Maryland Psychiatric Research Center, University of Maryland School of Medicine, Baltimore, MD 21228; and §Department of Statistics and Operations Research, The Sackler Faculty of Exact Sciences, and ¶Department of Zoology, George S. Wise Faculty of Life Sciences, Tel Aviv University, Tel Aviv 69978, Israel

Communicated by Philip Teitelbaum, University of Florida, Gainesville, FL, December 21, 2004 (received for review December 4, 2003)

In behavior genetics, behavioral patterns of mouse genotypes, such as inbred strains, crosses, and knockouts, are characterized and compared to associate them with particular gene loci. Such genotype differences, however, are usually established in single-laboratory experiments, and questions have been raised regarding the replicability of the results in other laboratories. A recent multilaboratory experiment found significant laboratory effects and genotype × laboratory interactions even after rigorous standardization, raising the concern that results are idiosyncratic to a particular laboratory. This finding may be regarded by some critics as a serious shortcoming in behavior genetics. A different strategy is offered here: (*i*) recognize that even after investing much effort in identifying and eliminating causes for laboratory differences, genotype × laboratory interaction is an unavoidable fact of life. (*ii*) Incorporate this understanding into the statistical analysis of multilaboratory experiments using the mixed model. Such a statistical approach sets a higher benchmark for finding significant genotype differences. (*iii*) Develop behavioral assays and endpoints that are able to discriminate genetic differences even over the background of the interaction. (*iv*) Use the publicly available multilaboratory results in single-laboratory experiments. We use

the A/J strain in two laboratories, whereas it is lower in the third (see Fig. 2). Such a genotype × laboratory interaction (*GxL*) might arise if a particular genotype reacts differently than another genotype, for no identifiable cause, to the peculiarities of a specific laboratory, and therefore cannot be eliminated by using a common genotype as a local control. Crabbe *et al.* (3) thus concluded: "experiments characterizing mutants may yield results that are idiosyncratic to a particular laboratory." The lack of across-laboratory replicability demonstrated in their study might be interpreted by some critics as a serious shortcoming in behavior genetics at large (4) because currently almost all experiments are conducted within a single laboratory.

When analysis reveals a substantial *GxL* effect, this effect might be caused by some methodological artifact in the test or the laboratory environment, which is in no way edifying and in every way misleading. It would be seen as bad science, once the artifact is traced to its origins. Successful correction of this artifact will be reflected by a great reduction in the size of the interaction.

The main remedy advocated to date for the *GxL* problem is thus a more careful standardization of test protocol, housing

Computational p/reproducibility

- Variation with analysis/methodology & *implementation* of tools
- Undesirable for analysis to be unstable, but algorithms matter, numerics matter, . . .
- Relying on packaged/commercial tools can be a problem
- Adopt tools from software development world:
  - revision control systems (not, eg, Dropbox or Google Docs)
  - documentation, documentation, documentation
  - coding standards/conventions
  - pair programming
  - issue trackers
  - code reviews (and in teaching, grade students' *code*, not just their *output*)
  - unit tests, regression tests, coverage checks
  - continuous integration; automation
  - scripted analyses: no point-and-click tools, *especially* spreadsheet calculations

Spreadsheets might be OK for data entry. Not for calculations

- Conflates input, code, output, presentation

Spreadsheets might be OK for data entry. Not for calculations

- Conflates input, code, output, presentation
- UI invites errors, then obscures them

Spreadsheets might be OK for data entry. Not for calculations

- Conflates input, code, output, presentation
- UI invites errors, then obscures them
- Debugging extremely hard

Spreadsheets might be OK for data entry. Not for calculations

- Conflates input, code, output, presentation
- UI invites errors, then obscures them
- Debugging extremely hard
- Unit testing hard/impossible

Spreadsheets might be OK for data entry. Not for calculations

- Conflates input, code, output, presentation
- UI invites errors, then obscures them
- Debugging extremely hard
- Unit testing hard/impossible
- Replication hard/impossible

Spreadsheets might be OK for data entry. Not for calculations

- Conflates input, code, output, presentation
- UI invites errors, then obscures them
- Debugging extremely hard
- Unit testing hard/impossible
- Replication hard/impossible
- Code review hard

Spreadsheets might be OK for data entry. Not for calculations

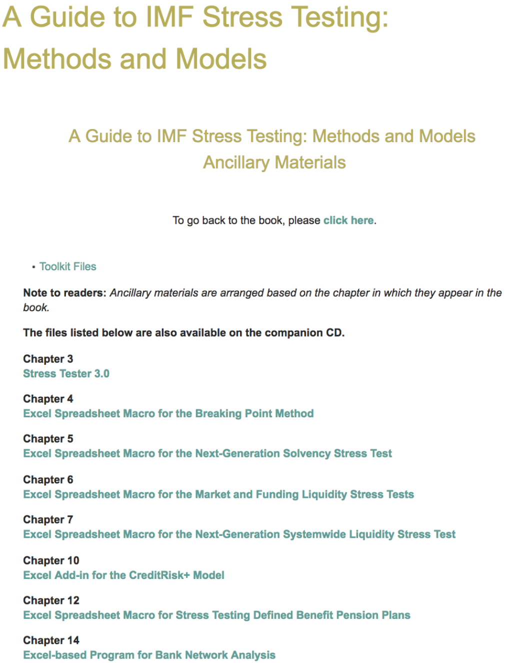
- Conflates input, code, output, presentation
- UI invites errors, then obscures them
- Debugging extremely hard
- Unit testing hard/impossible
- Replication hard/impossible
- Code review hard
- According to KPMG and PWC, over 90% of corporate spreadsheets have errors

