How feasible is it to abandon statistical significance? A reflection based on a short survey

Fredi Alexander Diaz-Quijano (frediazq@usp.br)
Universidade de Sao Paulo https://orcid.org/0000-0002-1134-1930
Fernando Morelli Calixto
Universidade de Sao Paulo
José Mário Nunes da Silva
Universidade de Sao Paulo

Research article

Keywords: p values, null-hypothesis, statistical inference, Statistical significance

Posted Date: September 9th, 2019

DOI: https://doi.org/10.21203/rs.2.14217/v1

Version of Record: A version of this preprint was published at BMC Medical Research Methodology on June 3rd, 2020. See the published version at https://doi.org/10.1186/s12874-020-01030-x.

Abstract

Background : There is a growing trend in using the “statistically significant” term in the scientific literature. However, harsh criticism of this concept motivated the recommendation to withdraw its use of scientific publications. We aimed to validate the support and the feasibility of adherence to this recommendation, among researchers having declared in favor of removing the statistical significance.

Methods : We surveyed signatories of an article published that defended this recommendation, to validate their opinion and ask them about how likely they will retire the concept of statistical significance.

Results: We obtained 151 responses which confirmed the support for the mentioned publication in aspects such as the adequate interpretation of the p-value, the degree of agreement, and the motivations to sign it. However, there was a wide distribution of answers about how likely are they to use the concept of "statistical significance” in future publications. About 42% declared being neutral, or that would likely use it again. We described arguments referred by several signatories and discussed aspects to be considered in the interpretation of research results. Conclusions : The declared position against the use of statistical significance had had legitimate support from numerous researchers. However, the full application of this recommendation does not seem feasible. The arguments related to the inappropriate use of statistical tests should promote more education among researchers and users of scientific evidence.

Background

The culture of testing the null hypothesis through the p-value has dominated the practice of statistical inference [1]. In this sense, the level of significance is defined according to the alpha error that we would be willing to accept when rejecting the hypothesis that there is no association between variables of interest [2]. This level (often set at 0.05) is used to define whether an association is “statistically significant” according to the p-value obtained from the tests [3].

However, the p-value varies depending on the sample size and the magnitude of the association, and the latter varies randomly even in the absence of biases. Due to a generalized application of the same level of significance, investigations on a particular subject can lead to different conclusions, especially with limited sample sizes or in the case of weaker associations [4–6]. Using lower levels of significance (e. g., 0.005 instead of 0.05) or calculating the posttest probability have been suggested to address lack of replication of the claimed associations [6–9].

With a more radical approach, an article recently published in the journal *Nature* called on the entire scientific community to abandon the concept of “statistical significance” in scientific publications. More than 800 signatories supported this paper [10]. Even though we agree with most of the arguments cited in this publication, we considered that the recommendation of entirely abandoning the statistical significance might be less useful than promoting its rational use.

Besides, there is a growing trend of using the term “statistically significant” in the biomedical literature, as seen in Pubmed searches (Figure 1). Therefore, we wonder how feasible it would be to abolish the use of this concept from our future publications. For this reason, we conducted a short survey with the signatories of the article to remove the statistical significance to consult the probability of them not using this term anymore. Also, we considered pertinent to validate the support of these researchers for the recommendation mentioned above.

Methods

We sent a short survey to each of the signatories by email between May 5 and 10 (approximately six weeks after publication). The questionnaire included three questions to avoid duplicate answers (country of residence, gender, and date of birth). Moreover, we added three questions to validate the support of the signatories:

- One of them presented a scenario for the interpretation of a value of p. This multichoice question had as a correct option one considering p-value as the probability of finding values at least as extreme as those observed, assuming that the null hypothesis was true. We also considered as right when participants did not check that answer but referred to a proper interpretation of the question in the comments.

- Another question aimed to confirm the support to the recommendation, which was formulated as “Currently, how much do you agree with the retiring of the statistical significance of future scientific publications?” The options were presented on a Likert scale with five possible responses.

- Also, we asked about factors influencing the decision to sign the paper on retiring of statistical significance. This question allowed choosing multiple answers from four suggested options (arguments against statistical significance, arguments in favor of alternative terms, the prestige of the authors and the prestige of the journals of the publication) and writing down other motivations.

