Reproducibility and Error

Charles J. Geyer

School of Statistics
Minnesota Center for Philosophy of Science
University of Minnesota

November 5, 2022

Slides for this talk:
http://users.stat.umn.edu/~geyer/repro.pdf
Draft paper for this talk:
http://users.stat.umn.edu/~geyer/repro-paper.pdf

Outline

- Most scientific papers with statistics have conclusions not actually supported because of
  - mathematical or computational error,
  - statistical procedures inappropriate for the data or
  - not leading to the inferences claimed.
- Good computing practices
  - version control,
  - testing (quality control),
  - code reviews,
  - literate programming

  are essential for correct computing.
- Failure to do all calculations from raw data to conclusions in a way that is fully reproducible and available in a permanent public repository is, by itself, a questionable research practice.
- Failure to do statistics as if it could have been pre-registered is a questionable research practice.

Outline (cont.)

- Journals that use $P < 0.05$ as a criterion of publication are not scientific journals (publishing only one side of a story is as unscientific as it is possible to be).
- Statistics should be adequately described, at least in the supplementary material.
- Scientific papers whose conclusions depend on nontrivial statistics should have statistical referees, and those referees should be heeded.
- Not all errors are describable by statistics. There is also what physicists call *systematic error* that is the same in every replication of an experiment. Physicists regularly attempt to quantify this. Others should too.

Misuse of Statistics

Statistical procedures are derived by rigorous math. Like theorems, they have *assumptions* and *conclusions* (also called inferences or interpretations).

Users of statistics (including most scientists) often pay no attention to this math. So that gives us the three kinds of mistakes listed in the outline.

- They botch the calculations.
- They ignore the assumptions (so the procedure is not validly applied).
- They ignore the conclusions (so the procedure does not say what they want it to say).

Misuse of Statistics (cont.)

I have a joke that to most scientists $P < 0.05$ means "statistics has proved that every idea I have ever had on this subject is correct."

No scientist would be brazen enough to say that, but many seem argue that.

Many statistical hypothesis tests have null hypotheses that are mere straw men to knock down. When the test duly knocks them down, the proper interpretation is that the test has shown that *something* is going on in the data, but not necessarily what the scientists want to claim.

But "these data are not worthless" is not a strong story, so scientists routinely go beyond what the statistics actually says.

Software Crisis

The term "software crisis" was coined in 1968. It is still ongoing.

The term "software engineering" was coined in 1966. It refers to all of the methodology used to ameliorate (*but not cure*) the software crisis.

Still today, large companies employing the best programmers available, produce software that is buggy, difficult to use, and late. And they are *using all* of the software engineering best practices.

Most scientists doing statistical computing *use none* of the software engineering best practices. Furthermore they have little or no training in statistics and computing.

Why should we expect they somehow magically do better than the professionals?

Software Crisis (cont.)

At the very least scientists should use

- version control,
- software testing,
- code reviews, and
- literate programming.

There are other things they could pick up from software engineering, but these are IMHO the most important.

Fully Reproducible Computing

All computing for all scientific papers that involve computing should have

- all code available in a permanent public repository,
- ditto for all quality control tests,
- all code fully explained and justified (literate programming),

Many journals now require (with loopholes) all data to be in a permanent public repository. Few do this for code. All should.

No paper that lacks the above can be trusted.

Even papers that do have the above can only be trusted after you have checked their code and quality control.

Data Snooping versus Pre-Registration

Long before there was any talk of the "reproducibility crisis" statisticians inveighed against data snooping, also called data dredging or cherry picking.

More formally, this is failure to report all statistical procedures tried plus failure to correct for the multiplicity of procedures tried (such corrections have been in the statistics literature for seventy years).

A large part of the "reproducibility crisis" is just scientists ignoring how statisticians have *always* said statistics should be done.

Data Snooping versus Pre-Registration (cont.)

Pre-registration is a device to prevent data snooping.

It only works if referees actually compare the paper to the pre-registration and insist that the analysis done is the pre-registered one.

But the main point is no data dredging.

Papers should be done as if they could have been pre-registered.

- All procedures tried reported.
- Corrections for multiple procedures.
- It is believable that the procedures could have been chosen before data were collected.

Publication Bias versus Pre-Acceptance

Publication bias is journals favoring "positive" results.

This is grossly unscientific (suppressing one side of the story is about as unscientific as it is possible to be).

So journals and their editors and referees should not do that.

They should accept papers describing experiments or observational studies well done and reject the poorly done. Acceptance should not depend on whether the results are "positive" or "negative" (in scare quotes because the whole notion of results being "positive" or "negative" is grossly unscientific).

Publication Bias versus Pre-Acceptance

Pre-acceptance is a device to prevent publication bias.

It ties the hands of editors and referees. They commit to accepting the paper (regardless of whether the results are "positive" or "negative" so long as the scientists do what their pre-registration said they would do).

Editors and referees should behave as if papers were pre-accepted. Accept good papers and reject bad ones *regardless of whether results are "positive" or "negative."*

Carelessness About Statistics

Authors, editors, and referees are often careless about statistics.

Most papers inadequately describe statistics. Often readers cannot guess what was done. Many editors and referees accept this as normal.

No such paper is worth reading.

Papers using nontrivial statistics rarely get referees competent to critique the statistics. What value refereeing then?

Systematic Error

Physicists often discuss systematic error, which is non-random error not describable or controllable with statistics.

Discussion of systematic error is partly theoretical and partly experimental: what is being measured and how well does the equipment measure it?

Operationally defined quantities (IQ is what IQ tests measure) cannot have systematic error. But theoretically defined quantities can.

There are theoretically defined quantities in areas other than physics. Other sciences should try to analyze systematic error. Ask how wrong could we be?