Embrace variation and accept uncertainty

Jesper W. Schneider
Danish Centre for Studies in Research and Research Policy,
Aarhus University, Denmark
jws@ps.au.dk

Somethings wrong … !


Open access, freely available online

Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis

Summary

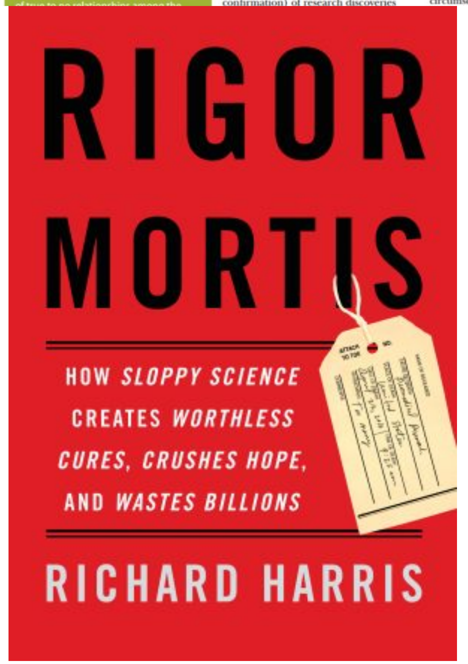
There is increasing concern that most current published research findings are false. The probability that a research claim is true may depend on study power and bias, the number of other studies on the same question, and, importantly, the ratio

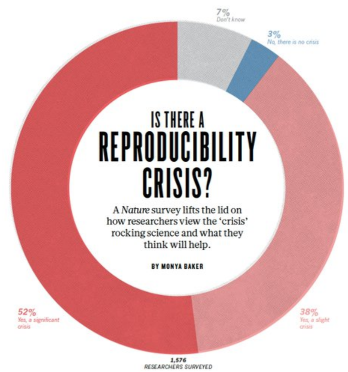
factors that influence this problem and some corollaries thereof.

Modeling the Framework for False Positive Findings

Several methodologists have pointed out [9–11] that the high rate of nonreplication (lack of confirmation) of research discoveries

is characteristic of the field and can vary a lot depending on whether the field targets highly likely relationships or searches for only one or a few true relationships among thousands and millions of hypotheses that may be postulated. Let us also consider, for computational simplicity, circumscribed fields where either there





But the scholarly process is fallible, it is natural that we are wrong sometimes!

Science is self-correcting!

Not true for all branches of the “sciences”

… especially not the soft sciences!

If you didn’t know …

The Hardest Science

Everything is fucked: The syllabus

PUBLISHED ON *August 11, 2016*

PSY 607: Everything is Fucked
Prof. Sanjay Srivastava
Class meetings: Mondays 9:00 – 10:50 in 257 Straub
Office hours: Held on Twitter at your convenience (@hardsci)

In a much-discussed article at Slate, social psychologist Michael Inzlicht told a reporter, "Meta-analyses are fucked" (Engber, 2016). What does it mean, in science, for something to be fucked? Fucked needs to mean more than that something is complicated or must be undertaken with thought and care, as that would be trivially true of everything in science. In this class we will go a step further and say that something is fucked if it presents hard conceptual challenges to which implementable, real-world solutions for working scientists are either not available or routinely ignored in practice.

The format of this seminar is as follows: Each week we will read and discuss 1-2 papers that raise the question of whether something is fucked. Our focus will be on things that may be fucked in research methods, scientific practice, and philosophy of science. The potential fuckedness of specific theories, research topics, etc. will not be the focus of this class per se, but rather will be used to illustrate these important topics. To that end, each week a different student will be assigned to find a paper that illustrates the fuckedness (or lack thereof) of that week's topic, and give a 15-minute presentation about whether it is indeed fucked.

Grading:

- 20% Attendance and participation
- 30% In-class presentation
- 50% Final exam

Week 1: Psychology is fucked

Meehl, P. E. (1990). Why summaries of research on psychological theories are often uninterpretable. *Psychological Reports, 66*, 195-244.

Sanjay Srivastava

About the blog

Follow new posts by email

Join 1,412 other followers

Enter your email address

SIGN ME UP!

On Twitter

Tweets by @hardsci

Sanjay Srivastava Retweeted
Christopher Chartier @CRChartier
It is time to begin building a "CERN for psychological science"
christopherchartier.com/2017/08/26/bui...