Spreadsheets might be OK for data entry. Not for calculations

- Conflates input, code, output, presentation
- UI invites errors, then obscures them
- Debugging extremely hard
- Unit testing hard/impossible
- Replication hard/impossible
- Code review hard
- According to KPMG and PWC, over 90% of corporate spreadsheets have errors
- Bug in the PRNG for many generations of Excel, allegedly fixed in Excel 2010.

Spreadsheets might be OK for data entry. Not for calculations

- Conflates input, code, output, presentation
- UI invites errors, then obscures them
- Debugging extremely hard
- Unit testing hard/impossible
- Replication hard/impossible
- Code review hard
- According to KPMG and PWC, over 90% of corporate spreadsheets have errors
- Bug in the PRNG for many generations of Excel, allegedly fixed in Excel 2010.
- Other bugs in Excel $+$, *, statistical routines; PRNG still won't accept a seed; etc.



A Guide to IMF Stress Testing: Methods and Models

A Guide to IMF Stress Testing: Methods and Models
Ancillary Materials

To go back to the book, please **click here**.

- Toolkit Files

Note to readers: *Ancillary materials are arranged based on the chapter in which they appear in the book.*

The files listed below are also available on the companion CD.

Chapter 3
Stress Tester 3.0

Chapter 4
Excel Spreadsheet Macro for the Breaking Point Method

Chapter 5
Excel Spreadsheet Macro for the Next-Generation Solvency Stress Test

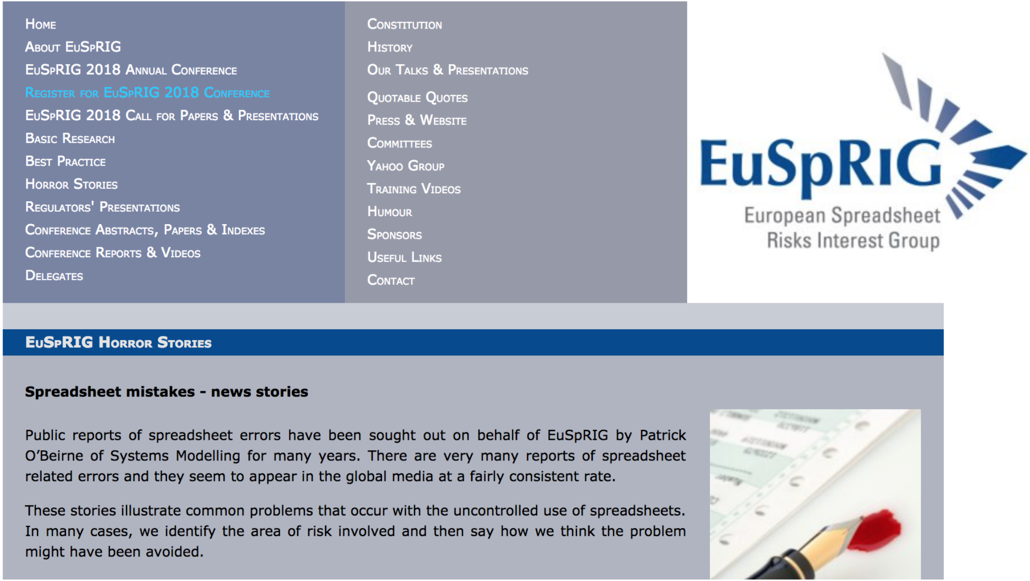
Chapter 6
Excel Spreadsheet Macro for the Market and Funding Liquidity Stress Tests

Chapter 7
Excel Spreadsheet Macro for the Next-Generation Systemwide Liquidity Stress Test

Chapter 10
Excel Add-in for the CreditRisk+ Model

Chapter 12
Excel Spreadsheet Macro for Stress Testing Defined Benefit Pension Plans

Chapter 14
Excel-based Program for Bank Network Analysis



Home
About EuSpRIG
EuSpRIG 2018 Annual Conference
Register for EuSpRIG 2018 Conference
EuSpRIG 2018 Call for Papers & Presentations
Basic Research
Best Practice
Horror Stories
Regulators' Presentations
Conference Abstracts, Papers & Indexes
Conference Reports & Videos
Delegates

Constitution
History
Our Talks & Presentations
Quotable Quotes
Press & Website
Committees
Yahoo Group
Training Videos
Humour
Sponsors
Useful Links
Contact