In addition to these questions, we asked the signatories how likely they are to use the concept of “statistical significance” in their future publications. The options included:

- Never (I expect never to use it again)
- Unlikely (It is unlikely that I will use it again)
- Neutral, or it depends on the occasion
- Likely (It is likely that I will use it again)
- Always (I will use it every time I have the chance)

We presented the distribution of answers to this question as raw frequencies, and we also calculated values weighted by the inverse the probability of responding. To calculate this probability (ρ), we considered ten categories of geographical origin, including the eight countries with the highest number of responders and two regions grouping the others. Four participants did not provide data from residence, so for them, the average probability of response was considered (ρ = 151 / 854). We discounted the weight

of these four participants (4×1/ρ) from the total and distributed the rest among the other participants. In this way, the sum of all the weights remained equal to the number of signatories.

In the end, the questionnaire had an open question to record additional comments, and those we considered related to the discussion are presented in the supplementary material. We excluded any data that could potentially identify the respondent or another person.

This study was evaluated and approved by the Ethics Committee of the School of Public Health of the University of São Paulo. A link for an informed consent form was sent to the participants in the invitation email. The survey was intended to be responded anonymously, and any data suggesting an individual identity was treated confidentially.

Results

We received 153 responses but excluded two (one because it claimed to be lying and another because it was considered duplicate). In total, we obtained 151 valid responses, mostly from men (n = 136) with a median age of 43 years old (interquartile interval: 36 to 56). About the question of interpretation of the value of p, we considered 136 correct answers (including 133 closed selections and three open responses), corresponding to 90.1% of the total of respondents. Five participants did not answer this question or make a justification.

Relating to the current degree of agreement with the decision to withdraw the statistical significance, 98 (65%) answered that they strongly agreed, 49 (32%) partially agreed, three neither agreed nor disagreed, and only one indicated a strong disagreement. The last one referred that did “not agree with the title of the essay” and “it is unfortunate that the press and colleagues (…) are focusing on the title”.

Concerning the motivations to sign, the majority (142/151, 94%) answered that it was because of the arguments against statistical significance, followed by the arguments in favor of the alternatives (91/151, 60.3%). Only a minority of the respondents recognized that the prestige of the authors (n = 9) or of the journal (n = 12) were part of the motivations; however, for none of them, this was the only motivation. Additionally, 20 respondents reported other motives such as problems of misinterpretation and misuse of p-values.

Regarding the probability of using the concept of statistical significance in future publications, 35 (23%) responded that they expect never to use it again; and, 52 (34%) said it would be unlikely. On the other hand, 34 (23%) answered as neutral, or it depends on the occasion; 29 (19%) indicated that it would be likely to be used again and one stated that they would use it whenever they had the opportunity (Figure 2A). We obtained similar results when we weighted the frequencies by the countries of residence (Figure 2B).

We received several comments about the matter (supplementary material), of which we highlighted the following:

- " “Significance” with firm thresholds is the problem. The credibility of a result is multi-determined. P-level is one of the determinants, but only one. The main - really the only - thing that’s needed to determine the credibility of a result is replication. There is no shortcut; you can’t know what the study would find if you repeated it, unless you repeat it. The use of “significance” and even exact p-levels typically is an attempt to avoid this stubborn truth.”
- “Although I signed in agreement with the article, I do not think the title was properly reflecting the spirit with which it was written. We are not advocating to drop statistical significance altogether, but to make a more mindful use of it. The main mistakes are 1) to think the p-value gives us a measure of the strength or magnitude of a relationship, for example. 2) a p value can help use supporting or rejecting alternative hypothesis. We need to incorporate measures that make sense in the system we are studying. Effect sizes, confidence intervals, Bayesian or Information Theory approaches in addition to the classical stats.”
- “It is impossible to interpret a p-value in the absence of some prior estimate of the probability of the null hypothesis being true (or false). I am much more in favor of presenting the Bayes Factor Bound.”
- “The paper in question proposed to stop using the term “statistical significance”. It did NOT propose that p values should be banned, but only that they should not be dichotomized. I proposed that p values should be supplemented by a number that represents the risk that a “positive” test is a false positive.”