Building a CERN for Psyc...
In response to the reproduc...
christopherchartier.com

Aug 26, 2017

Sanjay Srivastava Retweeted
The Dogfather @matt__nelson
BODY: i'm exhausted
BRAIN: neat
BODY: let's sleep

The syllabus:
Week 1: Psychology is fucked
Week 2: Significance testing is fucked
Week 3: Causal inference from experiments is fucked
Week 4: Mediation is fucked
Week 5: Covariates are fucked
Week 6: Replicability is fucked
Week 7: Interlude: Everything is fine, calm the fuck down
Week 8: Scientific publishing is fucked
Week 9: Meta-analysis is fucked
Week 10: The scientific profession is fucked

At least when it comes to the empirical knowledge production model in the soft sciences that ritualistically adheres to significance test

The culprit:

“GASSSPP”

Generally **A**ccepted **S**oft **S**ocial **S**cience **P**ublishing **P**rocess

- suffers from a number of deficiencies which not only make this kind of work less than scientific, it actually defeat the objectives of science and produces an endless stream of publications with unsupported or false claims
- … only useful for career progression!

Hot topic, but warnings are ol

SOME DIFFICULTIES OF INTERPRETATION EN-
COUNTERED IN THE APPLICATION
OF THE CHI-SQUARE TEST*

By Joseph Berkson, M.D.
Division of Biometry and Medical Statistics,
The Mayo Clinic, Rochester, Minnesota

THE remarks that I have to make are not
erations of the mathematics of the chi-s
siderable interest in mathematical statistics,
in it. You will not hear anything about card
from me. I shall speak as a practitioner w
the test to real observations, made seriousl
crete scientific problems. I have used the ch
decisions as to the character of experimer
which I had every reason to think it was a
in the same spirit in which we doctors use,
to help make decisions in situations where
have syphilis. In the course of these experi
numerous situations in which the test did n
function for which I thought I could use
few examples *seriatim:*

I. I believe that an observant statistici
siderable experience with applying the chi-s
agree with my statement that, as a matter
numbers in the data are quite large, the P
Having observed this, and on reflection, I
matic statement, referring for illustration
the normal curve is fitted to a body of d
observations whatever of quantities in the
number of observations is extremely large—
of 200,000—the chi-square P will be small
significance."

This dogmatic statement is made on the b
the observation referred to and can also be
from *a priori* considerations. For we may as
certain that any series of real observations
normal curve *with absolute exactitude* in all re

* A paper presented at the Ninety-ninth Annual Meeting o
Atlantic City, New Jersey, December 27, 1937.

[1] In this discussion I mean, by the chi-square test, the co
which chi-square is the sum of the terms $(o-t)^2/t$ calculated f
quencies, not other tests using the chi-square distribution, such
difference of an observed and theoretical variance.

Psychological Bulletin
1960, Vol. 57, No. 5, 416–428

THE FALLACY OF THE NULL-HYPOTHESIS SIGNIFICANCE TEST

WILLIAM W. ROZEBOOM
St. Olaf College

The theory of probability and statistical inference is various things to various people. To the mathematician, it is an intricate formal calculus, to be explored and developed with little professional concern for any empirical significance that might attach to the terms and propositions involved. To the philosopher, it is an embarrassing mystery whose justification and conceptual clarification have remained stubbornly refractory to philosophical insight. (A famous philosophical epigram has it that induction [a special case of statistical inference] is the glory of science and the scandal of philosophy.) To the experimental scientist, however, statistical inference is a research instrument, a processing device by which unwieldy masses of raw data may be refined into a product more suitable for assimilation into the corpus of science, and in this lies both strength and weakness. It is strength in that, as an ultimate *consumer* of statistical methods, the experimentalist is in position to demand that the techniques made available to him conform to his actual needs. But it is also weakness in that, in his need for the tools constructed by a highly technical formal discipline, the experimentalist, who has specialized along other lines, seldom feels competent to extend criticisms or even comments; he is much more likely to make unquestioning application of procedures learned more or less by rote from persons assumed to be more knowledgeable of statistics than he. There is, of course, nothing surprising or reprehensible about this—one need not understand the principles of a complicated tool in order to make effective use of it, and the research scientist can no more be expected to have sophistication in the theory of statistical inference than he can be held responsible for the principles of the computers, signal generators, timers, and other complex modern instruments to which he may have recourse during an experiment. Nonetheless, this leaves him particularly vulnerable to misinterpretation of his aims by those who build his instruments, not to mention the ever present dangers of selecting an inappropriate or outmoded tool for the job at hand, misusing the proper tool, or improvising a tool of unknown adequacy to meet a problem not conforming to the simple theoretical situations in terms of which existent instruments have been analyzed. Further, since behaviors once exercised tend to crystallize into habits and eventually traditions, it should come as no surprise to find that the tribal rituals for data-processing passed along in graduate courses in experimental method should contain elements justified more by custom than by reason.

