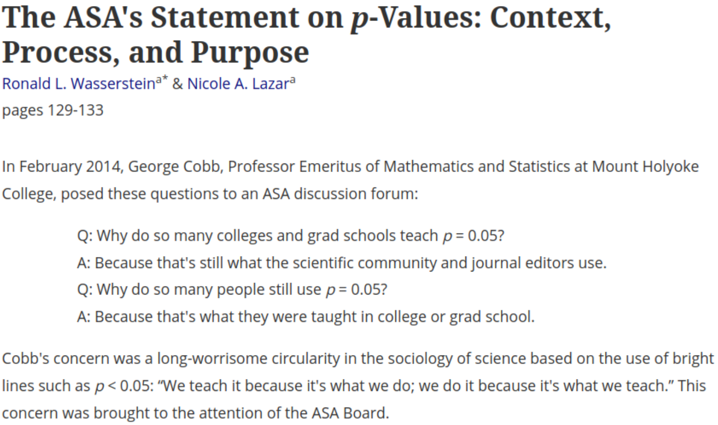
P-values, confidence intervals, and significance testing - navigating the minefield when communicating your findings

CBD talk

Stefanie Muff

November 10, 2020

The ongoing controversy around p-values



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

Ronald L. Wasserstein[a*] & Nicole A. Lazar[a]

pages 129-133

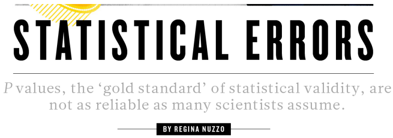
In February 2014, George Cobb, Professor Emeritus of Mathematics and Statistics at Mount Holyoke College, posed these questions to an ASA discussion forum:

> 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.

(Wasserstein and Lazar 2016)

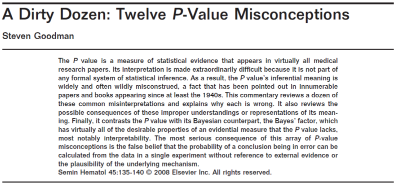
Lots of publications in the past decades...



STATISTICAL ERRORS

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

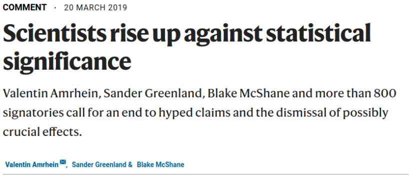
BY REGINA NUZZO



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 © 2008 Elsevier Inc. All rights reserved.

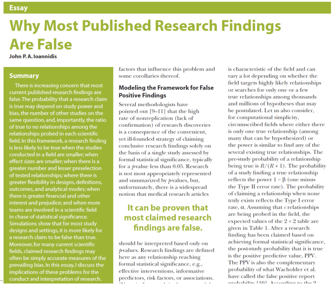


COMMENT · 20 MARCH 2019

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



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 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.

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

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 probability [10]. According to the 2

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* and significance testing according to it may lead to *mindless statistics* (Gigerenzer 2004).
- 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.

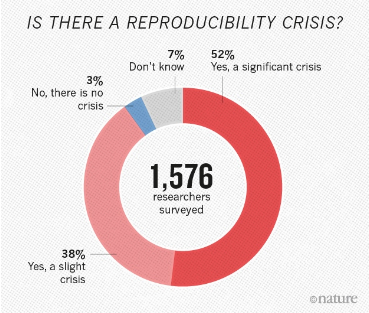
From Goodman (2016):

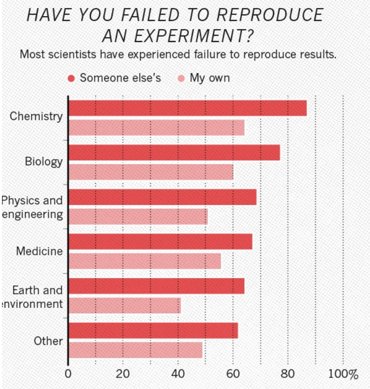
> *Fisher used "significance" merely* ***to indicate that an observation was worth following up, with refutation of the null hypothesis justified only if further experiments "rarely failed" to achieve significance****. This is in stark contrast to the modern practice of making claims based on a single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

The misuse of p-values is partially responsible for the reproducibility/replicability crisis in science!

