Best practices for reporting scientific results

Navigating the minefield around P-values and significance

Stefanie Muff

Open Science course, Finse, November 2022

The ongoing controversy around p-values

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.

(Wasserstein and Lazar 2016)

Lots of publications in the past decades...

STATISTICAL ERRORS

P values, the 'gold standard' of statistical validity, are not as reliable as many scientists assume.

BY REGINA NUZZO

COMMENT · 20 MARCH 2019

Scientists rise up against statistical significance

Valentin Amrhein, Sander Greenland, Blake McShane and more than 800 signatories call for an end to hyped claims and the dismissal of possibly crucial effects.

Valentin Amrhein, Sander Greenland & Blake McShane

A Dirty Dozen: Twelve *P*-Value Misconceptions

Steven Goodman

The *P* value is a measure of statistical evidence that appears in virtually all medical research papers. Its interpretation is made extraordinarily difficult because it is not part of any formal system of statistical inference. As a result, the *P* value's inferential meaning is widely and often wildly misconstrued, a fact that has been pointed out in innumerable papers and books appearing since at least the 1940s. This commentary reviews a dozen of these common misinterpretations and explains why each is wrong. It also reviews the possible consequences of these improper understandings or representations of its meaning. Finally, it contrasts the *P* value with its Bayesian counterpart, the Bayes' factor, which has virtually all of the desirable properties of an evidential measure that the *P* value lacks, most notably interpretability. The most serious consequence of this array of *P*-value misconceptions is the false belief that the probability of a conclusion being in error can be calculated from the data in a single experiment without reference to external evidence or the plausibility of the underlying mechanism.
Semin Hematol 45:135-140

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 of true to no relationships among the relationships probed in each scientific field. In this framework, a research finding is less likely to be true when the studies conducted in a field are smaller; when effect sizes are smaller; when there is a greater number and lesser preselection of tested relationships; where there is greater flexibility in designs, definitions, outcomes, and analytical modes; when there is greater financial and other interest and prejudice; and when more teams are involved in a scientific field in chase of statistical significance. Simulations show that for most study designs and settings, it is more likely for a research claim to be false than true. Moreover, for many current scientific fields, claimed research findings may often be simply accurate measures of the prevailing bias. In this essay, I discuss the implications of these problems for the conduct and interpretation of research.

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 a consequence of the convenient, yet ill-founded strategy of claiming conclusive research findings solely on the basis of a single study assessed by formal statistical significance, typically for a *p*-value less than 0.05. Research is not most appropriately represented and summarized by *p*-values, but, unfortunately, there is a widespread notion that medical research articles

It can be proven that most claimed research findings are false.

should be interpreted based only on *p*-values. Research findings are defined here as any relationship reaching formal statistical significance, e.g., effective interventions, informative predictors, risk factors, or associations.

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 is only one true relationship (among many that can be hypothesized) or the power is similar to find any of the several existing true relationships. The pre-study probability of a relationship being true is $R/(R+1)$. The probability of a study finding a true relationship reflects the power $1-\beta$ (one minus the Type II error rate). The probability of claiming a relationship when none truly exists reflects the Type I error rate, α. Assuming that c relationships are being probed in the field, the expected values of the 2×2 table are given in Table 1. After a research finding has been claimed based on achieving formal statistical significance, the post-study probability that it is true is the positive predictive value, PPV. The PPV is also the complementary probability of what Wacholder et al. have called the false positive report

Ioannidis (2005), Goodman (2008), Nuzzo (2014), Amrhein, Greenland, and McShane (2019), ...

P-values / statistical significance criticism

P-value **criticism is** as **old** as statistical significance testing (1920s!).
Issues:

- The sharp line $p < 0.05$ is *arbitrary*.
- P-hacking / data dredging: Search until you find a result with $p < 0.05$.
- Publication bias: Studies with $p < 0.05$ are more likely to be published than "non-significant'' results.
- HARKING: Hypothesizing After the Results are Known.
- Model selection using p-values → **model selection bias**.

Note: R.A. Fisher, the "inventor" of the p-value (1920s) didn't mean the p-value to be used in the way it is used today, which is: doing a single experiment and use $p < 0.05$ for a conclusion.

From Goodman (2016):

> *Fisher used "significance" merely* ***to indicate that an observation was worth following up, with refutation of the null hypothesis justified only if further experiments "rarely failed" to achieve significance****. This is in stark contrast to the modern practice of making claims based on a single demonstration of statistical significance.*

Note: R.A. Fisher, the "inventor" of the p-value (1920s) didn't mean the p-value to be used in the way it is used today, which is: doing a single experiment and use $p < 0.05$ for a conclusion.