In this paper, I wish to examine a dogma of inferential procedure which, for psychologists at least, has attained the status of a religious conviction. The dogma to be scrutinized is the "null-hypothesis significance test" orthodoxy that passing statistical judgment on a scientific hypothesis by means of experimental observa-

Journal of Consulting and Clinical Psychology
1978, Vol. 46, 806-834.

#11

Theoretical Risks and Tabular Asterisks:
Sir Karl, Sir Ronald, and the Slow Progress of Soft Psychol

Paul E. Meehl
University of Minnesota

Theories in "soft" areas of psychology lack the cumulative character of sci
knowledge. They tend neither to be refuted nor corroborated, but instead merely
away as people lose interest. Even though intrinsic subject matter difficulties (20
contribute to this, the excessive reliance on significance testing is partly respo
being a poor way of doing science. Karl Popper's approach, with modifications, wo
prophylactic. Since the null hypothesis is quasi-always false, tables summarizing re
in terms of patterns of "significant differences" are little more than complex, ca
uninterpretable outcomes of statistical power functions. Multiple paths to esti
numerical point values ("consistency tests") are better, even if approximate with
tolerances; and lacking this, ranges, orderings, second-order differences, curve peal
valleys, and function forms should be used. Such methods are usual in deve
sciences that seldom report statistical significance. Consistency tests of a conje
taxometric model yielded 94% success with zero false negatives.

I had supposed that the title gave an easy tipoff to my topic, but some puzzled reactions by my Minnesota colleagues show otherwise, which heartens me because it suggests that what I am about to say is not trivial and universally known. The two knights are Sir Karl Raimund Popper (1959, 1962, 1972; Schilpp, 1974) and Sir Ronald Aylmer Fisher (1956, 1966, 1967), whose respective emphases on subjecting scientific theories to grave danger of refutation (that's Sir Karl) and major reliance on tests of statistical significance (that's Sir Ronald) are, at least in current practice, not well integrated—perhaps even incompatible. If you have not been accustomed to thinking about this incoherency, and my remarks lead you to do so (whether or not you end up agreeing with me), this article will have served its scholarly function.

I consider it unnecessary to persuade you that most so-called "theories" in the soft areas

of psychology (clinical, counselin
personality, community, and schoo
ogy) are scientifically unimpres
technologically worthless. Docume
statement would of course require a
able amount of time, but you can q
the flavor by having a look at Brau
Fiske (1974); Gergen (1973); Hoga
and Solano (1977); McGuire (197
(1960/1973a, 1959/1973f); Mische
Schlenker (1974); Smith (1973); an
(1973). These are merely some hi
and forceful samples; I make no
bibliographic completeness on the la
of "What's wrong with 'soft' psych
beautiful hatchet job, which in m
should be required reading for all P
dates, is by the sociologist Andres
Perhaps the easiest way to convinc
is by scanning the literature of sof
ogy over the last 30 years and noti

This article is based on a lecture delivered at the meeting of the American Psychological Association, Washington, D.C., September 1976, on the occasion of the author's receiving Division 12, Section 3's Distinguished Scientist Award. The research reported here was assisted by Grant MH 24224 from the National Institute of Mental Health and the University of Minnesota Psychiatry Research Fund. Completion of the article was aided by a James McKeen Cattell Fund sabbatical award.

Requests for reprints should be sent to P. E. Meehl, University of Minnesota, Department of Psychology, 75 East River Road, Minneapolis MN 55455-0344.