Discussion

This independent survey may be considered a validation of the support of a group of researchers to a recommendation to abandon the use of the concept of statistical significance. With very few exceptions, the signatories correctly interpreted the p-value. This result is not superfluous because some studies suggest that the misinterpretation of the p-value can be frequent even among academics [11–13]. Besides, most responders strongly agree to abandon the use of “statistical significance” and, for the most part, were motivated by the arguments presented against this concept.

However, regarding the feasibility of abandoning statistical significance, close to a quarter are fully convinced that they will never use this concept again. On the other hand, about 42% declared being neutral or that would likely use it in future publications. Assuming that the researchers surveyed represent those against the concept, the distribution of answers to this question suggests that the fully retire of the statistical significance does not seem feasible.

Because we were looking for a high response rate in our survey, we did not include questions related to the causes for which signatories would again use statistical significance in future publications. Therefore, the fact of using the concept of statistical significance does not mean that they are going to

base their conclusions solely or primarily on this result. Also, it is possible those continuing to use this term would be motivated by compliance the expectations of journals, reviewers, or readers, more than by their way of interpreting the results.

The p-value will continue to be presented, and dichotomization results seem to be inevitable regardless of the criterion chosen. Despite this, based on the validation we have made, we consider that Amhrein et al. materialized in their paper a legitimate concern of researchers from different areas. In that sense, we agree with the importance of a research finding not being based only on statistical significance [14, 15].

An aspect to highlight is to differentiate the application scenarios from statistical significance [12]. For example, there is a critical distinction between studies of causal inference vs. those for prediction purposes. In the latter, the interpretability of the estimates may be optional, and the statistical criteria can command decisions to use or not a predictor [16, 17]. However, in studies of causal inference, the concept of statistical significance should not be a primary concern. Before looking at a p-value, the researcher will have to avoid biases and look on conceptual structures to control confusion and consider contexts in which effects can be modified [18, 19]. After that, the measures of association and impact are those that must define when a result is significant in the clinical and public health scopes [20].

Therefore, it is not surprising that one of the major concerns expressed by several of the signatories is the misuse and misinterpretation of the value of p. Also, well-documented publication biases in favor of “positive results” are a consequence of overvaluation of statistical significance [21, 22]. These are often concerns among editors and statistical consultants of biomedical journals. For example, *The New England Journal of Medicine* recently modified its guidelines for statistical reporting by including a requirement to replace P values with estimates and confidence intervals when neither the protocol nor the analysis plan has specified methods to adjust for multiplicity [23].

We agree that the value of a result must be based on the interpretation of the spectrum of values compatible with the data, such as Amrhein *et al.* suggested [10]. However, removing a term such as statistical significance is far to be a solution to avoid the publication biases. We regret that, even based on point and interval estimates, associations compatible with the null value undoubtedly would continue being under-reported. Conversely, the absence of a preset threshold to interpret a p-value could subtract objectivity [24].

Faced with the seemingly inevitable use of statistical significance[21], we must give due value to statistical tests, promoting the understanding of their limitations [25]. In that sense, one critical issue is the widespread application of an arbitrary significance level (i.e., 0.05) [12, 26]. As an analogy, diagnostic tests may need different cut-offs according to the disease prevalence to maintain high predictive values [27]. Similarly, it would be negligent in using the same significance level for all research problems. The pre-test probability of an association would help to define a cut-off to increase the chance of both identifying the true associations and discarding those spurious [7]. Nevertheless, no value should become a new thumb rule applicable to all situations.

Reducing the significance level can reduce the false positive rate but increase the false negative rate, which is reducing the power of a study. This can be a problem when decisions have to be based on studies with small sample sizes, such as in preliminary outbreak investigations or in the research of extremely rare but severe diseases.

For another purpose, a study aimed to replicate or confirm results from other well-designed studies would not need to use the same level of significance, since the state of knowledge has changed. A higher significance level could be justified when previous studies suggested a high pre-test probability.

Moreover, other issues may be necessary to consider in each case, such as the implications of a false positive and false negative result. For example, it does not seem sensible to use the same significance level to approve a drug with a high risk of adverse effects as for a low-risk educational intervention. Probably in the former, we were more interested in ruling out the alpha error. As with diagnostic tests, the cut-off point for significance should also be adjusted to increase the likelihood that our research will cause more benefit than harm.

