From Frequentist Problems Towards Bayesian Solutions

Part 1: Replication crisis in Psychology
(and what does Statistics have to do with it)

Jorge N. Tendeiro

10 August 2019

Slides at: https://rebrand.ly/Nagoya2019-Part1
GitHub: https://github.com/jorgetendeiro/Nagoya-Workshop-10-Aug-2019

Today

- Fraud
- Questionable research practices
- But why
- Ireproducibility
- Didnt we see this coming
- P values
- Confidence intervals
- Publication policies
- What do statistical associations advice
- What to avoid
- Bayesian statistics
- References

Fraud

Definition

Fraud = scientifc misconduct.

- Falsifying or fabricating data.
- This is intentional, not accidental.
- Puts all science under a bad light.
- Markedly different from QRPs (next).

Notable examples

- Diederik Stapel, social psychologist. Suspended in 2011. Fabricating and manipulating data.
- Marc Hauser, psychologist at Harvard. Resigned in 2011. Scientific misconduct.
- Jens Förster, social psychologist. Resigned in 2017. Data tampering.

Today we don't talk about fraud *explicitly*.

We talk about something much harder to identify and erradicate:

Questionable research practices (QRPs).

Questionable research practices

QRPs

Term coined by John, Loewenstein, & Prelec (2012).
See also Simmons, Nelson, & Simonsohn (2011).

- Not necessarily fraud.
- Includes the (ab)use of actually *acceptable* research practices.
- Problem with QRPs:
  - Introduce bias (typically, in *favor* of the researcher's intentions…).
  - Inflated power at the cost of inflated Type I error probability ($\gg 5\%$).
  - Results not replicable.

Example of QRPs

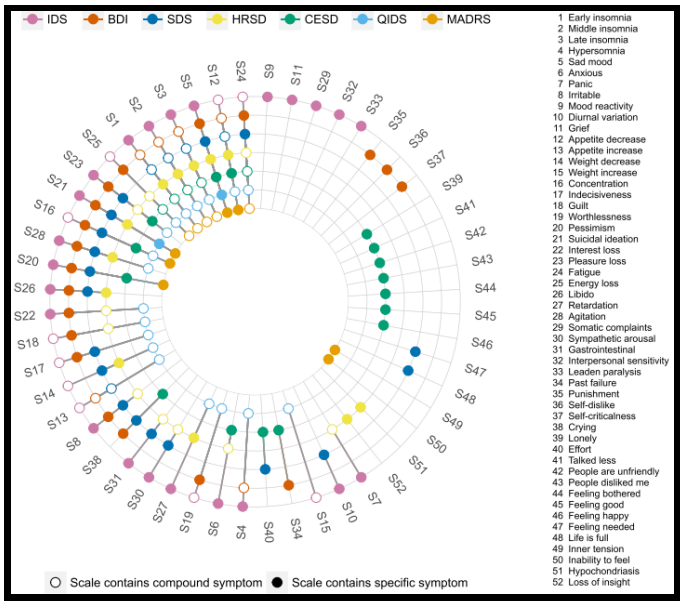
(John et al., 2012; Schimmack, 2015).

- Omit some DVs.
- Omit some conditions.
- Peeking: Sequential testing — Look and decide:
  - $p > .05$: Collect more.
  - $p < .05$: Stop.
- Only report $p < .05$ results.
- p-hacking: E.g.,
  - Exclusion of outliers depending on whether $p < .05$.
  - $p = .054 \longrightarrow p = .05$.
- HARKing (Kerr, 1998): Convert exploratory results into research questions.
- ...

Researcher's degrees of freedom

- Researchers have a multitude of decisions to make (experiment design, data collection, analyses performed); Wicherts et al. (2016), Steegen, Tuerlinckx, Gelman, & Vanpaemel (2016).
- It is very possible to manipulate results *in favor* of one's interests.
- This is now known as *researcher's degrees of freedom* (Simmons et al., 2011).
- Consequence: Inflated false positive findings (Ioannidis, 2005).

Fried (2017)



- The 7 most common depression scales contain 52 symptoms.
- That's 7 different sum scores.
- Yet, all are interpreted as `level of depression'.

Turning exploratory into confirmatory analysis

From Bem (2004):

> "(…) [L]et us (…) become intimately familiar with (…) the data. Examine them from every angle. Analyze the sexes separately. Make up new composite indices. If a datum suggests a new hypothesis, try to find further evidence for it elsewhere in the data. If you see dim traces of interesting patterns, try to reorganize the data to bring them into bolder relief. If there are participants you don't like, or trials, observers, or interviewers who gave you anomalous results, drop them (temporarily). Go on a fishing expedition for something– anything– interesting."

This is not OK *unless* the exploration is explicity stated.

Daryl Bem is the author of the famous 2011 precognition paper (data used in Part 2 of today's workshop).

A now famous example...

Prof. Brian Wansink at Cornell University.

His description of the efforts of a visiting Ph.D student:

> I gave her a data set of a self-funded, failed study which had null results (…). I said, "This cost us a lot of time and our own money to collect. There's got to be something here we can salvage because it's a cool (rich & unique) data set." I had three ideas for potential Plan B, C, & D directions (since Plan A had failed). I told her what the analyses should be and what the tables should look like. I then asked her if she wanted to do them.

> Every day she came back with puzzling new results, and every day we would scratch our heads, ask "Why," and come up with another way to reanalyze the data with yet another set of plausible hypotheses. Eventually we started discovering solutions that held up regardless of how we pressure-tested them. I outlined the first paper, and she wrote it up (…). This happened with a second paper, and then a third paper (which was one that was based on her own discovery while digging through the data).

This isn't being *creative* or *thinking outside the box*.

This is QRPing.

What happened to Wansink?

- He was severely criticized, his work was scrutinized (e.g., van der Zee, Anaya, & Brown, 2017).
- Over 100 (!!) errors in a set of four papers…
- Has now 40 (!!) publications retracted (as of July 2019).
- After a year-long internal investigation, he was forced to resign.