From Goodman (2016):

> *Fisher used "significance" merely* ***to indicate that an observation was worth following up, with refutation of the null hypothesis justified only if further experiments "rarely failed" to achieve significance****. This is in stark contrast to the modern practice of making claims based on a single demonstration of statistical significance.*

Right or wrong?

Go to www.menti.com and use code 3944 9342.
Which of these statements are right or wrong?

1. The p-value is the probability that the null hypothesis is true.
2. $p = 0.02$ means that the alternative hypothesis is true with 98% probability.
3. The p-value is the type-1 error rate.
4. The p-value is the probability that the result happened by chance.
5. If $p > 0.05$, we can conclude that there is no effect.
6. Two studies with $p > 0.05$ and $p < 0.05$ are in a conflict.

Significance thresholding is arbitrary

- Is there a significant difference between $p = 0.049$ and $p = 0.051$...??

Significance thresholding is arbitrary

- Is there a significant difference between $p = 0.049$ and $p = 0.051$...??

  No: *The difference between significant and non-significant is not necessarily significant.*

Significance thresholding is arbitrary

- Is there a significant difference between $p = 0.049$ and $p = 0.051$...??

  No: *The difference between significant and non-significant is not necessarily significant.*

- Does $p > 0.05$ automatically imply that a variable is unimportant or that it has no effect?

Significance thresholding is arbitrary

- Is there a significant difference between $p = 0.049$ and $p = 0.051$...??

  No: *The difference between significant and non-significant is not necessarily significant.*

- Does $p > 0.05$ automatically imply that a variable is unimportant or that it has no effect?

  No: *Absence of evidence is not evidence of absence (Altman and Bland 1995). The null hypothesis cannot be proved.*

Significance thresholding is arbitrary

- Is there a significant difference between $p = 0.049$ and $p = 0.051$...??

  No: *The difference between significant and non-significant is not necessarily significant.*

- Does $p > 0.05$ automatically imply that a variable is unimportant or that it has no effect?

  No*: Absence of evidence is not evidence of absence (Altman and Bland 1995). The null hypothesis cannot be proved.*

Reasons for large p-values:

- Low sample size ($\rightarrow$ low power).
- The truth is not far from the null hypothesis.
- Collinear covariates.

- “Statistical significance” is often used almost synonymously with “there is an effect”.
- But we all know: Correlation is not causation.

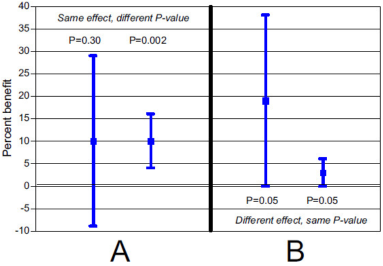
Significance vs relevance

Paul D. Ellis in *The Essential Guide to Effect Sizes* (2010, chapter 2):

> *Indeed, statistical significance, which partly reflects sample size, may say nothing at all about the practical significance of a result. [....] To extract meaning from their results [...] scientists need to look beyond p values and effect sizes and* ***make informed judgments about what they see****.*

- A low p-value does not automatically imply that a variable is "important" – and vice versa.

- "Is there an effect?" v.s. "How much of an effect is there?".



Goodman (2008)

Problem: The p-value blands the estimated effect size with its uncertainty.

Shall we abolish p-values?

NATURE | RESEARCH HIGHLIGHTS: SOCIAL SELECTION

Psychology journal bans *P* values

Test for reliability of results 'too easy to pass', say editors.

Chris Woolston

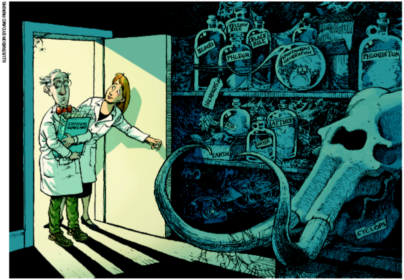
26 February 2015 | Clarified: 09 March 2015

PDF | Rights & Permissions

A controversial statistical test has finally met its end, at least in one journal. Earlier this month, the editors of *Basic and Applied Social Psychology* (*BASP*) announced that the journal would no longer publish papers containing *P* values because the statistics were too often used to support lower-quality research [1].

- But that throws the baby out with the bath water. It's as if we would forbid trains because they cannot fly to South America...
- p-values are not "good" or "bad". They have **strengths** and **weaknesses**.

What should we do then?



Retire statistical significance

Valentin Amrhein, Sander Greenland, Blake McShane and more than 800 signatories call for an end to hyped claims and the dismissal of possibly crucial effects.

- In many situations it is not justified to make a strict yes/no decision.[1]
- **Instead**: accumulating evidence over more and more studies.[2]

[1] And we are usually not forced to! In contrast to e.g. clinical trials.