For all the above, it is likely that we have not yet found a magical formula to choose levels of significance. Therefore, we share the frustration of decisions being guided by an arbitrary or poorly justified rule. Statistical significance may play a supporting role, but not a leading one. However, better than trying to abolish this concept, we consider it is necessary to develop strategies to justify and predefine the significance levels considering both the evidence and the implications of errors resulting from the statistical tests. Moreover, efforts to define what is clinically or epidemiologically significant may be more useful to guide research and interventions [18, 19].

Conclusions

The declared position against the use of statistical significance has had legitimate support from numerous researchers. However, the full application of this recommendation does not seem feasible (at least shortly). The arguments against the inappropriate use of statistical tests should promote more education among researchers and users of scientific evidence. Probably, the main problem does not rely on choosing a cutoff for the p-value, but on our difficulty recognizing the limitations of both statistics and rules.

Declarations

Ethical approval and consent to participate.

This study was evaluated and approved by the Ethics Committee of the School of Public Health of the University of São Paulo. The committee waived the need for written or verbal consent. However, a link for an informed consent form was sent to the participants in the invitation email. Moreover, in the text of this email was stated that “by answering the survey, I [FADQ] am assuming you have read the form, and you consent freely to participate.”

Consent for publication

Not applicable because the manuscript does not contain individual data.

Availability of data and material

All data generated and analyzed during this study are included in this published article and its supplementary material file.

Competing interest

None of the authors has any competing interests.

Funding

The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Authors' Contributions

FADQ conceived the study, participated in the design and coordination thereof, analyzed the data and prepared the first draft of the manuscript. FMC and JMNS participated in the preparation of the questionnaire, data collection and the interpretation and discussion of results. All authors read, carried out critical reviews of the manuscript and approved the final manuscript.

Acknowledgments

The authors would like to thank the signatories that kindly responded to the survey.

References

1. Lash TL. The Harm Done to Reproducibility by the Culture of Null Hypothesis Significance Testing. Am J Epidemiol. 2017;186:627–35.
2. Fisher R. Statistical Methods for Research Workers. 14th edition. Edinburgh: Oliver and Boyd; 1970.
3. Goodman S. A Dirty Dozen: Twelve P-Value Misconceptions. Semin Hematol. 2008;45:135–40.
4. Altman N, Krzywinski M. Interpreting P values. 2017. doi:10.1038/nmeth.4210.
5. Greenland S. Nonsignificance Plus High Power Does Not Imply Support for the Null Over the Alternative. Ann Epidemiol. 2012;22:364–8.
6. Mark DB, Lee KL, Frank;, Harrell E. Understanding the Role of P Values and Hypothesis Tests in Clinical Research. Clin Rev Educ JAMA Cardiol | Spec Commun. 2016;1:1048–54.
7. Ioannidis JPA. The Proposal to Lower P Value Thresholds to.005. JAMA. 2018. doi:10.1001/jama.2018.1536.