Is it really *that* bad?...

Yes.

- Martinson, Anderson, & Vries (2005): "Scientists behaving badly".
- Fanelli (2009): Meta-analysis shows evidence of science misconduct.
- John et al. (2012): Evidence for QRPs in psychology.
- Mobley, Linder, Braeuer, Ellis, & Zwelling (2013): Reported evidence of pressure to *find* significant results.
- Agnoli, Wicherts, Veldkamp, Albiero, & Cubelli (2017): Evidence of QRPs, now in Italy.
- Fraser, Parker, Nakagawa, Barnett, & Fidler (2018): In other fields of science.

Interestingly, science misconduct has been a longtime concern (see Babbage, 1830).

And for the sake of balance:
There are also some voices *against* this description of the current state of affairs (e.g., Fiedler & Schwarz, 2016).

Preregistration eliminates QRPs, right?...

Well, maybe not (yet).

Here's an interesting preprint (from July 2019!) from a Japanese research group (Kyushu University):

> Ikeda, A., Xu, H., Fuji, N., Zhu, S., & Yamada, Y. (2019). *Questionable research practices following pre-registration* [Preprint]. doi: 10.31234/osf.io/b8pw9

But *why*?...

Why are QRPs so prevalent?

It is strongly related to incentives (Nosek, Spies, & Motyl, 2012; Schönbrodt, 2015).

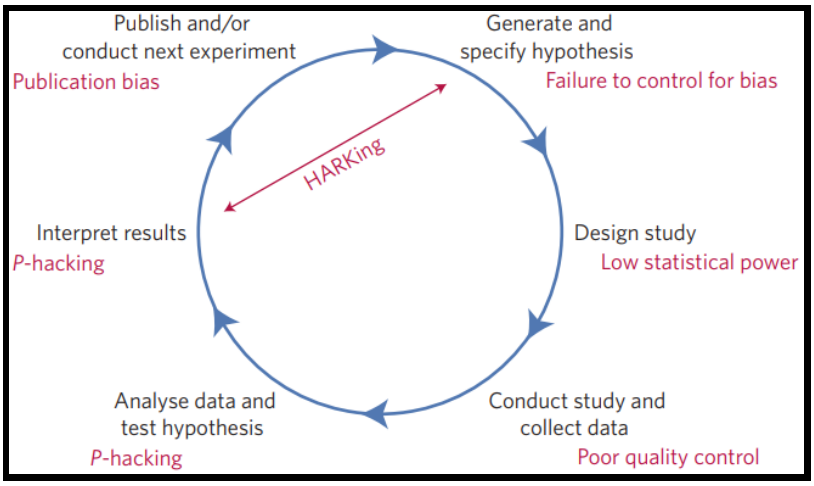
- "Publish or perish":
  Publish a lot, at highly prestigious journals.
  - Journals only publish a fraction of all manuscripts.
  - Journals don't like publishing null findings…
- Get tenured.
- Get research grant.
- Fame (prizes, press coverage, …).
- …

But, very importantly, it also happens in spite of the researcher's best intentions.

- Deficient statistics education (yes, statisticians need to acknowledge this!…).
- Perpetuating traditions in the field.

(I)reproducibility

Threats to reproducible science



Munafò et al. (2017)

- Hypothetico-deductive model of the scientific method.
- In red: Potential threats to this model.

Lack of replications

Until very recently (Makel, Plucker, & Hegarty, 2012).

aps ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

Replications in Psychology Research: How Often Do They Really Occur?

Perspectives on Psychological Science
7(6) 537–542

DOI: 10.1177/1745691612460688
http://pps.sagepub.com
SAGE

Matthew C. Makel[1], Jonathan A. Plucker[2], and Boyd Hegarty[3]
[1]Duke University, [2]University of Connecticut, and [3]University of New Hampshire

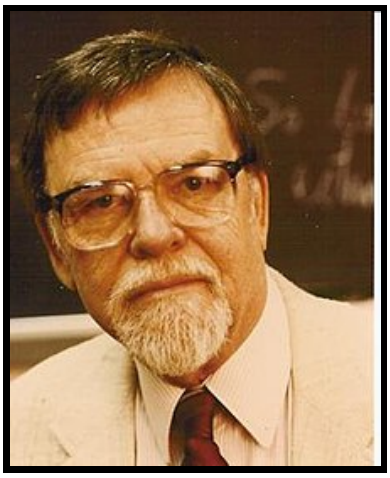
- Very low rate of replications in Psychology (estimated ~1%).
- Until 2012, majority of replications were actually successful!!
- However, in most cases both the original and replication studies shared authorship…
- Conflict of interest?…

Famous replication failures

- The Marshmallow Test (Watts, Duncan, & Quan, 2018)
- Ego depletion (Friese, Loschelder, Gieseler, Frankenbach, & Inzlicht, 2019; Hagger et al., 2016; Vadillo, Gold, & Osman, 2018)
- Power posing (Ranehill et al., 2015)
- The Stanford Prison Experiment (Griggs, 2014; Reicher & Haslam, 2006)
- The facial feedback hypothesis (Wagenmakers et al., 2016)
- Newborn babies' imitation (Oostenbroek et al., 2016)
- The blocking effect (Maes et al., 2016)
- The stereotype threat (Flore, Mulder, & Wicherts, 2019)
- The facial expression (Gendron, Crivelli, & Barrett, 2018)
- ESP, of course! (Galak, LeBoeuf, Nelson, & Simmons, 2012; Ritchie, Wiseman, & French, 2012)
- The Mozart Effect (McKelvie & Low, 2002; Steele, Bass, & Crook, 1999)
- ...

Didn’t we see this coming?