A reproducibility crisis?

A survey carried out by *Nature* in 2016, sheds light on researcher's experiences and thoughts.



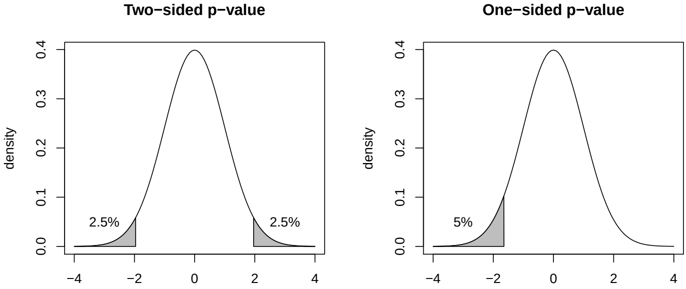


But wait, what is the problem with the *p*-value?

The *p*-value is not a very intuitive concept...

Formal definition

The *p*-value is the probability to observe a data summary (e.g., a *t*-value) that is at least as extreme as the one observed, given that the Null Hypothesis is correct.



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.

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.

Sorry, but this is all bullshit...

Some explanations

The p-value is

$$p = \mathrm{P}(\mathrm{Result} \,|\, H_0) \ ,$$

[1] $\mathrm{P}(H_0 \,|\, \mathrm{Result}) = \frac{\mathrm{P}(\mathrm{Result} \,|\, H_0) \cdot \mathrm{P}(H_0)}{\mathrm{P}(\mathrm{Result})}$

Some explanations

The p-value is

$$p = \mathrm{P}(\mathrm{Result} \,|\, H_0) \ ,$$

but unfortunately it is not the reverse (why?[1]):

$$p \neq \mathrm{P}(H_0 \,|\, \mathrm{Result}) \ .$$

[1] $\mathrm{P}(H_0 \,|\, \mathrm{Result}) = \frac{\mathrm{P}(\mathrm{Result} \,|\, H_0) \cdot \mathrm{P}(H_0)}{\mathrm{P}(\mathrm{Result})}$

Some explanations

The p-value is

$$p = \mathrm{P}(\mathrm{Result} \,|\, H_0) \ ,$$

but unfortunately it is not the reverse (why?[1]):

$$p \neq \mathrm{P}(H_0 \,|\, \mathrm{Result}) \ .$$

Similarly:

$$1 - p \neq \mathrm{P}(H_1 \,|\, \mathrm{Result}) \ .$$

[1] $\mathrm{P}(H_0 \,|\, \mathrm{Result}) = \frac{\mathrm{P}(\mathrm{Result} \,|\, H_0) \cdot \mathrm{P}(H_0)}{\mathrm{P}(\mathrm{Result})}$

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 (→ low power).
- The truth is not far from the null hypothesis.
- Collinear covariates.

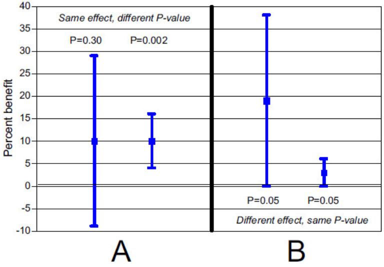
Significance vs relevance

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****.*

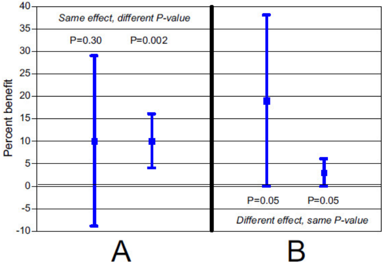
- A low *p*-value does not automatically imply that a variable is "important" – and vice versa.

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



Goodman (2008)

- 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.