8. Colquhoun D. The False Positive Risk: A Proposal Concerning What to Do About p-Values. Am Stat. 2019;2019:192–201.
9. Benjamin DJ, Berger JO, Johannesson M, Nosek BA, Wagenmakers E-J, Berk R, et al. Redefine statistical significance. Nat Hum Behav. 2018;2:6–10. doi:10.1038/s41562–017–0189-z.
10. Amrhein V, Greenland S, McShane BB. Retire statistical significance. Nature. 2019;567:305–7. https://www.nature.com/articles/d41586–019–00857–9.
11. Badenes-Ribera L, Frías-Navarro D, Monterde-I-Bort H, Pascual-Soler M. Interpretation of the p value: A national survey study in academic psychologists from Spain. Psicothema. 2015;27:290–5.
12. Greenland S, Senn SJ, Rothman KJ, Carlin JB, Poole C, Goodman SN, et al. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. Eur J Epidemiol. 2016;31:337–50. doi:10.1007/s10654–016–0149–3.
13. Cassidy SA, Dimova R, Giguère B, Spence JR, Stanley DJ. Failing Grade: 89% of Introduction-to-Psychology Textbooks That Define or Explain Statistical Significance Do So Incorrectly. Adv Methods Pract Psychol Sci. 2019.
14. Wasserstein RL, Schirm AL, Lazar NA. Moving to a World Beyond “p < 0.05.” Am Stat. 2019;73:1–19.
15. Ho J, Tumkaya T, Aryal S, Choi H, Claridge-Chang A. Moving beyond P values: Everyday data analysis with estimation plots. bioRxiv. 2019;:377978. doi:10.1101/377978.
16. Johansson U, Sönströd C, Norinder U, Boström H. Trade-off between accuracy and interpretability for predictive in silico modeling. Futur Med Chem. 2011;3:647–63.
17. Schooling CM, Jones HE. Clarifying questions about “risk factors”: predictors versus explanation. Emerg Themes Epidemiol. 2018;15:10.
18. Jakobsen JC, Gluud C, Lange T, Wetterslev J. The thresholds for statistical and clinical significance – a five-step procedure for evaluation of intervention effects in randomised clinical trials. BMC Med Res Methodol. 2014;14:1–12.
19. Koretz RL. Assessing the Evidence in Evidence-Based Medicine. Nutr Clin Pract. 2019;34:60–72.
20. Glass T, Goodman S, Hernán MA, Samet JM. Causal Inference in Public Health. Ssrn. 2013;:61–75.
21. Kyriacou DN. The Enduring Evolution of the P Value. JAMA. 2016;15:1113.
22. Perneger T V., Combescure C. The distribution of P -values in medical research articles suggested selective reporting associated with statistical significance. J Clin Epidemiol. 2017;87:70–7. doi:10.1016/j.jclinepi.2017.04.003.
23. Harrington D, D’Agostino RB, Gatsonis C, Hogan JW, Hunter DJ, Normand S-LT, et al. New Guidelines for Statistical Reporting in the Journal. N Engl J Med. 2019;381:285–6. doi:10.1056/NEJMe1906559.
24. Ioannidis JPA. Retiring significance: a free pass to bias. Nature. 2019;567. https://www.nature.com/magazine-assets/d41586–019–00969–2/d41586–019–00969–2.pdf. Accessed 6 Apr 2019.

25. Wasserstein RL, Lazar NA. The ASA's Statement on p-Values: Context, Process, and Purpose. Am Stat. 2016;70:129–33.
26. Lakens D. On the challenges of drawing conclusions from p-values just below 0.05. Peer J. 2015;3:e1142.
27. Weitkunat R, Kaelin E, Vuillaume G, Kallischnigg G. Effectiveness of strategies to increase the validity of findings from association studies: size vs. replication. BMC Med Res Methodol. 2010;10:47.