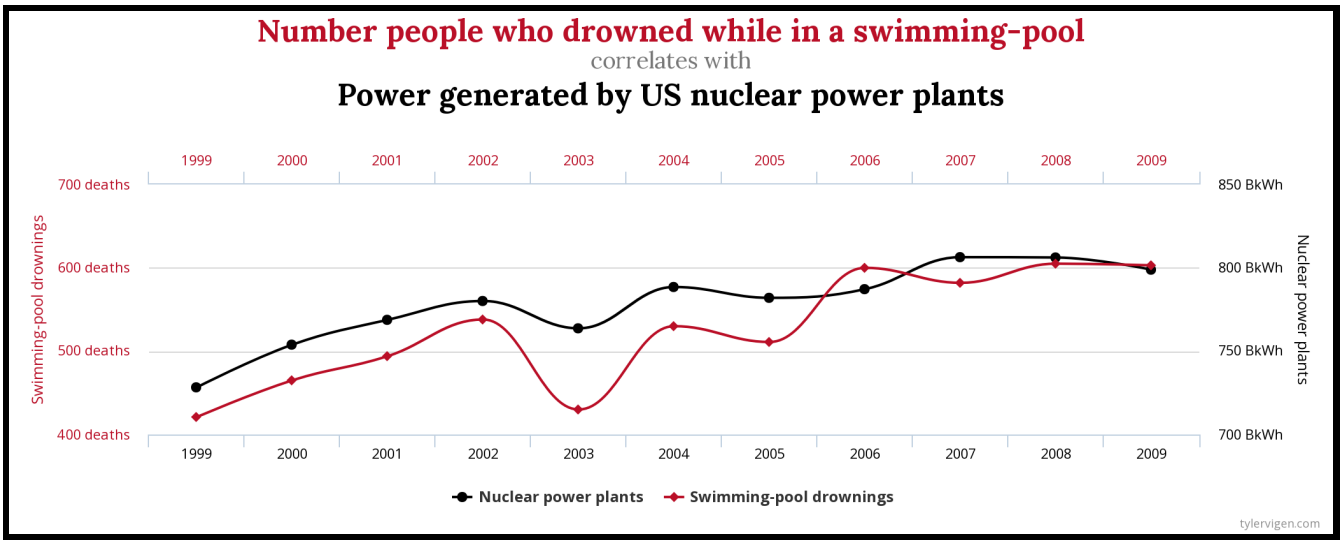
Meehl (1967)



How poorly we build theory (see Gelman):

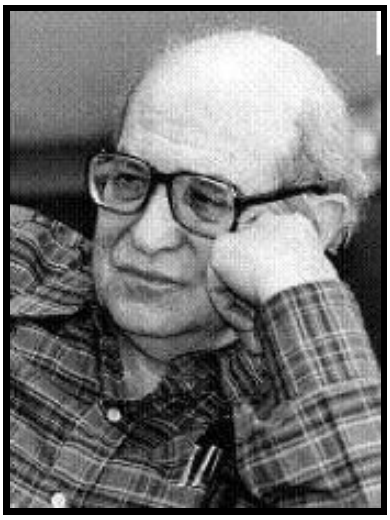
> "It is not unusual that (…) this *ad hoc* challenging of auxiliary hypotheses is repeated in the course of a series of related experiments, in which the auxiliary hypothesis involved in Experiment 1 (…) becomes the focus of interest in Experiment 2, which in turn utilizes further plausible but easily challenged auxiliary hypotheses, and so forth. In this fashion a zealous and clever investigator can slowly wend his way through (…) a long series of related experiments (…) without ever once refuting or corroborating so much as a single strand of the network."

Say what?...



http://www.tylervigen.com/spurious-correlations

Cohen (1962)

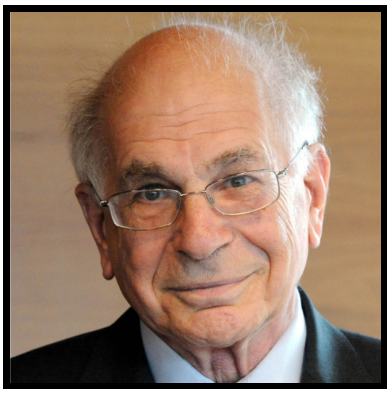


Low-powered experiments:

> "(…) It was found that the average power (probability of rejecting false null hypotheses) over the 70 research studies was .18 for small effects, .48 for medium effects, and .83 for large effects. These values are deemed to be far too small."
>
> "(…) it is recommended that investigators use larger sample sizes than they customarily do."

Kahneman (2012)

See here.

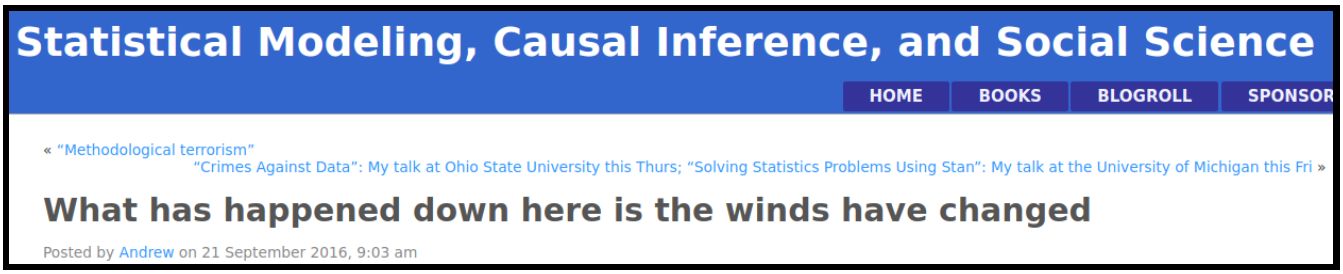


Nobel prize winner, 2002.

About priming effects (but quite general remarks…):

> "The storm of doubts is fed by (…) the recent exposure of fraudulent researchers, general concerns with replicability (…), multiple reported failures to replicate salient results (…), and the growing belief in the existence of a pervasive file drawer problem (…)."
>
> "My reason for writing this letter is that I see a train wreck looming."
>
> "I believe that you should collectively do something about this mess."