Journal of Informetrics 7 (2013) 50–62

Contents lists available at SciVerse ScienceDirect

Journal of Informetrics

journal homepage: www.elsevier.com/locate/joi

ELSEVIER

Review

Caveats for using statistical significance tests in research assessments

Jesper W. Schneider*

Danish Centre for Studies in Research and Research Policy, Department of Political Science & Government, Aarhus University, Finlandsgade 4, DK-8200, Aarhus N, Denmark

A R T I C L E   I N F O

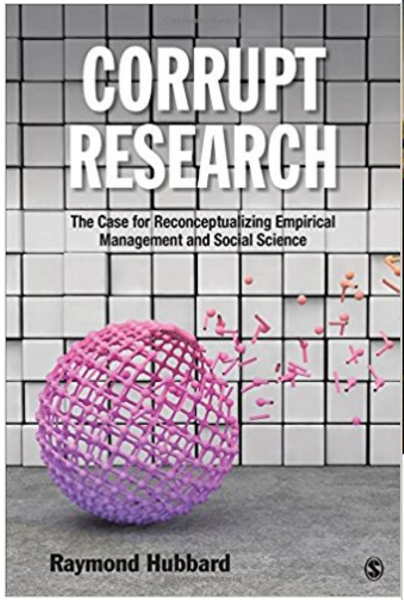
Article history:
Received 12 December 2011
Received in revised form 8 August 2012
Accepted 16 August 2012

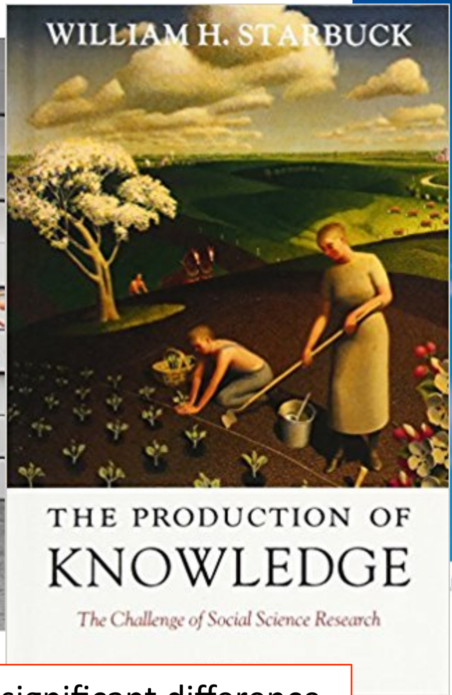
Keywords:
Statistical significance tests
Problems and misuse
Research assessment

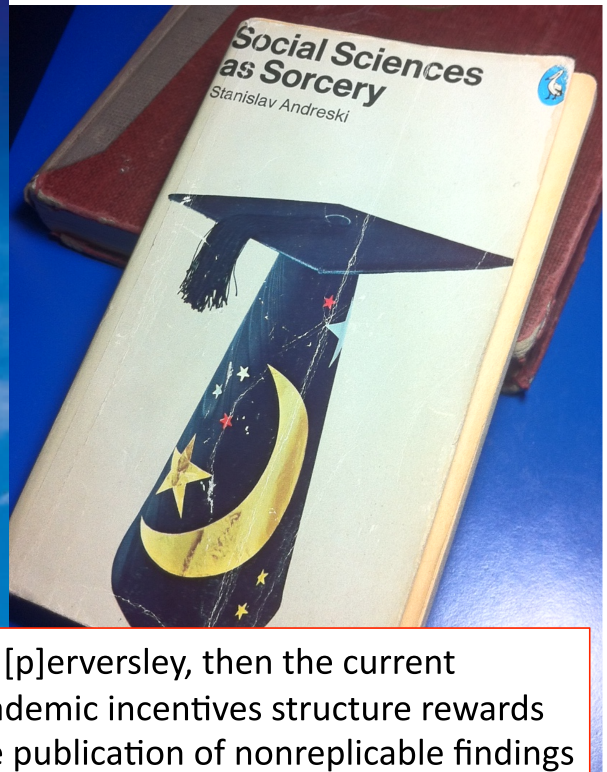
A B S T R A C T

This article raises concerns about the advantages of using statistical significance tests in research assessments as has recently been suggested in the debate about proper normalization procedures for citation indicators by Opthof and Leydesdorff (2010). Statistical significance tests are highly controversial and numerous criticisms have been leveled against their use. Based on examples from articles by proponents of the use statistical significance tests in research assessments, we address some of the numerous problems with such tests. The issues specifically discussed are the ritual practice of such tests, their dichotomous application in decision making, the difference between statistical and substantive significance, the implausibility of most null hypotheses, the crucial assumption of randomness, as well as the utility of standard errors and confidence intervals for inferential purposes. We argue that applying statistical significance tests and mechanically adhering to their results are highly problematic and detrimental to critical thinking. We claim that the use of such tests do not provide any advantages in relation to deciding whether differences between citation indicators are important or not. On the contrary their use may be harmful. Like many other critics, we generally believe that statistical significance tests are over- and misused in the empirical sciences including scientometrics and we encourage a reform on these matters.