An example

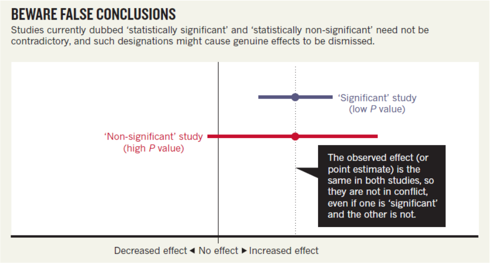
The WHO recommendation concerning smoking and the consumption of processed meat. Both, smoking and meat consumption, are “significantly” increasing the probability to get cancer.

- 50g processed meat per day increases the risk for colon cancer by a factor of 1.18 (+18%).
- Smoking increases the risk to get any type of cancer by a factor of 3.6 (+260%).

Thus: Although both, meat consumption and smoking, are carcinogenic (“significant”), their *effect sizes are vastly different*!

Are two studies in conflict?

- In the following example, two studies find the same effect size, but one is significant ($p = 0.02$) and the other one is not ($p = 0.09$).



BEWARE FALSE CONCLUSIONS

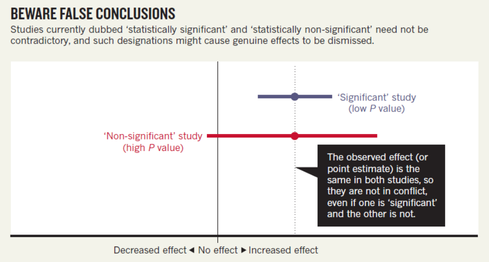
Studies currently dubbed 'statistically significant' and 'statistically non-significant' need not be contradictory, and such designations might cause genuine effects to be dismissed.

The observed effect (or point estimate) is the same in both studies, so they are not in conflict, even if one is 'significant' and the other is not.

Amrhein, Greenland, and McShane (2019)

Are two studies in conflict?

- In the following example, two studies find the same effect size, but one is significant ($p = 0.02$) and the other one is not ($p = 0.09$).



BEWARE FALSE CONCLUSIONS

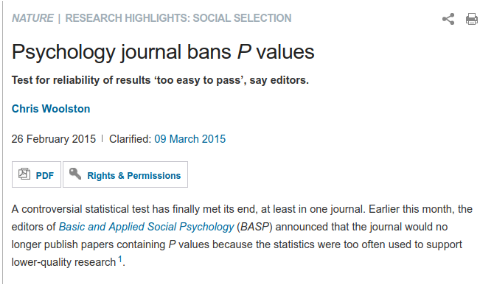
Studies currently dubbed 'statistically significant' and 'statistically non-significant' need not be contradictory, and such designations might cause genuine effects to be dismissed.

The observed effect (or point estimate) is the same in both studies, so they are not in conflict, even if one is 'significant' and the other is not.

Amrhein, Greenland, and McShane (2019)

- This is obviously no conflict, only the uncertainty is different.
- **Again:** The p-value blands the effect size with its uncertainty.

Shall we abolish p-values?



NATURE | RESEARCH HIGHLIGHTS: SOCIAL SELECTION

Psychology journal bans *P* values

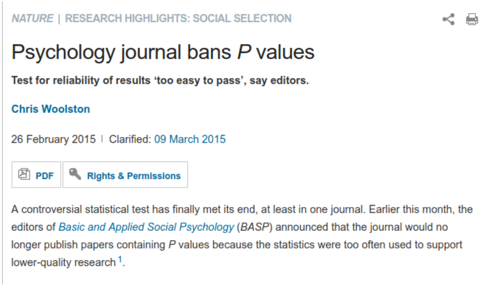
Test for reliability of results ‘too easy to pass’, say editors.

Chris Woolston

26 February 2015 | Clarified: 09 March 2015

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].

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

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 contain important information, and they have **strengths** and **weaknesses**.

What should we do then?

ILLUSTRATION BY DAVID PARKINS
YELLOW BILE
BLACK BILE
BLOOD
PHLEGM
SPONTANEOUS GENERATION
PHLOGISTON
HUMOURS
STATISTICAL SIGNIFICANCE
AIR
AETHER
WATER
EARTH
CYCLOPS

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.[2]
- **Instead**: accumulating evidence over more and more studies.[3]

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

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

But... an (almost) randomly selected example

- Last week I went to the *Evolution* website and just picked the most recent paper (published November 1st, first on the list).
- Guess how many times the word *significant* was used.