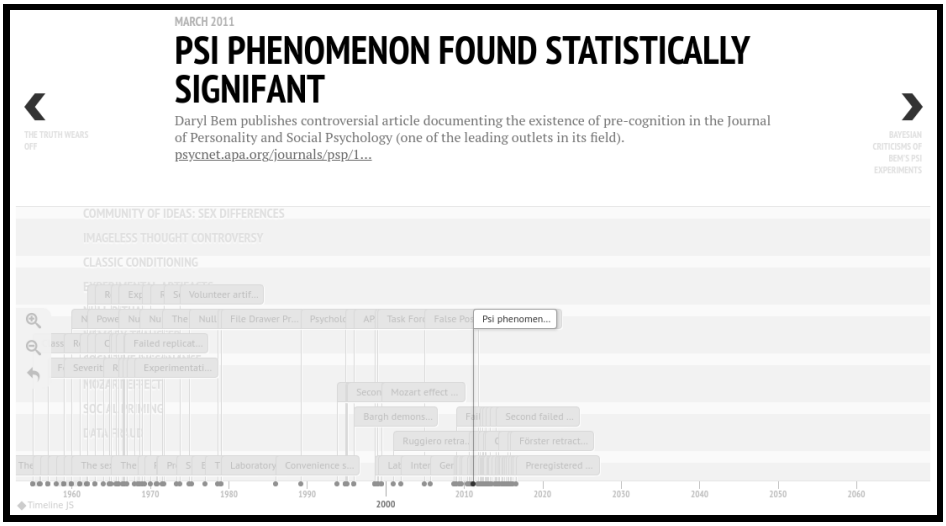
Timeline of a train wreck



- Gelman blogged about an impressive timeline about the replication crisis.
- The whole blog post is worth reading for many reasons, including Gelman's criticism over criticism! (versus Susan Fiske's position).

Timeline of a train wreck

See also this impressive dynamic plot:
https://psyborgs.github.io/projects/replication-in-psychology/



p-values

Definition

> Probability of an effect at least as extreme as the one we observed, *given that $\mathcal{H}_0$ is true*.

$$\boxed{p\text{-value} = P\left(X_{\text{obs}} \text{ or more extreme} | \mathcal{H}_0\right)}$$

The definition is simple enough, right?...

Test yourself

Consider the following statement (Falk & Greenbaum, 1995; Gigerenzer, Krauss, & Vitouch, 2004; Haller & Kraus, 2002; Oakes, 1986):

> *Suppose you have a treatment that you suspect may alter performance on a certain task. You compare the means of your control and experimental groups (say, 20 subjects in each sample). Furthermore, suppose you use a simple independent means t-test and your result is significant (* $t = 2.7$, $df = 18$, $p = .01$*). Please mark each of the statements below as "true" or "false."* False *means that the statement does not follow logically from the above premises. Also note that several or none of the statements may be correct.*

Test yourself

(1) You have absolutely disproved the null hypothesis (i.e., there is no difference between the population means).	☐ True	False ☐
(2) You have found the probability of the null hypothesis being true.	☐ True	False ☐
(3) You have absolutely proved your experimental hypothesis (that there is a difference between the population means).	☐ True	False ☐
(4) You can deduce the probability of the experimental hypothesis being true.	☐ True	False ☐
(5) You know, if you decide to reject the null hypothesis, the probability that you are making the wrong decision.	☐ True	False ☐
(6) You have a reliable experimental finding in the sense that if, hypothetically, the experiment were repeated a great number of times, you would obtain a significant result on 99% of occasions.	☐ True	False ☐

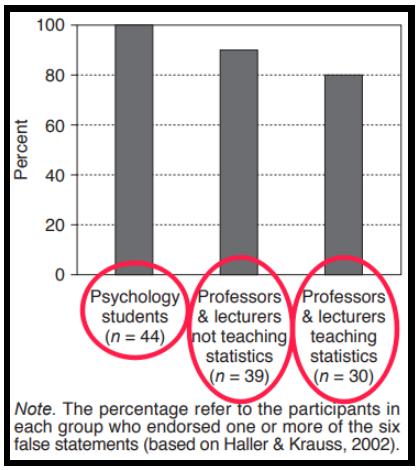
Try it!: rebrand.ly/pvalue

Results

All statements are incorrect.

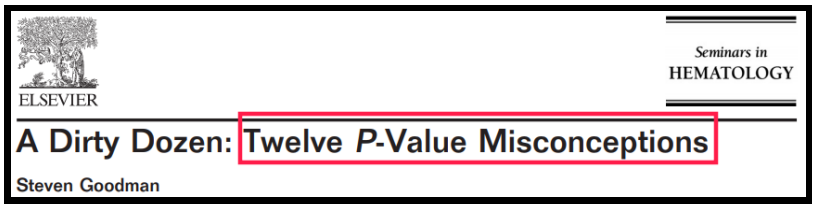
Results

But how did students and teachers perceive these statements?