Contents







"… science following a significant difference philosophy tilts toward the production of suspect empirical literatures"

"… [p]erversley, then the current academic incentives structure rewards the publication of nonreplicable findings … a conclusion so disheartening that some contest as mythical the idea that science is self-correcting"

Notice, this is not a qualitative versus quantitative issue or positivist versus constructivist nor what social science is all about

It is simply the fact that the empirical knowledge production model centred on significance testing is broken!

… and this has severe and important consequences

Are we then f….d?

Take a look at our journals
…

Thirteen Dutch universities and ten principles in the Leiden Ranking 2017.

This is a reaction to https://www.cwts.nl/blog?article=n-r2q274&title=ten-principles-for-the-responsible-use-of-university-rankings

Under principle 6, you formulate as follows: “To some extent it may be possible to quantify uncertainty in university rankings (e.g., using stability intervals in the Leiden Ranking), but to a large extent one needs to make an intuitive assessment of this uncertainty. In practice, this means that it is best not to pay attention to small performance differences between universities.”

It seems to me of some relevance whether minor differences are significant or not. The results can be counter-intuitive. At the occasion of the Leiden Ranking 2011, Lutz Bornmann and I therefore developed a tool in Excel that enables the user to test (i) the difference between two universities on its significance and (ii) for each university the difference between its participation in the top-10% cited publications versus the *ceteris-paribus* expectation of 10% participation (Leydesdorff & Bornmann, 2012). Does the university perform above or below expectation?

The Excel sheet containing the test can be retrieved at http://www.leydesdorff.net/leiden11/leiden11.xls . In response to concerns similar to yours about using significance tests expressed by (Cohen, 1994; Schneider, 2013; Waltman, 2016), we added effect sizes to the tool (Cohen, 1988) . However, the weights of effect sizes are more difficult to interpret than *p*-values indicating a significance level.

For example, one can raise the question of whether the relatively small differences among Dutch universities indicate that they can be considered as a homogenous set. This is the intuitive assessment which dominates in the Netherlands. Using the stability intervals on your website, however, one can show that there are two groups: one in the western part of the country (the “randstad”) and another in more peripheral regions with significantly lower scores in terms of the top-10 publication (PP10). Figure 1 shows the division.

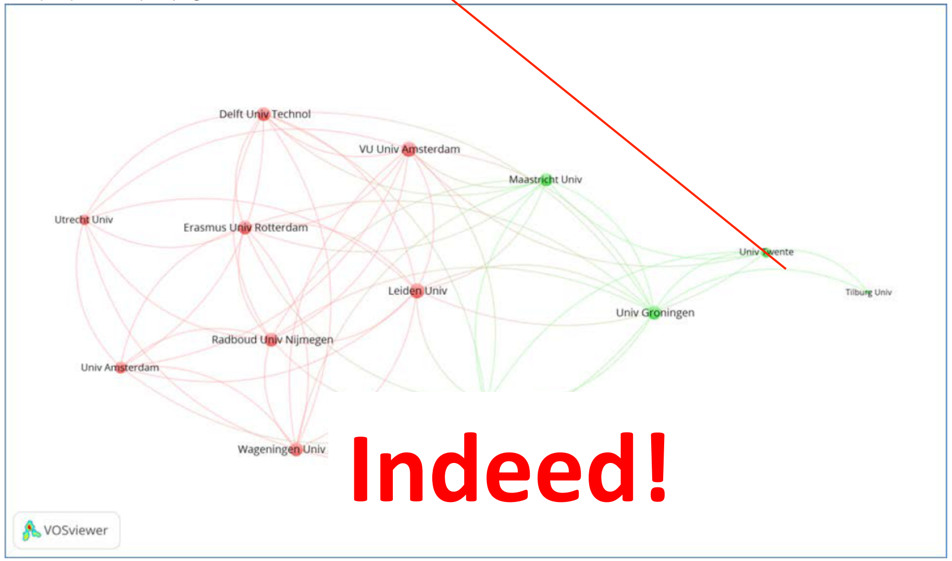


Figure 1: Thirteen Dutch universities grouped into two statistically homogenous sets on the basis of the Leiden Rankings 2017. Stability intervals used as methodology. (If not visible, see the version at http://www.leydesdorff.net/leiden17/index.htm)

Researchers want certainty!