[2] That's why it is so important to publish non-significant results, too! And: the importance of meta-analyses.

A small literature review

Did reporting behavior change?

Has the debate had an impact on how we report and interpret our findings in the ecology and evolution research community? In order to get a better feeling for this question, we carried out a small literature review. We used the January 2021 issues (December 2020 if January 2021 was a special issue) of eight major journals in ecology and evolution and checked all research papers containing at least one statistical analysis (n = 137, see the supplemental information online). Of those, 113 (82.5%) reported results based on the NHST philosophy: 104/113 (92%) of the dichotomous decisions were based on the P-value, while seven used the 95% CIs, and two used an information criterion. A total of 110/113 (97.3%) reported their findings using the 'significance' terminology. It appears as if the decades with waving warning flags had relatively little impact on the routines in our field when it comes to writing the results sections of scientific papers.

Suggestion 1: Language matters!

Rewrite your results and use a *gradual interpretation of the p-value*.

For single (observational) studies, the following has been suggested already decades ago (Bland 1986):

Interpreting the P value

As a rough and ready guide, we can think of P values as indicating the strength of evidence like this:

P value	**Evidence for a difference or relationship**
Greater than 0.1:	Little or no evidence
Between 0.05 and 0.1:	Weak evidence
Between 0.01 and 0.05:	Evidence
Less than 0.01:	Strong evidence
Less than 0.001:	Very strong evidence

Opinion

Rewriting results sections in the language of evidence

Stefanie Muff [1,2,*,@] Erlend B. Nilsen,[2,3,4,@] Robert B. O'Hara,[1,2,@] and Chloé R. Nater[2,3,@]

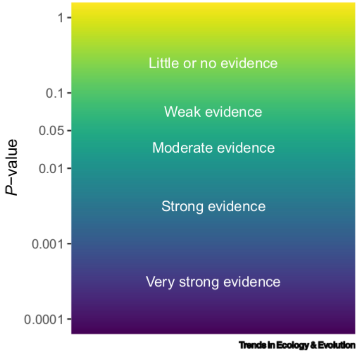


Figure 1. Suggested ranges to approximately translate the *P*-value into the language of evidence. The ranges are based on Bland (1986) [27], but the boundaries should not be understood as hard thresholds.

Suggestion 2: Report effect sizes, 95% CIs, and figures

Ask:

- Is the effect size (biologically, medically, socially...) relevant?
- Which range of true effects is statistically consistent with the observed data?

$\rightarrow$ 95% confidence interval

However

- The choice of the 95% is again somewhat arbitrary. We could also go for 90% or 99% or any other interval.
- The 95% CI should **not be misused for simple hypothesis testing** in the sense of "Is 0 in the confidence interval or not?" – that is just significance testing.

A results table from an example where I was involved (Imo et al. 2018):

Table 4. Evidence for the association with log-transformed mercury values in urine (µg/g creatinine).

n = 164	Variable	Coefficient	95% CI	p-Value
Very strong evidence	Amalgam fillings	0.33	0.24, 0.42	<0.001
	Last time sea fish	0.32	0.17, 0.47	<0.001
	Age	−0.04	−0.06, −0.02	<0.001
	Interaction age × mother	0.05	0.02, 0.08	<0.001
Strong evidence	Mother (indicator)	−0.97	−1.64, −0.31	0.004
	Smoking	0.30	0.09, 0.50	0.005
	Sea fish	0.08	0.03, 0.13	0.003
Little or no evidence	Log_{10} Hg soil	0.02	−0.06, 0.10	0.64
	Limit of quantification	−0.08	−0.25, 0.09	0.37
	Country of birth near the sea	−0.01	−0.16, 0.15	0.93
	Eats vegetables from region	0.07	−0.03, 0.18	0.18

CI: Confidence interval.

We found very strong evidence for a positive association of mercury in urine with amalgam fillings (regression coefficient: 0.33; 95% CI: 0.24–0.42; $p < 0.001$).

We found no evidence for an association of log-transformed mercury concentrations in soil with log-transformed concentrations in urine (regression coefficient: 0.02; 95% CI: −0.06–0.10; $p = 0.64$).

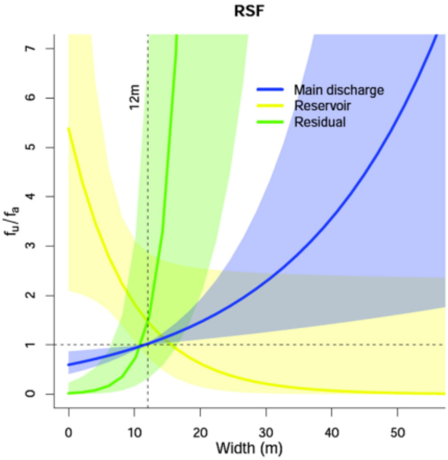
A graphical description often says more than thousand words...