This was in 2004. But things did not improve since…

Goodman (2008)



ELSEVIER

Seminars in HEMATOLOGY

A Dirty Dozen: Twelve *P*-Value Misconceptions

Steven Goodman

Table 1 Twelve *P*-Value Misconceptions

1	*If P = .05, the null hypothesis has only a 5% chance of being true.*
2	*A nonsignificant difference (eg, P ≥.05) means there is no difference between groups.*
3	*A statistically significant finding is clinically important.*
4	*Studies with P values on opposite sides of .05 are conflicting.*
5	*Studies with the same P value provide the same evidence against the null hypothesis.*
6	*P = .05 means that we have observed data that would occur only 5% of the time under the null hypothesis.*
7	*P = .05 and P ≤.05 mean the same thing.*
8	*P values are properly written as inequalities (eg, "P ≤.02" when P = .015)*
9	*P = .05 means that if you reject the null hypothesis, the probability of a type I error is only 5%.*
10	*With a P = .05 threshold for significance, the chance of a type I error will be 5%.*
11	*You should use a one-sided P value when you don't care about a result in one direction, or a difference in that direction is impossible.*
12	*A scientific conclusion or treatment policy should be based on whether or not the P value is significant.*

Greenland et al. (2016)

Eur J Epidemiol (2016) 31:337–350
DOI 10.1007/s10654-016-0149-3

CrossMark

ESSAY

Statistical tests, *P* values, confidence intervals, and power: a guide to misinterpretations

Sander Greenland[1] · Stephen J. Senn[2] · Kenneth J. Rothman[3] · John B. Carlin[4] · Charles Poole[5] · Steven N. Goodman[6] · Douglas G. Altman[7]

This paper expands Goodman (2008) and elaborates on 25 misinterpretations.

The American Statistician (2019)

Special issue with 43 (!!) papers (Wasserstein, Schirm, & Lazar, 2019).

Moving to a world beyond "$p < .05$"

Confidence intervals

A better alternative?

- Confidence intervals (CIs) have been often advocated as the best inferential alternative to NHST.
- Recall, for example the Wilkinson Task Force (Wilkinson & Task Force on Statistical Inference, 1999):

> "(...) it is hard to imagine a situation in which a dichotomous accept–reject decision is better than reporting an actual p value or, better still, a confidence interval."

- But, are CIs really a better alternative?

Definition

See, for instance, Hoekstra, Morey, Rouder, & Wagenmakers (2014).

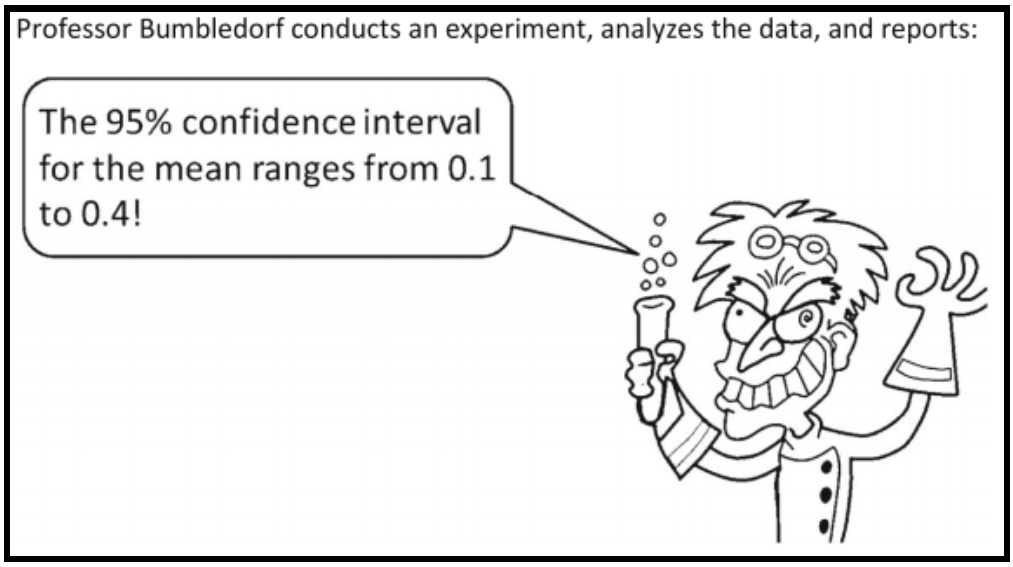
> A (say) 95% CI is a numerical interval found through a procedure that, if repeated across a series of hypothetical data, leads to an interval covering the true parameter 95% of the times.

- A CI indicates a property of the performance of the procedure used to compute it:
  How is the procedure expected to perform in the long run?
- A CI for a parameter is constructed around the parameter's estimate.
- However, a CI does not (really not!) directly indicate a property of the parameter being estimated!

Confused?
So is the vast majority of psychologists…

Test yourself

From Hoekstra et al. (2014), mimicking the p value study by Gigerenzer et al. (2004).



Test yourself

Please mark each of the statements below as "true" or "false". False means that the statement does not follow logically from Bumbledorf's result. Also note that all, several, or none of the statements may be correct:

1. The probability that the true mean is greater than 0 is at least 95%. ☐True ☐False
2. The probability that the true mean equals 0 is smaller than 5%. ☐True ☐False
3. The "null hypothesis" that the true mean equals 0 is likely to be incorrect. ☐True ☐False
4. There is a 95% probability that the true mean lies between 0.1 and 0.4. ☐True ☐False
5. We can be 95% confident that the true mean lies between 0.1 and 0.4. ☐True ☐False
6. If we were to repeat the experiment over and over, then 95% of the time the true mean falls between 0.1 and 0.4. ☐True ☐False

Please indicate the level of your statistical experience from 1 (no stats courses taken, no practical experience) to 10 (teaching statistics at a university): _____

Try it!: rebrand.ly/confint

Results

All statements are incorrect.

Results

But how did students and teachers perceive these statements?

Table 1 Percentages of students and teachers endorsing an item

Statement	First Years ($n = 442$)	Master Students ($n = 34$)	Researchers ($n = 118$)
The probability that the true mean is greater than 0 is at least 95 %	51 %	32 %	38 %
The probability that the true mean equals 0 is smaller than 5 %	55 %	44 %	47 %
The "null hypothesis" that the true mean equals 0 is likely to be incorrect	73 %	68 %	86 %
There is a 95 % probability that the true mean lies between 0.1 and 0.4	58 %	50 %	59 %
We can be 95 % confident that the true mean lies between 0.1 and 0.4	49 %	50 %	55 %
If we were to repeat the experiment over and over, then 95 % of the time the true mean falls between 0.1 and 0.4	66 %	79 %	58 %

What would be correct, then?...