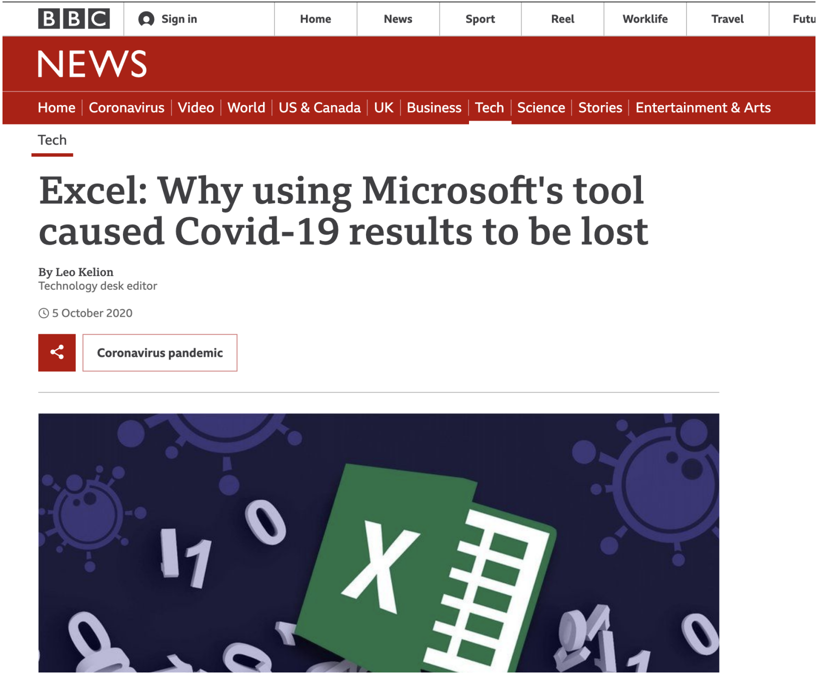
EuSpRiG
European Spreadsheet
Risks Interest Group

EuSpRIG Horror Stories

Spreadsheet mistakes - news stories

Public reports of spreadsheet errors have been sought out on behalf of EuSpRIG by Patrick O'Beirne of Systems Modelling for many years. There are very many reports of spreadsheet related errors and they seem to appear in the global media at a fairly consistent rate.

These stories illustrate common problems that occur with the uncontrolled use of spreadsheets. In many cases, we identify the area of risk involved and then say how we think the problem might have been avoided.



Excel: Why using Microsoft's tool caused Covid-19 results to be lost

By Leo Kelion
Technology desk editor

5 October 2020

Relying on spreadsheets for important calculations is like driving drunk:

No matter how carefully you do it, a wreck is likely.

2014 Coverity study

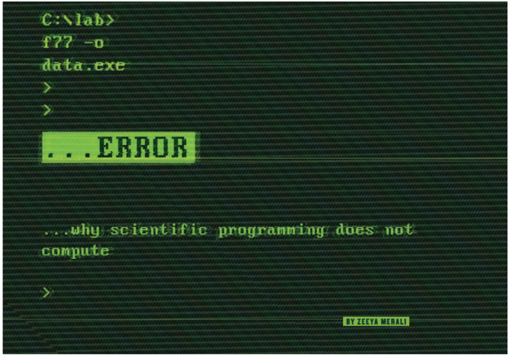
- 0.61 errors per 1,000 lines of source code in open-source projects

2014 Coverity study

- 0.61 errors per 1,000 lines of source code in open-source projects
- 0.76 errors per 1,000 lines of source code in commercial software

2014 Coverity study

- 0.61 errors per 1,000 lines of source code in open-source projects
- 0.76 errors per 1,000 lines of source code in commercial software
- Scientists generally don't use good software engineering practices, so expect worse in practice.



```
C:\lab>
f77 -o
data.exe
>
>
```

...ERROR

...why scientific programming does not compute

```
>
```

BY ZEEYA MERALI

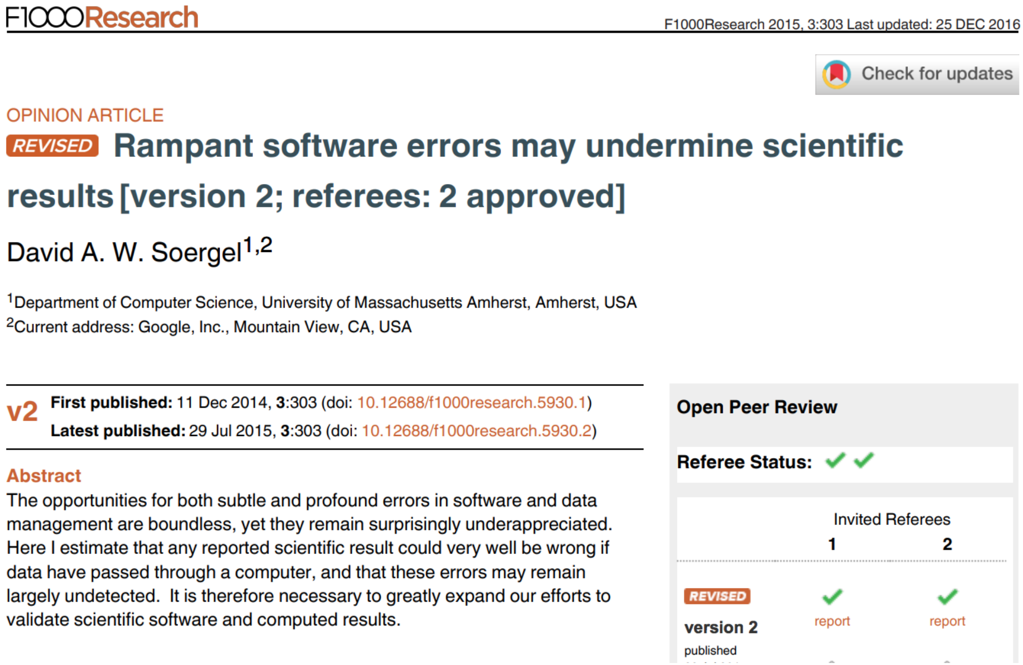
When hackers leaked thousands of e-mails from the Climatic Research Unit (CRU) at the University of East Anglia in Norwich, UK, last year, global-warming sceptics pored over the documents for signs that researchers had manipulated data. No such evidence emerged, but the e-mails did reveal another problem — one described by a CRU employee named "Harry", who often wrote of his wrestling matches with wonky computer software.