… what we have produced is “over-certainty”
… what we have to accept is the presence of uncertainty (and variation)
… no statistical tool can remove it!

Something has to happen

- The ritualistic use, misuse and misunderstanding of significance tests is one of the most damaging factors to the flawed soft science knowledge production model

This has never happened before!

THE AMERICAN STATISTICIAN
2016, VOL. 70, NO. 2, 129–133
http://dx.doi.org/10.1080/00031305.2016.1154108

EDITORIAL

The ASA's Statement on *p*-Values: Context, Process, and Purpose

In February 2014, George Cobb, Professor Emeritus of Mathematics and Statistics at Mount Holyoke College, posed these questions to an ASA discussion forum:

> Q: Why do so many colleges and grad schools teach $p = 0.05$?
> A: Because that's still what the scientific community and journal editors use.
> Q: Why do so many people still use $p = 0.05$?
> A: Because that's what they were taught in college or grad school.

Cobb's concern was a long-worrisome circularity in the sociology of science based on the use of bright lines such as $p < 0.05$: "We teach it because it's what we do; we do it because it's what we teach." This concern was brought to the attention of the ASA Board.

The ASA Board was also stimulated by highly visible discussions over the last few years. For example, ScienceNews (Siegfried 2010) wrote: "It's science's dirtiest secret: The 'scientific method' of testing hypotheses by statistical analysis stands on a flimsy foundation." A November 2013, article in Phys.org Science News Wire (2013) cited "numerous deep flaws" in null hypothesis significance testing. A ScienceNews article (Siegfried 2014) on February 7, 2014, said "statistical techniques for testing hypotheses …have more flaws than Facebook's privacy policies." A week later, statistician and "Simply Statistics" blogger Jeff Leek responded. "The problem is not that people use P-values poorly," Leek wrote, "it is that the vast majority of data analysis is not performed by people properly trained to perform data analysis" (Leek 2014). That same week, statistician and science writer Regina Nuzzo published an article in *Nature* entitled "Scientific Method: Statistical Errors" (Nuzzo 2014). That article is now one of the most highly viewed *Nature* articles, as reported by altmetric.com (*http://www.altmetric.com/details/2115792#score*).

Of course, it was not simply a matter of responding to some articles in print. The statistical community has been deeply concerned about issues of *reproducibility* and *replicability* of scientific conclusions. Without getting into definitions and distinctions of these terms, we observe that much confusion and even doubt about the validity of science is arising. Such doubt can lead to radical choices, such as the one taken by the editors of *Basic and Applied Social Psychology*, who decided to ban *p*-values (null hypothesis significance testing) (Trafimow and Marks 2015). Misunderstanding or misuse of statistical inference is only one cause of the "reproducibility crisis" (Peng 2015), but to our community, it is an important one.

When the ASA Board decided to take up the challenge of developing a policy statement on *p*-values and statistical significance, it did so recognizing this was not a lightly taken step. The ASA has not previously taken positions on specific matters of statistical practice. The closest the association has come to this is a statement on the use of value-added models (VAM) for educational assessment (Morganstein and Wasserstein 2014) and a statement on risk-limiting post-election audits (American Statistical Association 2010). However, these were truly policy-related statements. The VAM statement addressed a key educational policy issue, acknowledging the complexity of the issues involved, citing limitations of VAMs as effective performance models, and urging that they be developed and interpreted with the involvement of statisticians. The statement on election auditing was also in response to a major but specific policy issue (close elections in 2008), and said that statistically based election audits should become a routine part of election processes.

By contrast, the Board envisioned that the ASA statement on *p*-values and statistical significance would shed light on an aspect of our field that is too often misunderstood and misused in the broader research community, and, in the process, provides the community a service. The intended audience would be researchers, practitioners, and science writers who are not primarily statisticians. Thus, this statement would be quite different from anything previously attempted.

The Board tasked Wasserstein with assembling a group of experts representing a wide variety of points of view. On behalf of the Board, he reached out to more than two dozen such people, all of whom said they would be happy to be involved. Several expressed doubt about whether agreement could be reached, but those who did said, in effect, that if there was going to be a discussion, they wanted to be involved.

Over the course of many months, group members discussed what format the statement should take, tried to more concretely visualize the audience for the statement, and began to find points of agreement. That turned out to be relatively easy to do, but it was just as easy to find points of intense disagreement.