> "If we were to repeat the experiment over and over, then 95% of the time the confidence intervals contain the true mean."

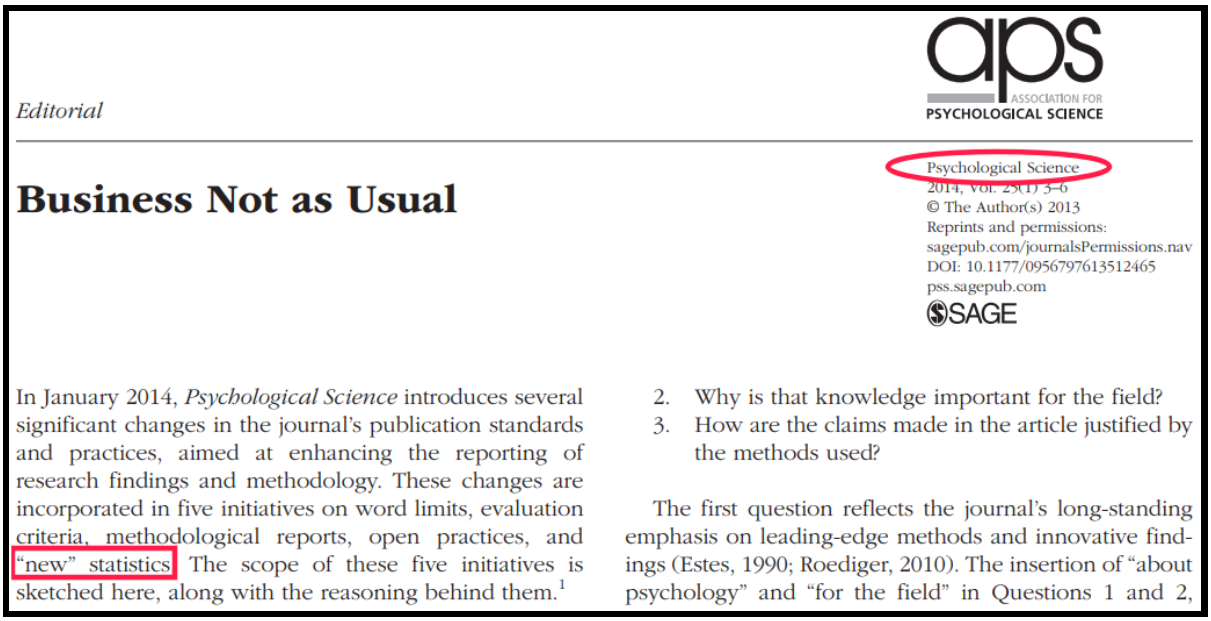
How informative is this?!

Mental note:
Remember this when interpreting Bayesian credible intervals in part 2 of today's workshop!

For completeness, not everyone agrees with the Hoekstra study (García-Pérez & Alcalá-Quintana, 2016; Miller & Ulrich, 2016; see also a reply by Morey, Hoekstra, Rouder, & Wagenmakers, 2016).

Publication policies

Psychological Science (Eich, 2014)



Editorial

aps ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

Business Not as Usual

Psychological Science
2014, Vol. 25(1) 3–6

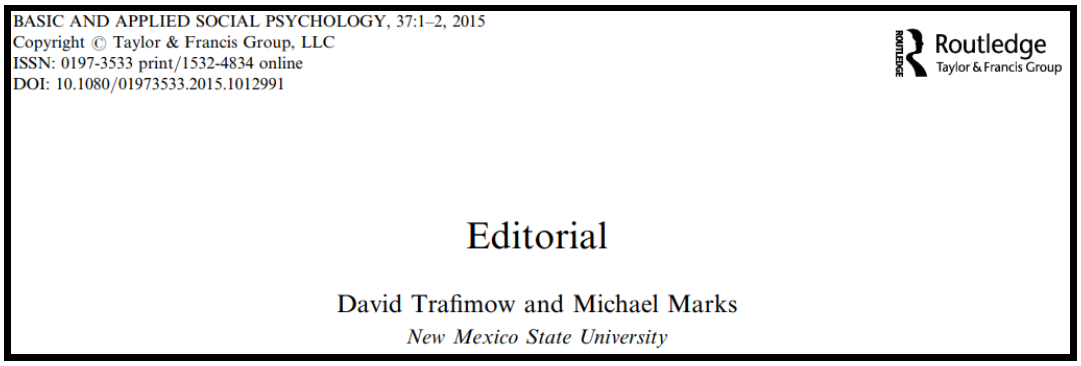
DOI: 10.1177/0956797613512465
pss.sagepub.com
SAGE

In January 2014, *Psychological Science* introduces several significant changes in the journal's publication standards and practices, aimed at enhancing the reporting of research findings and methodology. These changes are incorporated in five initiatives on word limits, evaluation criteria, methodological reports, open practices, and "new" statistics. The scope of these five initiatives is sketched here, along with the reasoning behind them.[1]

2. Why is that knowledge important for the field?
3. How are the claims made in the article justified by the methods used?