"Yup, my awful programming strikes again," Harry lamented in one of his notes, as he attempted to correct a code analysing weather-station data from Mexico.

Although Harry's frustrations did not ultimately compromise CRU's work, his difficulties will strike a chord with scientists in a wide range of disciplines who do a large amount of coding. Researchers are spending more and more time writing computer software to model biological structures, simulate the early evolution of the Universe and analyse past climate data, among other topics. But programming experts have little faith that most scientists are up to the task.

A quarter of a century ago, most of the computing work done by scientists was relatively straightforward. But as computers and programming tools have grown more complex, scientists have hit a "steep learning curve", says James Hack, director of the US National Center for Computational Sciences at Oak Ridge National Laboratory in Tennessee. "The level of effort and skills needed to keep up aren't in the wheelhouse of the average scientist."

As a general rule, researchers do not test or document their programs rigorously, and they rarely release their codes, making it almost impossible to reproduce and verify published results generated by scientific software, say computer scientists. At best, poorly written



Check for updates

OPINION ARTICLE

REVISED Rampant software errors may undermine scientific results [version 2; referees: 2 approved]

David A. W. Soergel[1,2]

[1]Department of Computer Science, University of Massachusetts Amherst, Amherst, USA
[2]Current address: Google, Inc., Mountain View, CA, USA

v2 **First published:** 11 Dec 2014, **3**:303 (doi: 10.12688/f1000research.5930.1)
Latest published: 29 Jul 2015, **3**:303 (doi: 10.12688/f1000research.5930.2)

Abstract

The opportunities for both subtle and profound errors in software and data management are boundless, yet they remain surprisingly underappreciated. Here I estimate that any reported scientific result could very well be wrong if data have passed through a computer, and that these errors may remain largely undetected. It is therefore necessary to greatly expand our efforts to validate scientific software and computed results.

Open Peer Review

Referee Status: ✓ ✓

	Invited Referees	
	1	**2**
REVISED **version 2** published	✓ report	✓ report

Thermo ML: $\sim$20% of papers that otherwise would have been accepted had serious errors.

Stodden (2010) Survey of NIPS re code & data:

Excuse	code	data
Time to document and clean up	77%	54%
Dealing with questions from users	52%	34%
Not receiving attribution	44%	42%
Possibility of patents	40%	N/A
Legal Barriers (i.e. copyright)	34%	41%
Time to verify release with admin	N/A	38%
Potential loss of future publications	30%	35%
Competitors may get an advantage	30%	33%
Web/disk space limitations	20%	29%

Stodden (2010) Survey of NIPS re code & data:

Excuse	code	data
Time to document and clean up	77%	54%
Dealing with questions from users	52%	34%
Not receiving attribution	44%	42%
Possibility of patents	40%	N/A
Legal Barriers (i.e. copyright)	34%	41%
Time to verify release with admin	N/A	38%
Potential loss of future publications	30%	35%
Competitors may get an advantage	30%	33%
Web/disk space limitations	20%	29%

I.e., Fear, greed, ignorance, & sloth

Hacking the limbic system

If I say *just trust me* and I'm wrong, I'm untrustworthy.

Hacking the limbic system

If I say *just trust me* and I'm wrong, I'm untrustworthy.

If I say *here's my work* and it's wrong, I'm honest, human, and serving scientific progress.

Hacking the limbic system

If I say *just trust me* and I'm wrong, I'm untrustworthy.

If I say *here's my work* and it's wrong, I'm honest, human, and serving scientific progress.

Science should be "help me if you can," not "catch me if you can."

Revision-control systems for teaching, research, collaboration

- Teaching use cases:
  - submit homework by pull request (can see commits)
  - collaborate on term projects, create project wikis
  - use for timed exams: push at a coordinated time, pull requests
  - supports automation, including code testing
- Research use cases
  - 1st step of new project: create a repo
  - commits leave breadcrumbs; all changes visible & attributable
  - notes, code, manuscripts, etc. (not ideal for large datasets)
  - issue trackers
  - automated testing & package deployment
  - know last version that worked
- Collaboration use cases
  - parallel development & feature implementation through branches
  - can find last working version of code; *blame*

Scripts & notebook-style tools

- Jupyter (Sweave and knitR are great for papers; less good for workflow), . . .