The time came for the group to sit down together to hash out these points, and so in October 2015, 20 members of the group met at the ASA Office in Alexandria, Virginia. The 2-day meeting was facilitated by Regina Nuzzo, and by the end of the meeting, a good set of points around which the statement could be built was developed.

The next 3 months saw multiple drafts of the statement, reviewed by group members, by Board members (in a lengthy discussion at the November 2015 ASA Board meeting), and by members of the target audience. Finally, on January 29, 2016, the Executive Committee of the ASA approved the statement.

The statement development process was lengthier and more controversial than anticipated. For example, there was considerable discussion about how best to address the issue of multiple *potential* comparisons (Gelman and Loken 2014). We debated at some length the issues behind the words "a *p*-value near 0.05 taken by itself offers only weak evidence against the null

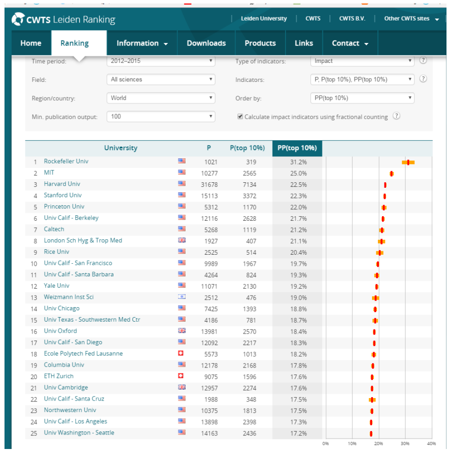
I have at least three concerns (1 and 2)

- Research practice is scientometrics and research evaluation
  - False claims, one-off studies, “garden-of-forking-paths”, “p-hacking” … (epistemic concerns)
- The use of significance tests (or pseudo tests) for decision making in research evaluations
  - Arbitrary, numerous fallacies, does not bring “certainty” or indications of “importance”, irresponsible surrogates for sound judgements about importance (ethical concerns)

We do address “uncertainty” …

Much more attention should be given to what exactly it is we think we quantify and what we can quantify

Intentions are fine, but the “solutions” are so far, unsatisfactory, both
- theoretically
- practically



CWTS Leiden Ranking | Leiden University | CWTS | CWTS B.V. | Other CWTS sites

Home | Ranking | Information | Downloads | Products | Links | Contact

Time period: 2012–2015 | Type of indicators: Impact

Field: All sciences | Indicators: P, P(top 10%), PP(top 10%)

Region/country: World | Order by: PP(top 10%)

Min. publication output: 100 | Calculate impact indicators using fractional counting

	University	P	P(top 10%)	PP(top 10%)
1	Rockefeller Univ	1021	319	31.2%
2	MIT	10277	2565	25.0%
3	Harvard Univ	31678	7134	22.5%
4	Stanford Univ	15113	3372	22.3%
5	Princeton Univ	5312	1170	22.0%
6	Univ Calif - Berkeley	12116	2628	21.7%
7	Caltech	5268	1119	21.2%
8	London Sch Hyg & Trop Med	1927	407	21.1%
9	Rice Univ	2525	514	20.4%
10	Univ Calif - San Francisco	9989	1967	19.7%
11	Univ Calif - Santa Barbara	4264	824	19.3%
12	Yale Univ	11071	2130	19.2%
13	Weizmann Inst Sci	2512	476	19.0%
14	Univ Chicago	7425	1393	18.8%
15	Univ Texas - Southwestern Med Ctr	4186	781	18.7%
16	Univ Oxford	13981	2570	18.4%
17	Univ Calif - San Diego	12092	2217	18.3%
18	Ecole Polytech Fed Lausanne	5573	1013	18.2%
19	Columbia Univ	12178	2168	17.8%
20	ETH Zurich	9075	1596	17.6%
21	Univ Cambridge	12957	2274	17.6%
22	Univ Calif - Santa Cruz	1988	348	17.5%
23	Northwestern Univ	10375	1813	17.5%
24	Univ Calif - Los Angeles	13898	2398	17.3%
25	Univ Washington - Seattle	14163	2436	17.2%

0% 10% 20% 30% 40%

COMMENT

SUSTAINABILITY Data needed to drive UN development goals p.432

CONSERVATION Economics and environmental catastrophe p.434

GEOLOGY Questions raised over proposed Anthropocene dates p.436

HISTORY Music inspired Newton to add more colours to the rainbow p.436

The Leiden Manifesto for research metrics

Use these ten principles to guide research evaluation, urge **Diana Hicks**, **Paul Wouters** and colleagues.