The first question reflects the journal's long-standing emphasis on leading-edge methods and innovative findings (Estes, 1990; Roediger, 2010). The insertion of "about psychology" and "for the field" in Questions 1 and 2,

Basic and Applied Social Psychology



BASIC AND APPLIED SOCIAL PSYCHOLOGY, 37:1–2, 2015

ISSN: 0197-3533 print/1532-4834 online
DOI: 10.1080/01973533.2015.1012991

Routledge
Taylor & Francis Group

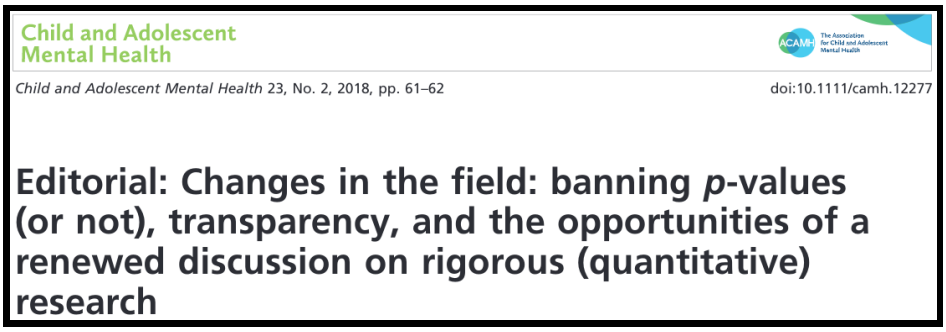
Editorial

David Trafimow and Michael Marks
New Mexico State University

> "The Basic and Applied Social Psychology (BASP) (…) emphasized that the null hypothesis significance testing procedure (NHSTP) is invalid (…). From now on, BASP is banning the NHSTP."

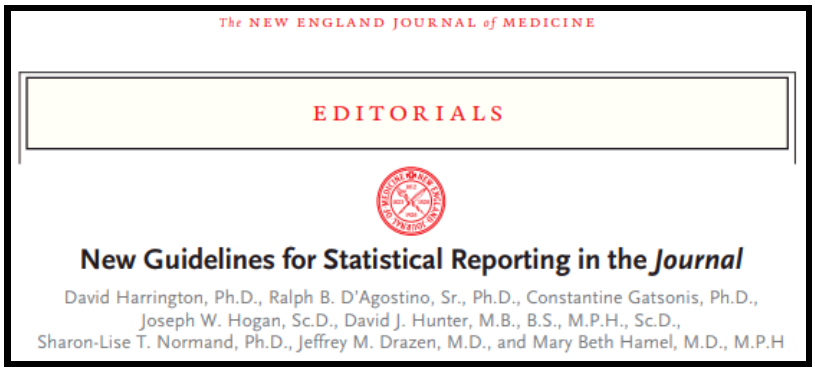
Did it actually *work*? For a reflection, see Fricker, Burke, Han, & Woodall (2019).

Child Adolescent Mental Health



> (…) I will encourage authors to provide replication syntax and data through public repositories. Moreover, I will encourage the journal to focus on a manuscript's research design and the author's justification thereof, rather than the results, with the aim of ensuring that transparent studies that explore a research question with equipoise, will be published.

The New England Journal of Medicine


The NEW ENGLAND JOURNAL of MEDICINE

EDITORIALS

New Guidelines for Statistical Reporting in the *Journal*

David Harrington, Ph.D., Ralph B. D'Agostino, Sr., Ph.D., Constantine Gatsonis, Ph.D., Joseph W. Hogan, Sc.D., David J. Hunter, M.B., B.S., M.P.H., Sc.D., Sharon-Lise T. Normand, Ph.D., Jeffrey M. Drazen, M.D., and Mary Beth Hamel, M.D., M.P.H

Editorial (Harrington et al., 2019).

> "(…) a requirement to replace p values with estimates of effects or association and 95% confidence intervals"

What do statistical associations advice?

Wilkinson Task Force 1999

Among many many, advices,

- Do not focus on p values.
- Report effect sizes.
- Report power analyses.
- Check model assumptions.

> “Novice researchers err either by overgeneralizing their results or, equally unfortunately, by overparticularizing.”

ASA 2016 (Wasserstein & Lazar, 2016)


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

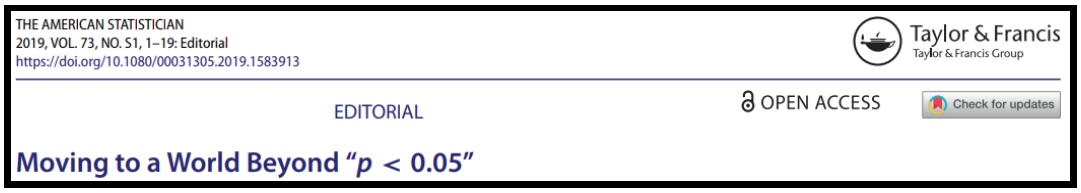
EDITORIAL

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

Six principles:

1. p-values can indicate how incompatible the data are with a specified statistical model.
2. p-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.
3. Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.
4. Proper inference requires full reporting and transparency.
5. A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.
6. By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.

ASA 2019 (Wasserstein et al., 2019)


THE AMERICAN STATISTICIAN
2019, VOL. 73, NO. S1, 1–19: Editorial
https://doi.org/10.1080/00031305.2019.1583913
Taylor & Francis
Taylor & Francis Group
EDITORIAL
OPEN ACCESS
Check for updates

Moving to a World Beyond "$p < 0.05$"

This is an editorial of a special issue consisting of 43 (!!) papers.

Main ideas:

- "Don't" is not enough – Some *what to do* advices are provided.
- However… Don't say "statistically significant" – Just **don't**.

> "(…) it is time to stop using the term "statistically significant" entirely. Nor should variants such as "significantly different," "$p < 0.05$," and "nonsignificant" survive, whether expressed in words, by asterisks in a table, or in some other way."

But:

> "Despite the limitations of p-values (…), however, we are not recommending that the calculation and use of continuous p-values be discontinued. Where p-values are used, they should be reported as continuous quantities (e.g., $p = 0.08$). They should also be described in language stating what the value means in the scientific context."

- There is no unique "do":

> "What you will NOT find in this issue is one solution that majestically replaces the outsized role that statistical significance has come to play."

- Accept uncertainty (I cannot stress this enough!).
  Be thoughtful, open, and modest.
- Editorial, educational, and other institutional practices will have to change.
  This includes: Journals, funding agencies, education, career system.
- Value replicability, open materials and data, and reliable practices (which all take time) over "publish or perish".