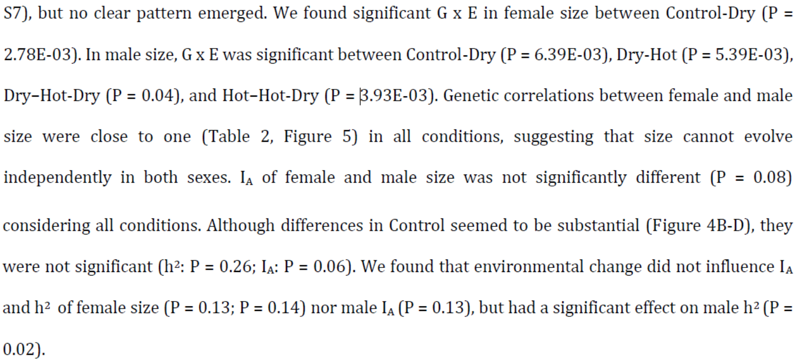
But... an (almost) randomly selected example

- Last week I went to the *Evolution* website and just picked the most recent paper (published November 1st, first on the list).

- Guess how many times the word *significant* was used.

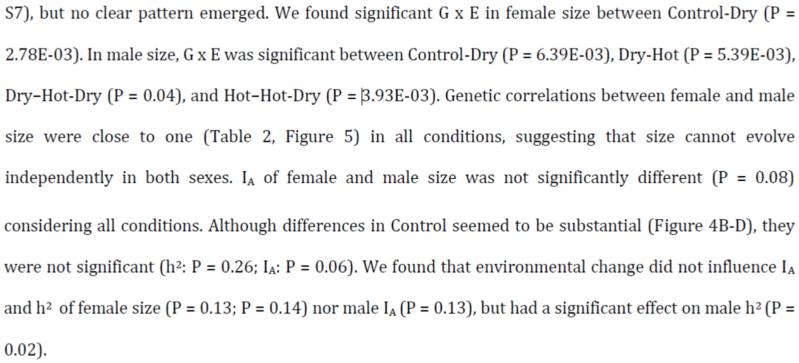
Answer: 55

A snapshot from the paper:



S7), but no clear pattern emerged. We found significant G x E in female size between Control-Dry (P = 2.78E-03). In male size, G x E was significant between Control-Dry (P = 6.39E-03), Dry-Hot (P = 5.39E-03), Dry–Hot-Dry (P = 0.04), and Hot–Hot-Dry (P = 3.93E-03). Genetic correlations between female and male size were close to one (Table 2, Figure 5) in all conditions, suggesting that size cannot evolve independently in both sexes. I_A of female and male size was not significantly different (P = 0.08) considering all conditions. Although differences in Control seemed to be substantial (Figure 4B-D), they were not significant (h^2: P = 0.26; I_A: P = 0.06). We found that environmental change did not influence I_A and h^2 of female size (P = 0.13; P = 0.14) nor male I_A (P = 0.13), but had a significant effect on male h^2 (P = 0.02).

A snapshot from the paper:



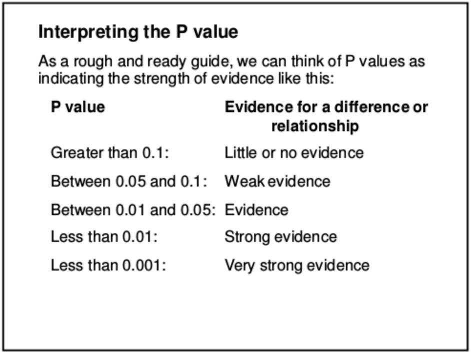
S7), but no clear pattern emerged. We found significant G x E in female size between Control-Dry (P = 2.78E-03). In male size, G x E was significant between Control-Dry (P = 6.39E-03), Dry-Hot (P = 5.39E-03), Dry–Hot-Dry (P = 0.04), and Hot–Hot-Dry (P = 3.93E-03). Genetic correlations between female and male size were close to one (Table 2, Figure 5) in all conditions, suggesting that size cannot evolve independently in both sexes. I_A of female and male size was not significantly different (P = 0.08) considering all conditions. Although differences in Control seemed to be substantial (Figure 4B-D), they were not significant (h^2: P = 0.26; I_A: P = 0.06). We found that environmental change did not influence I_A and h^2 of female size (P = 0.13; P = 0.14) nor male I_A (P = 0.13), but had a significant effect on male h^2 (P = 0.02).