Scripts & notebook-style tools

- Jupyter (Sweave and knitR are great for papers; less good for workflow), ...
- leave breadcrumbs

Scripts & notebook-style tools

- Jupyter (Sweave and knitR are great for papers; less good for workflow), . . .
- leave breadcrumbs
- readable

Scripts & notebook-style tools

- Jupyter (Sweave and knitR are great for papers; less good for workflow), . . .
- leave breadcrumbs
- readable
- easy to re-run and modify analysis

Scripts & notebook-style tools

- Jupyter (Sweave and knitR are great for papers; less good for workflow), . . .
- leave breadcrumbs
- readable
- easy to re-run and modify analysis
- easy to build on previous analyses

Scripts & notebook-style tools

- Jupyter (Sweave and knitR are great for papers; less good for workflow), . . .
- leave breadcrumbs
- readable
- easy to re-run and modify analysis
- easy to build on previous analyses
- not ideal for production code, packages/libraries, testing

Preproducibility is collaboration w/ people you don’t know,

Preproducibility is collaboration w/ people you don't know,

including yourself next week.

Preproducibility is collaboration w/ people you don't know,

including yourself next week.

Preproducibility & collaboration

- same habits, attitudes, principles, and tools facilitate both
- develop better work habits, *computational hygiene*
- analogue of good lab technique in wet labs

Why work p/reproducibly?

> *There is only one argument for doing something; the rest are arguments for doing nothing. The argument for doing something is that it is the right thing to do.*
>
> *—Cornford, 1908. Microcosmographia Academica*

My top reasons:

1. I feel good about it.

My top reasons:

1. I feel good about it.
2. Others can check my work and correct it if it’s wrong.

My top reasons:

1. I feel good about it.
2. Others can check my work and correct it if it’s wrong.
3. Others (including me) can re-use and extend my work more easily.

How can we do better?

- Scripted analyses: no point-and-click tools, *especially* spreadsheet calculations

How can we do better?

- Scripted analyses: no point-and-click tools, *especially* spreadsheet calculations
- Revision control systems

How can we do better?

- Scripted analyses: no point-and-click tools, *especially* spreadsheet calculations
- Revision control systems
- Documentation, documentation, documentation

How can we do better?

- Scripted analyses: no point-and-click tools, *especially* spreadsheet calculations
- Revision control systems
- Documentation, documentation, documentation
- Coding standards/conventions

How can we do better?

- Scripted analyses: no point-and-click tools, *especially* spreadsheet calculations
- Revision control systems
- Documentation, documentation, documentation
- Coding standards/conventions
- Pair programming

How can we do better?

- Scripted analyses: no point-and-click tools, *especially* spreadsheet calculations
- Revision control systems
- Documentation, documentation, documentation
- Coding standards/conventions
- Pair programming
- Issue trackers

How can we do better?

- Scripted analyses: no point-and-click tools, *especially* spreadsheet calculations
- Revision control systems
- Documentation, documentation, documentation
- Coding standards/conventions
- Pair programming
- Issue trackers
- Code reviews (and in teaching, grade students' *code*, not just their *output*)

How can we do better?

- Scripted analyses: no point-and-click tools, *especially* spreadsheet calculations
- Revision control systems
- Documentation, documentation, documentation
- Coding standards/conventions
- Pair programming
- Issue trackers
- Code reviews (and in teaching, grade students' *code*, not just their *output*)
- Code tests: unit, integration, coverage, regression; automation

Integration tests

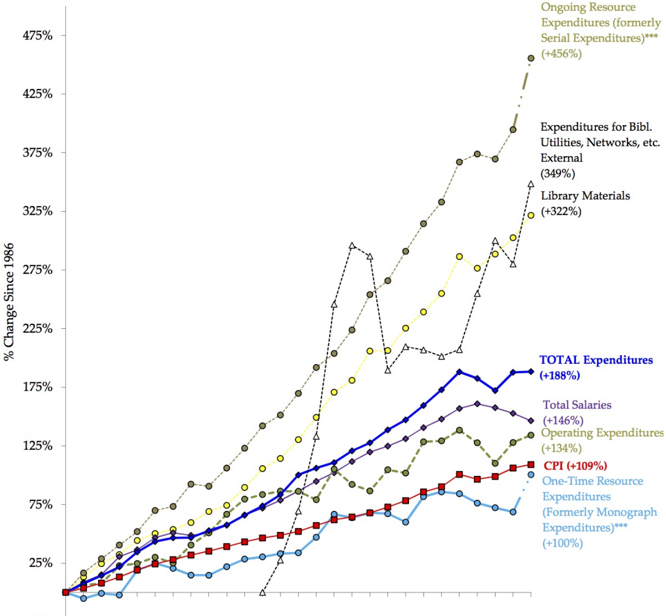
Checklist

1. Don't use spreadsheets for calculations.
2. Script your analyses, including data cleaning and munging.
3. Document your code.
4. Record and report software versions, including library dependencies.
5. Use unit tests, integration tests, coverage tests, regression tests.
6. Avoid proprietary software that doen't have an open-source equivalent.
7. Report all analyses tried (transformations, tests, variable selection, model selection, etc.), not just the final analysis
8. Make code and code tests available.
9. Make data available in an open format; provide data dictionary.
10. Publish in open journals.

Why open publication?

- Research funded by agencies
- Conducted at universities by faculty et al.
- Refereed/edited for journal by faculty at no cost to journal
- Pages charges paid by agencies
- Exclusionary & morally questionable for readers have to pay to view

ARL 1986-2016 Also CFUCBL rept



What’s the role of a journal?