ASA 2019: Also advocate Bayesian statistics

Ruberg, S., Harrell, F., Gamalo-Siebers, M., LaVange, L., Lee J., Price K., and Peck C., Inference and Decision-Making for 21st Century Drug Development and Approval

1. Apply Bayesian paradigm as a framework for improving statistical inference and regulatory decision making by using probability assertions about the magnitude of a treatment effect.
2. Incorporate prior data and available information formally into the analysis of the confirmatory trials.
3. Justify and pre-specify how priors are derived and perform sensitivity analysis for a better understanding of the impact of the choice of prior distribution.
4. Employ quantitative utility functions to reflect key considerations from all stakeholders for optimal decisions via a probability-based evaluation of the treatment effects.
5. Intensify training in Bayesian approaches, particularly for decision makers and clinical trialists (e.g., physician scientists in FDA, industry and academia).

What to avoid

Bullying

- Debates in blogs, Twitter, and journals can be fierce.
- Criticism should be part of science, of course.
- It's not bullying to criticize, of course, in particular, with grounded reasons (vide Wansink).
- But sometimes criticism gets too carried away, IMHO.

NYT, 2017

(Interestingly: A recent comeback in Psychological Science.)

Self-appointed police

Most likely, each of us has some skeleton's in their scientific closets.

We've all fallen prey to one or more of the problems mentioned today.

Full disclosure:

> I have too!!

So:

> *No one is better than anyone.*

Or in the words of Brian Nosek (as quoted here):

> "We're not here to *be right*. We're here to *get it right*."

No time today for...

No time today for...

- Replications projects
- Registered reports
- Preregistrations
- Education
- ...

(But we can talk about it too!...)

Today we focus on *statistics*.

Bayesian statistics

Alternative approach to statistical inference

After the break:

Gentle introduction to Bayesian statistics.

References

Agnoli, F., Wicherts, J. M., Veldkamp, C. L. S., Albiero, P., & Cubelli, R. (2017). Questionable research practices among italian research psychologists. *PLOS ONE*, *12*(3), e0172792. doi: 10.1371/journal.pone.0172792

Babbage, C. (1830). *Reflections on the Decline of Science in England: And on Some of Its Causes*. Retrieved from http://www.gutenberg.org/files/1216/1216-h/1216-h.htm

Bem, D. J. (2004). Writing the empirical journal article. In *The compleat academic: A career guide, 2nd ed* (pp. 185–219). Washington, DC, US: American Psychological Association.

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *The Journal of Abnormal and Social Psychology*, *65*(3), 145–153. doi: 10.1037/h0045186

Cuddy, A. J. C., Schultz, S. J., & Fosse, N. E. (2018). P-Curving a More Comprehensive Body of Research on Postural Feedback Reveals Clear Evidential Value for Power-Posing Effects: Reply to Simmons and Simonsohn (2017) - Amy J. C. Cuddy, S. Jack Schultz, Nathan E. Fosse, 2018. *Psychological Science*. doi: 10.1177/0956797617746749

Eich, E. (2014). Business Not as Usual. *Psychological Science*, *25*(1), 3–6. doi: 10.1177/0956797613512465

Falk, R., & Greenbaum, C. (1995). Significance Tests Die Hard - the Amazing Persistence of a Probabilistic Misconception. *Theory & Psychology*, *5*(1), 75–98. doi: 10.1177/0959354395051004

Fanelli, D. (2009). How Many Scientists Fabricate and Falsify Research? A Systematic Review and Meta-Analysis of Survey Data. *PLOS ONE*, *4*(5), e5738. doi: 10.1371/journal.pone.0005738

Fiedler, K., & Schwarz, N. (2016). Questionable Research Practices Revisited. *Social Psychological and Personality Science*, *7*(1), 45–52. doi: 10.1177/1948550615612150

Flore, P. C., Mulder, J., & Wicherts, J. M. (2019). The influence of gender stereotype threat on mathematics test scores of Dutch high school students: A registered report. *Comprehensive Results in Social Psychology*, 1–35. doi: 10.1080/23743603.2018.1559647

Fraser, H., Parker, T., Nakagawa, S., Barnett, A., & Fidler, F. (2018). Questionable research practices in ecology and evolution. *PLOS ONE*, *13*(7), e0200303. doi: 10.1371/journal.pone.0200303

Fricker, R. D., Burke, K., Han, X., & Woodall, W. H. (2019). Assessing the Statistical Analyses Used in *Basic and Applied Social Psychology* After Their *p* -Value Ban. *The American Statistician*, *73*(sup1), 374–384. doi: 10.1080/00031305.2018.1537892

Fried, E. I. (2017). The 52 symptoms of major depression: Lack of content overlap among seven common depression scales. *Journal of Affective Disorders*, *208*, 191–197. doi: 10.1016/j.jad.2016.10.019

Friese, M., Loschelder, D. D., Gieseler, K., Frankenbach, J., & Inzlicht, M. (2019). Is Ego Depletion Real? An Analysis of Arguments. *Personality and Social Psychology Review*, *23*(2), 107–131. doi: 10.1177/1088868318762183

Galak, J., LeBoeuf, R. A., Nelson, L. D., & Simmons, J. P. (2012). *Correcting the Past: Failures to Replicate Psi* (SSRN Scholarly Paper No. ID 2001721). Retrieved from Social Science Research Network website: https://papers.ssrn.com/abstract=2001721

García-Pérez, M. A., & Alcalá-Quintana, R. (2016). The Interpretation of Scholars' Interpretations of Confidence Intervals: Criticism, Replication, and Extension of Hoekstra et al. (2014). *Frontiers in Psychology*, *7*. doi: 10.3389/fpsyg.2016.01042

Gendron, M., Crivelli, C., & Barrett, L. F. (2018). Universality Reconsidered: Diversity in Making Meaning of Facial Expressions. *Current Directions in Psychological Science*, *27*(4), 211–219. doi: 10.1177/0963721417746794