Data are increasingly used to govern science. Research evaluations that were once bespoke and performed by peers are now routine and reliant on metrics[1]. The problem is that evaluation is now led by the data rather than by judgement. Metrics have proliferated: usually well intentioned, not always well informed, often ill applied. We risk damaging the system with the very tools designed to improve it, as evaluation is increasingly implemented by organizations without knowledge of, or advice on, good practice and interpretation.

Before 2000, there was the Science Citation Index on CD-ROM from the Institute for Scientific Information (ISI), used by experts for specialist analyses. In 2002, Thomson Reuters launched an integrated web platform, making the Web of Science database widely accessible. Competing citation indices were created: Elsevier's Scopus (released in 2004) and Google Scholar (beta version released in 2004). Web-based tools to easily compare institutional research productivity and impact were introduced, such as InCites (using the Web of Science) and SciVal (using Scopus), as well as software to analyse individual citation profiles using Google Scholar (Publish or Perish, released in 2007).

In 2005, Jorge Hirsch, a physicist at the University of California, San Diego, proposed the *h*-index, popularizing citation counting for individual researchers. Interest in the journal impact factor grew steadily after 1995 (see 'Impact-factor obsession').

Lately, metrics related to social usage

ILLUSTRATION BY DAVID PARKINS

23 APRIL 2015 | VOL 520 | NATURE | 429

8. Avoid misplaced concreteness and false precision

…

picture. If uncertainty and error can be quantified, for instance using error bars, this information should accompany published indicator values. If this is not possible, indicator producers should at least avoid false precision. For example, the journal impact factor is published to three decimal places to avoid ties. However, given the conceptual ambiguity and random variability of citation counts, it makes no sense to distinguish between journals on the basis of very small impact factor differences. Avoid false precision: only one decimal is warranted.

I have at least three concerns (3)

- The meaning of citations
  - Normalization for publication type, publication and “field” bring publications on an equal footing … this is what we recommend
  - Strong assumption = normalized citations have the same meaning across knowledge production models

  - … but if we have a blatantly flawed knowledge production model that evidentiary produces numerous claims that turn out be unsupported or false, that seem to invoke questionable research practices,
  - why then should such publications count equally to comparable empirical publications based on more sound knowledge production models?

  - Including social science publications in impact analyses is not only a question of their coverage

  - **Citations are not equal and our current counting regime reward flawed research!**

Some solutions

- We need to be more humble and acknowledge there are things which we cannot get such as certainty
- We need to embrace variation, it is all around
- Social science settings are noisy and rife with non-random biases
- We need to address our cognitive fallacies
- We need to estimate rather than arbitrarily decide between all or nothing
- We need theory
- We need to have a broader statistical toolbox
- We need to understand that nothing certain comes out of one study
- We need strong will to stand up against demands for surrogate numbers
- And our journal editors should demand much more from the manuscripts they accept!

Thank you very much

"No scientific worker has a fixed level of significance at which from year to year, and in all circumstances, he rejects hypotheses; he rather gives his mind to each particular case in the light of his evidence and his ideas."

—Sir Ronald A. Fisher (1956: 42)

Calling Bullshit:
Data Reasoning for the Digital Age

Logistics

Course: INFO 198 / BIOL 106B. University of Washington
To be offered: Autumn Quarter 2017
Credit: 3 credits, graded
Enrollment: 180 students
Instructors: Carl T. Bergstrom and Jevin West
Synopsis: Our world is saturated with bullshit. Learn to detect and defuse it.

Learning Objectives

Our learning objectives are straightforward. After taking the course, you should be able to:

- Remain vigilant for bullshit contaminating your information diet.
- Recognize said bullshit whenever and wherever you encounter it.
- Figure out for yourself precisely *why* a particular bit of bullshit is bullshit.
- Provide a statistician or fellow scientist with a technical explanation of why a claim is bullshit.
- Provide your crystals-and-homeopathy aunt or casually racist uncle with an accessible and persuasive explanation of why a claim is bullshit.

We will be astonished if these skills do not turn out to be among the most useful and most broadly applicable of those that you acquire during the course of your college education.

Lectures

1. Introduction to bullshit
2. Spotting bullshit
3. The natural ecology of bullshit
4. Causality
5. Statistical traps
6. Visualization
7. Big data
8. Publication bias
9. Predatory publishing and scientific misconduct
10. The ethics of calling bullshit.
11. Fake news
12. Refuting bullshit