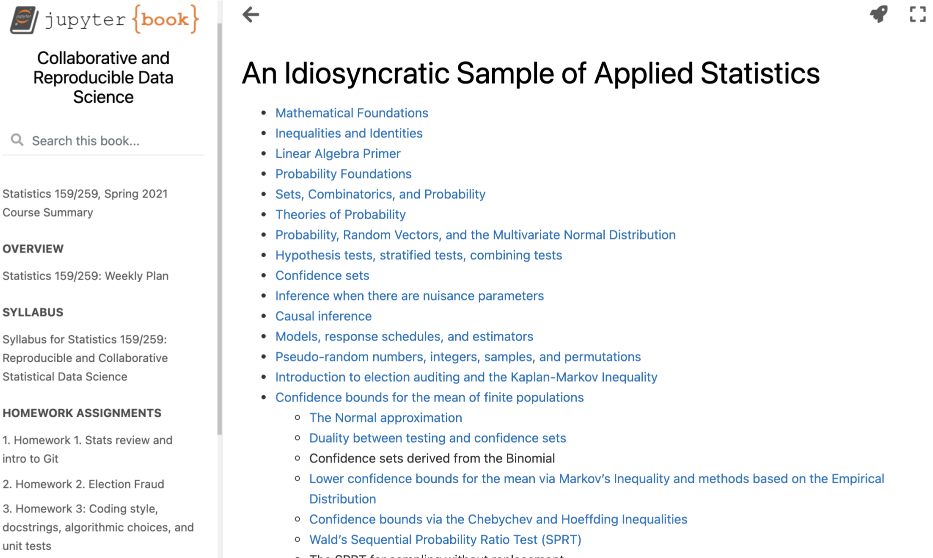
- Gatekeeping/QC by editors & referees
- Dissemination/discoverability
- Archive

cess were not immediately fulfilled; second, it assured the reader that the relator was not wilfully suppressing inconvenient evidence, that he was in fact being faithful to reality. Complex and circumstantial accounts were to be taken as undistorted mirrors of complex experimental outcomes.[87] So, for example, it was not legitimate to hide the fact that air-pumps sometimes did not work properly or that they often leaked: ". . . I think it becomes one, that professeth himself a faithful relator of experiments not to conceal" such unfortunate contingencies.[88] It is, however, vital to keep in mind that in his circumstantial accounts Boyle proffered only a *selection* of possible contingencies. There was not, nor can there be, any such thing as a report that notes *all* circumstances that might affect an experiment. Circumstantial, or stylized, accounts do not, therefore, exist as pure forms but as publicly acknowl-

It’s hard to teach an old dog new tricks.

It's hard to teach an old dog new tricks.

So work with puppies!



Collaborative and Reproducible Data Science

Statistics 159/259, Spring 2021 Course Summary

OVERVIEW

Statistics 159/259: Weekly Plan

SYLLABUS

Syllabus for Statistics 159/259: Reproducible and Collaborative Statistical Data Science

HOMEWORK ASSIGNMENTS

1. Homework 1. Stats review and intro to Git
2. Homework 2. Election Fraud
3. Homework 3: Coding style, docstrings, algorithmic choices, and unit tests

An Idiosyncratic Sample of Applied Statistics

- Mathematical Foundations
- Inequalities and Identities
- Linear Algebra Primer
- Probability Foundations
- Sets, Combinatorics, and Probability
- Theories of Probability
- Probability, Random Vectors, and the Multivariate Normal Distribution
- Hypothesis tests, stratified tests, combining tests
- Confidence sets
- Inference when there are nuisance parameters
- Causal inference
- Models, response schedules, and estimators
- Pseudo-random numbers, integers, samples, and permutations
- Introduction to election auditing and the Kaplan-Markov Inequality
- Confidence bounds for the mean of finite populations
  - The Normal approximation
  - Duality between testing and confidence sets
  - Confidence sets derived from the Binomial
  - Lower confidence bounds for the mean via Markov's Inequality and methods based on the Empirical Distribution
  - Confidence bounds via the Chebychev and Hoeffding Inequalities
  - Wald's Sequential Probability Ratio Test (SPRT)
  - The SPRT for sampling without replacement

Created in 2015. Joint undergrad/grad

- Git, Github, issue trackers
- pair programming; in-class code reviews
- unit tests, integration tests
- automation, make, Github actions
- Jupyter, python, PEP-8, PEP-257
- virtual machines, containers, myBinder
- yaml for requirements
- contribute, code, documentation, & tests to open-source software projects

Teach by doing science preproducibly; don't focus on tools.

Eyes on the code, not just the output!

Main assignments in 2021:

- reproducing calculations in Cicchetti's expert report re voter fraud
- reproducing SIR models of COVID-19 spread and mortality
- implementing several approaches to exact confidence bounds for binomial an hypergeometric
- implementing nonparametric confidence bounds for causal inference w binary outcomes
- application to the efficacy of Regeneron antibody cocktail for preventing COVID-19
- contributing to `cryptorandom` and `permute`

All work done using Git, JupyterLab (python & some R); collaboration by Slack.