Figures

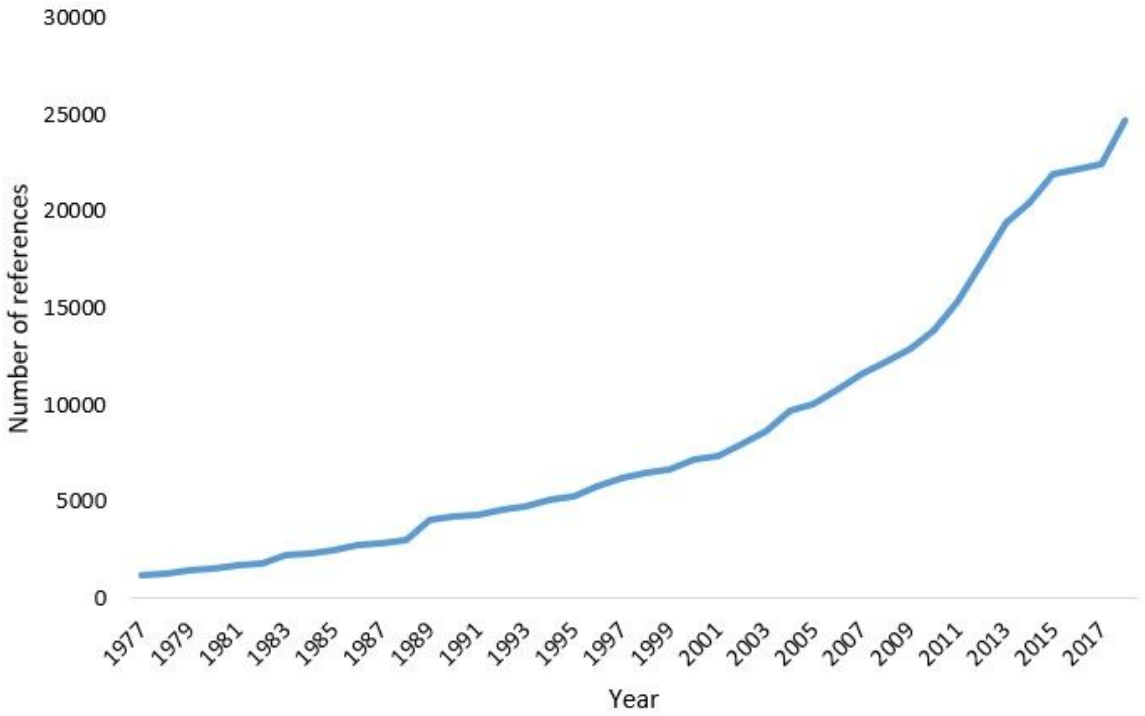

Figure 1

Count of references obtained in Pubmed with the term "statistically significant" (1977-2018)

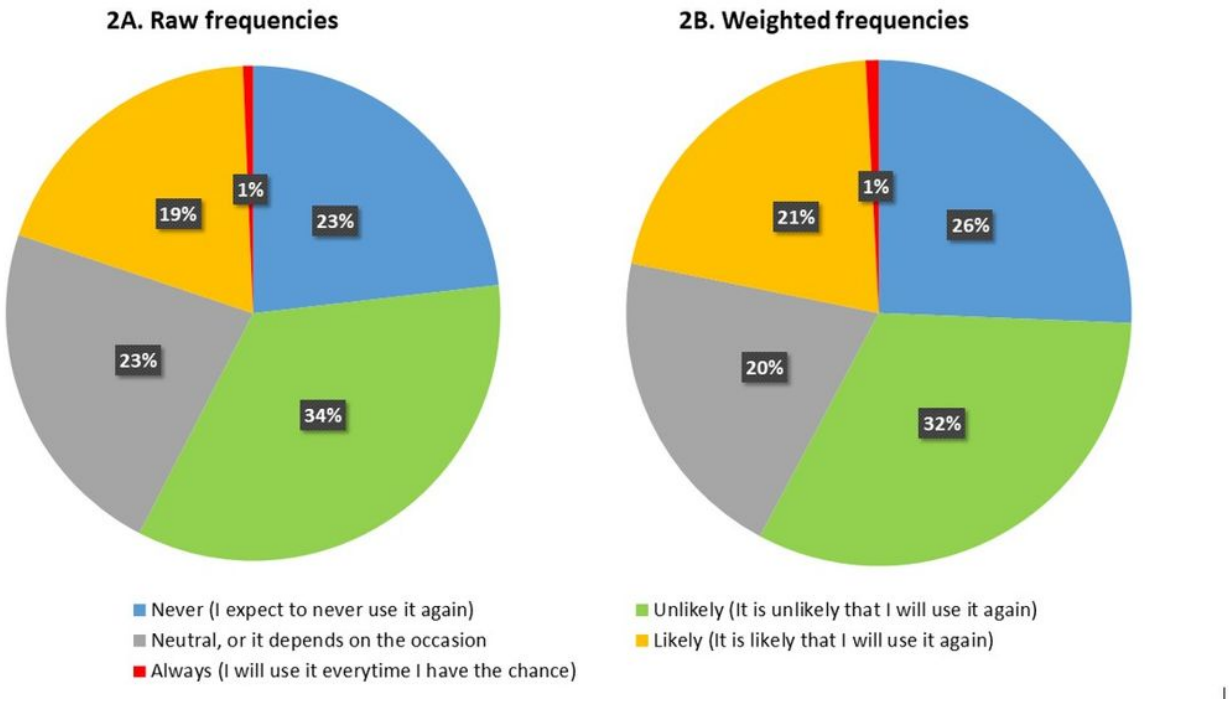

Figure 2

In your future publications, how likely are you to use the concept of "statistical significance"?

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- supplement1.docx