Do you prefer

Two-step conditional logit over all nine animals. Significant factors are in bold.

Covariates	Beta	SD	p-Value (Wald)
Distance to road	**0.063**	**0.031**	**0.020**
Function of riverbed: width			
(Main discharge as reference category)			
Residual water: width	**3.115**	**1.621**	**0.027**
Reservoir: width	**−2.036**	**1.126**	**0.035**
Distance to dam	−0.103	0.077	0.090
River width	0.599	0.45	0.092
Algae	0.057	0.058	0.162
Distance to fishpond	−0.098	0.101	0.166
Type riparian vegetation	−0.035	0.041	0.194
Width riparian vegetation	−0.038	0.073	0.303
Function of riverbed			
(Main discharge as reference category)			
Reservoir	0.207	0.515	0.344
Residual water	0.288	1.285	0.411
Wood debris	0.027	0.086	0.377
Riverbank modifications	−0.002	0.038	0.474
Variability in depth	−0.002	0.054	0.483
Material bank side	0.000	0.033	0.500

or … ?



(Weinberger et al. 2016)

The interpretation of the p-value depends!

- Observational vs experimental study
- Exploratory vs confirmatory analysis

Practice in drug regulation

Clinical trials (CTs) for **drug approval** underlie strict requirements – since decades.

- CTs are **randomized controlled trials**.
- **Study protocols** that are published even before any patient is treated.
- **Preregistration** of study protocols and analysis plans.
- **Two Trials Rule**:

  "at least two adequate and well-controlled studies, each convincing on its own, to establish effectiveness."

- Clinical trials are *experimental* and *confirmatory*, and there are very strict regulations.

  → *We can draw a causal conclusion.*

- On the other hand, in Ecology: (Often) observational studies, lots of researchers degrees of freedom, usually no preregistration, exploratory data analysis, no study protocols, model selection,...

  → *We are mostly detecting correlations.*

Exercise

- Work in teams of 2-3 and choose one of the papers I will give you.
- Check how the authors reported their results.
- Make concrete suggestions (e.g., example sentences) how the authors could have better presented their results.

The material can be found here:

https://github.com/stefaniemuff/statlearning/tree/master/OpenScience

"Homework"

I recommend you to read the following short articles (you find the pdfs on the literature list):

- Scientists rise up against statistical significance (2019). Amrhein et al., *Nature*, 567, p. 305–307, https://doi.org/10.1038/d41586-019-00857-9
- The ASA statement on p-values: context, process, and purpose (2016). Wasserstein and Lazar, *The American Statistician*, 70:2, 129-133, https://doi.org/10.1080/00031305.2016.1154108
- Rewriting results sections in the language of evidence (2022). Muff et al., *Trends in Ecology and Evolution*, 37, 203–210, https://doi.org/10.1016/j.tree.2021.10.009

References

Altman, D. G., and J. M. Bland. 1995. "Absence of Evidence Is Not Evidence of Absence." *British Medical Journal* 311: 485.

Amrhein, V., S. Greenland, and B. McShane. 2019. "Retire Statistical Significance." *Nature* 567: 305–7.

Bland, J. M. 1986. *An Introduction to Medical Statistics*. Oxford: Oxford Medical Publications.

Goodman, S. N. 2008. "A Dirty Dozen: Twelve P-Value Misconceptions." *Seminars in Hematology* 45: 135–40.

———. 2016. "Aligning Statistical and Scientific Reasoning." *Science* 352: 1180–82.

Imo, D., S. Muff, R. Schierl, K. Byber, Ch. Hitzke, M. Bopp, M. Maggi, S. Bose-O'Reilly, L. Held, and H. Dressel. 2018. "Human-Biomonitoring and Individual Soil Measurements for Children and Mothers in an Area with Recently Detected Mercury-Contaminations and Public Health Concerns: A Cross-Sectional Study." *Nternational Journal Of Environmental Health Research* 28: 1–16.

Ioannidis, J. P. A. 2005. "Why Most Published Research Findings Are False." *PLoS Medicine* 2: e124.

Nuzzo, R. 2014. "Scientific Method: Statistical Errors." *Nature* 506: 150–52.

Wasserstein, R. L., and N. A. Lazar. 2016. "The ASA's Statement on p-Values: Context, Process, and Purpose." *The American Statistician*.

Weinberger, I. C., S. Muff, A. Kranz, and F. Bontadina. 2016. "Flexible Habitat Selection Paves the Way for a Recovery of Otter Populations in the European Alps." *Biological Conservation* 199: 88–95.