Work submitted using Git Classroom

Daily practice of good "computational hygiene" to do useful work

2019

- reproduce/critique study of impact of land use on food contamination
- reproduce/critique study of impact of climate change on violent crime
- critique study of impact of player skin tone on soccer penalties

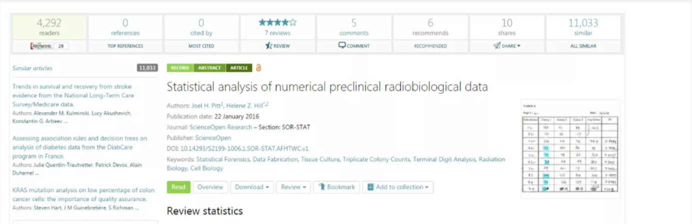
In: Peer Review ▸

A post-publication peer review success story

September 29, 2017 * Author: Jon Tennant

In 2016, Dr. Joel Pitt and Prof. Helene Hill published an important paper in ScienceOpen Research. In their paper, they propose new statistical methods to detect scientific fraudulent data. Pitt and Hill demonstrate the use of their method on a single case of suspected fraud. Crucially, in their excellent effort to combat fraud, Pitt and Hill make the raw data on which they tested their method publicly available on the Open Science Framework (OSF). Considering that a single case of scientific fraud can cost institutions and private citizens a huge amount of money, their result is provocative, and it emphasizes how important it is to make the raw data of research papers publicly available.

The Pitt and Hill (2016) article was read and downloaded almost 100 times a day since its publication on ScienceOpen. More importantly, it now has 7 independent post-publication peer reviews and 5 comments. Although this is a single paper in ScienceOpen's vast index of 28 million research articles (all open to post-publication peer review!), the story of how this article got so much attention is worth re-telling.



4,292 readers	0 references	0 cited by	★★★★☆ 7 reviews	5 comments	6 recommends	10 shares	11,033 similar
Altmetric 28	TOP REFERENCES	MOST CITED	REVIEW	COMMENT	RECOMMENDED	SHARE ▾	ALL SIMILAR

Similar articles 11,033

Trends in survival and recovery from stroke: evidence from the National Long-Term Care Survey/Medicare data.
Authors: Alexander M. Kulminski, Lucy Akushevich, Konstantin G. Arbeev ...

Assessing association rules and decision trees on analysis of diabetes data from the DiabCare program in France.
Authors: Julie Quentin-Trautvetter, Patrick Devos, Alain Duhamel ...

KRAS mutation analysis on low percentage of colon cancer cells: the importance of quality assurance.
Authors: Steven Hart, J M Guinebretière, S Richman ...

Statistical analysis of numerical preclinical radiobiological data

Authors: Joel H. Pitt[1], Helene Z. Hill[*,2]

Publication date: 22 January 2016

Journal: ScienceOpen Research – Section: SOR-STAT

Publisher: ScienceOpen

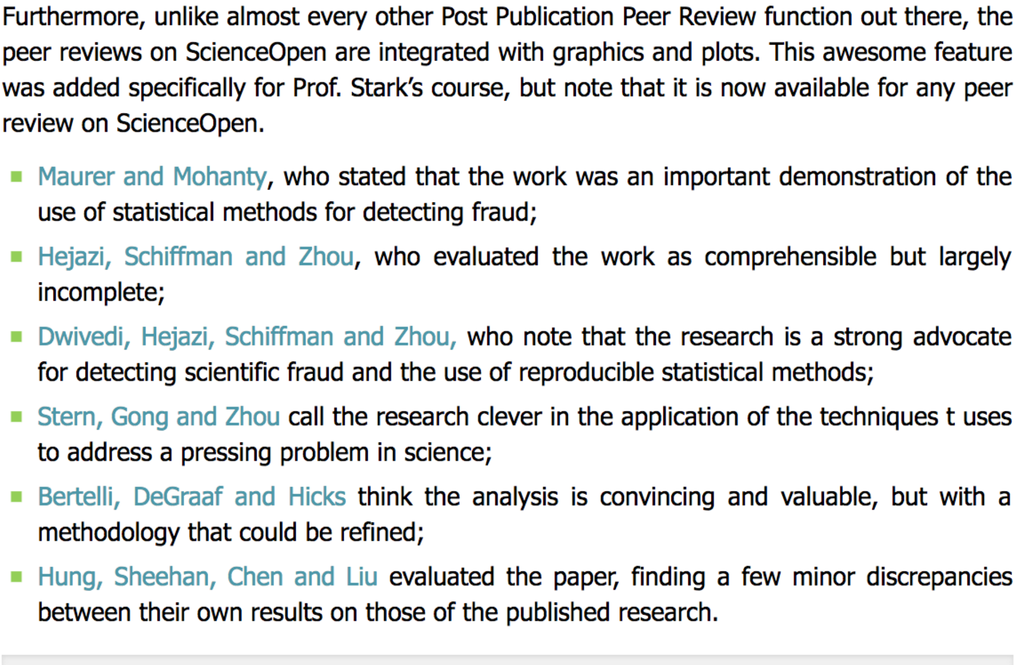
DOI: 10.14293/S2199-1006.1.SOR-STAT.AFHTWC.v1

Keywords: Statistical Forensics, Data Fabrication, Tissue Culture, Triplicate Colony Counts, Terminal Digit Analysis, Radiation Biology, Cell Biology