How could we do better?

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

Suggestion 2: Report effect sizes and 95% CIs

Ask:

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

  → 95% confidence interval

Suggestion 2: Report effect sizes and 95% CIs

Ask:

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

  → 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):

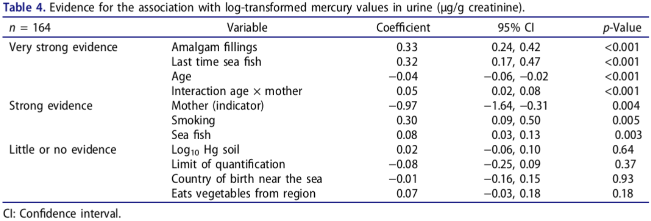


Table 4. Evidence for the association with log-transformed mercury values in urine (µg/g creatinine).

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.

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

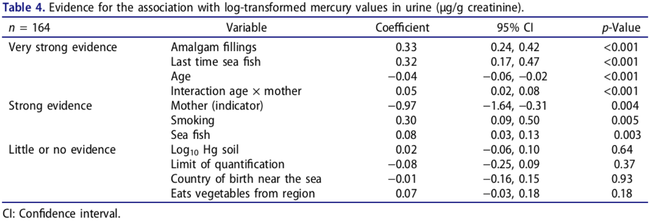


Table 4. Evidence for the association with log-transformed mercury values in urine (µg/g creatinine).

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).

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

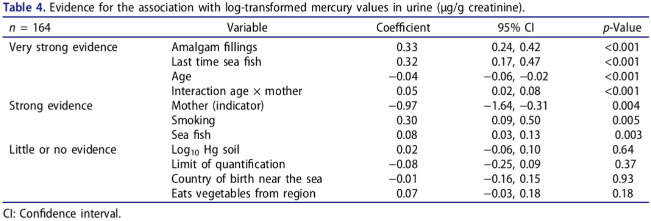


Table 4. Evidence for the association with log-transformed mercury values in urine (µg/g creatinine).

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; $p = 0.64$).*

The interpretation of the p-value depends!

- Observational vs experimental study
- Confirmatory vs exploratory 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."

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*, thus it is ok to draw causal conclusions.

Our situation is – for many reasons – not comparable to clinical trials:

- Clinical trials: Always experimental intervention and very strict regulations.
- Ecology: (Often) observational studies, lots of researchers degrees of freedom, usually no preregistration, exploratory data analysis, no study protocols, model selection,...

But, why not pick the good things from medical statistics?

Final suggestions:

- Use p-values, but *no significance testing*[4].
- Always report *effect sizes* and *confidence intervals*.
- Think about the *relevance*, not only the *significance*.
- More *gradual interpretation* of p-values, rewrite your results sections.
- Do not use p-values for model selection.

[4] unless you analyze a randomized controlled trial

Final suggestions:

- Use p-values, but *no significance testing*[4].
- Always report *effect sizes* and *confidence intervals*.
- Think about the *relevance*, not only the *significance*.
- More *gradual interpretation* of p-values, rewrite your results sections.
- Do not use p-values for model selection.

Question: Would you find a "translation guideline" to rewrite your results useful?

Thanks for feedback

[4] unless you analyze a randomized controlled trial

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.

Gigerenzer, G. 2004. “Mindless Statistics.” *The Journal of Socio-Economics* 33: 587–606.

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: 1180–2.

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*.