Review statistics

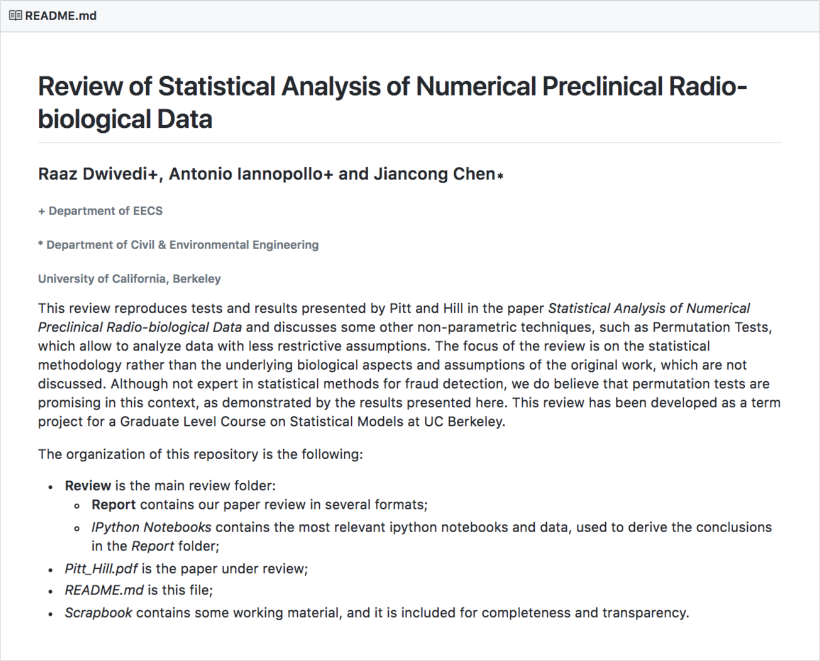
- Get students thinking about alternative models for scholarly publication;
- Get students thinking about reproducibility and open science;
- Get students to work collaboratively on a data analysis project that involves thinking hard about the underlying science;
- Get students to register with ORCID;
- Get students to post their analyses on GitHub so that their own work is reproducible/extensible;
- Get students their first scientific publication.

For another step of Open Science brilliance, the reviews themselves sought to be completely reproducible, with the code for all the students’ calculations is available on GitHub (eg here and here)!



Furthermore, unlike almost every other Post Publication Peer Review function out there, the peer reviews on ScienceOpen are integrated with graphics and plots. This awesome feature was added specifically for Prof. Stark's course, but note that it is now available for any peer review on ScienceOpen.

- Maurer and Mohanty, who stated that the work was an important demonstration of the use of statistical methods for detecting fraud;
- Hejazi, Schiffman and Zhou, who evaluated the work as comprehensible but largely incomplete;
- Dwivedi, Hejazi, Schiffman and Zhou, who note that the research is a strong advocate for detecting scientific fraud and the use of reproducible statistical methods;
- Stern, Gong and Zhou call the research clever in the application of the techniques t uses to address a pressing problem in science;
- Bertelli, DeGraaf and Hicks think the analysis is convincing and valuable, but with a methodology that could be refined;
- Hung, Sheehan, Chen and Liu evaluated the paper, finding a few minor discrepancies between their own results on those of the published research.



README.md

Review of Statistical Analysis of Numerical Preclinical Radio-biological Data

Raaz Dwivedi+, Antonio Iannopollo+ and Jiancong Chen*

+ Department of EECS

* Department of Civil & Environmental Engineering

University of California, Berkeley

This review reproduces tests and results presented by Pitt and Hill in the paper *Statistical Analysis of Numerical Preclinical Radio-biological Data* and discusses some other non-parametric techniques, such as Permutation Tests, which allow to analyze data with less restrictive assumptions. The focus of the review is on the statistical methodology rather than the underlying biological aspects and assumptions of the original work, which are not discussed. Although not expert in statistical methods for fraud detection, we do believe that permutation tests are promising in this context, as demonstrated by the results presented here. This review has been developed as a term project for a Graduate Level Course on Statistical Models at UC Berkeley.

The organization of this repository is the following:

- **Review** is the main review folder:
  - **Report** contains our paper review in several formats;
  - *IPython Notebooks* contains the most relevant ipython notebooks and data, used to derive the conclusions in the *Report* folder;
- *Pitt_Hill.pdf* is the paper under review;
- *README.md* is this file;
- *Scrapbook* contains some working material, and it is included for completeness and transparency.

Pledge

A. I will not referee any article that does not contain enough information to tell whether it is correct.
B. Nor will I submit any such article for publication.
C. Nor will I cite any such article published after [date].

See also *Open science peer review oath* http://f1000research.com/articles/3-271/v2

Show your work!

Resources

- Data Carpentry, Software Carpentry
- RunMyCode, Research Compendia, FigShare
- Jupyter (>40 languages!), Sweave, RStudio, knitr
- Reproducibility initiative http://validation.scienceexchange.com/#/reproducibility-initiative
- Best practices for scientific software dev http://arxiv.org/pdf/1210.0530v4.pdf
- Federation of American Societies for Experimental Biology
- Was ist open science? http://opensciencesasap.org/open-science/