Editorial

Toward a More Nuanced Interpretation of Statistical Significance in Biomedical Research

Ram Bajpai[1] Himanshu K. Chaturvedi[2]

[1]School of Medicine, Keele University, Staffordshire, United Kingdom
[2]National Institute of Medical Statistics, Indian Council of Medical Research, New Delhi, India

Asian J Oncol 2021;7:49–51.

Introduction

Statistical significance is the bedrock of evidence-based medicine, crucial for decision-making and framing biomedical science policies. The emergence of complex data in many disciplines, resulting from the analysis of large datasets, has amplified the popularity of *p*-values. Its simplicity allows investigators to conclude and disseminate their research findings in a manner understood by most. Thus, obtaining a *p*-value that indicates "statistical significance" against the null hypothesis is often required for publishing in medical journals. However, it creates challenges due to nonreproducibility, misuse and overinterpretation, which lead to serious methodological errors. This article aims to draw biomedical researchers' attention toward the appropriate use of *p*-values in clinical decision-making.

Historical Development

Fisher first introduced null hypothesis significance testing (NHST) in 1925.[1] Subsequently, *p*-values became the standard of reporting and judging scientific evidence's strength when testing the null hypotheses against the alternative proposition in most scientific disciplines, including biomedical research. The recent development of big data research has made *p*-values even more popular to test the significance of study findings and has become a sine qua non for publishing in medical journals. The *basis of the p*-value is that it computes the probability of observing results at least as extreme as the ones observed, given that the null hypothesis is correct and compared against a predetermined significance level (α). If the reported *p*-value is lesser than α, the test result is to be considered statistically significant (►**Fig. 1**). Typically, α is set arbitrarily at 0.05 level to control false-positive rate, and other commonly used significance levels are 0.01 and 0.001.

Debate on Statistical Significance

Many researchers have questioned the acceptability of *p*-values in medical decision-making, and considerable research exists into how *p*-values are misused.[2,3] For example, in his seminal paper, Cohen[4] argued that: "NHST is highly flawed as it is relatively easy to achieve results that can be labeled significant when a 'nil' hypothesis is used rather than a true 'null' hypothesis." In recent years, despite its success, there is an emerging debate about whether to use *p*-values to describe statistically significant scientific results due to its frequent failure to reproduce and replicate similar statistically significant findings. Halsey et al[5] argued that: "the *p*-value is often used without the realization that in most cases the statistical power of a study is too low for *p* to assist the interpretation of the data.… Researchers would do better to discard the *p*-value and use alternative statistical measures for data interpretation." In agreement with this thinking, the journal of *Basic and Applied Social Psychology* recently barred *p*-values and hypothesis testing from articles published in their journal.[6] *p*-Values are also susceptible to "hacking," to demonstrate statistical significance when no association exists and encourages selective reporting of only positive findings. A recent methodological review of articles published in high impact journals suggests that significant results are about twice as likely to be reported as nonsignificant results.[7]

In our opinion, *p*-values alone cannot be responsible for the lack of reproducibility of research findings, as it is often a combination of methodological errors and interpretation. Like other statistical measures, the *p*-value is also a one-dimensional metric and based on data; thus, it could be misleading when calculated from relatively small samples. The overall selection of statistical methods, including the lack of randomness in the sample and missing data, can

Address for correspondence
Ram Bajpai, PhD, School of Medicine, Keele University, Staffordshire ST5 5BG, United Kingdom (e-mail: r.bajpai@keele.ac.uk).

DOI https://doi.org/10.1055/s-0041-1727066
ISSN 2454-6798

Thieme Medical and Scientific Publishers Pvt. Ltd., A-12, 2nd Floor, Sector 2, Noida-201301 UP, India

also influence the statistical significance that may result in misleading *p*-values. Misinterpretation of the *p*-value as a measure of the strength of association, rather than its true meaning (assessing the probability for a given result arising due to chance), is a glaring indictment of the current biomedical teaching standards. A recent methodological survey of statistical methods in Indian journals suggests that no significant progress has been achieved regarding the correct use of statistical analyses.[8]

Suggested Alternatives on Statistical Significance

Several alternatives to *p*-values have been suggested in the literature, such as confidence intervals and Bayesian statistics.[9] A confidence interval provides the point estimate with uncertainty bounds that can be more informative than a *p*-value. However, confidence intervals are like *p*-values when testing the null hypothesis's acceptance or rejection and challenging to compare between studies due to unit-dependence. In Bayesian statistics, the credible interval, equivalent to the frequentist approach's confidence interval, is another possible alternative to the *p*-value. Both alternative methods are like the *p*-value when testing the null hypothesis for clinical decision-making and can misinterpret the results' clinical or biological importance. Recently, Benjamin et al proposed

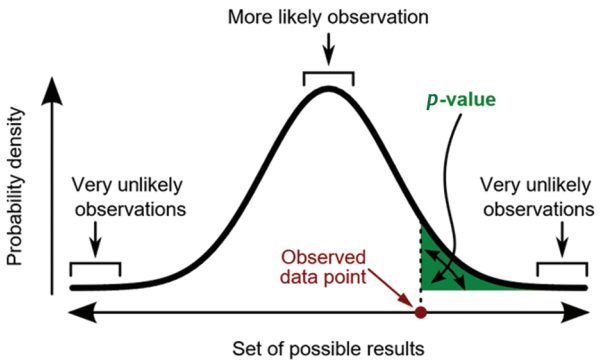

Fig. 1 Definition of *p*-value (https://en.wikipedia.org/wiki/Statistics).

to lower the *p*-value (the conventional "statistical significance") threshold from 0.05 to 0.005 for all novel claims with relatively low prior odds, to avoid high false positives and improve the reproducibility of scientific research.[10] However, others have argued that research should be guided by rigorous scientific principles, not by heuristics and arbitrary thresholds. These principles include sound statistical analyses, replication, validation and generalization, avoidance of logical traps, intellectual honesty, research workflow transparency, and accounting for potential sources of error.[11,12]

Guidance on Appropriate Interpretation of Statistical Significance

Therefore, educating researchers with appropriate training on concepts and relevant software could be one alternative to prevent misinterpretation of the *p*-value. It is worth reiterating Fisher's initial view that *p*-values should be one part of the evidence used when deciding whether to reject the null hypothesis. Appropriate guidance should also be taken during the design, process and data analysis of a study from various available resources such as Consolidated Standards of Reporting Trials (CONSORT)[13] for clinical trials, Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA)[14] for systematic reviews and meta-analysis, Strengthening the Reporting of Observational Studies in Epidemiology (STROBE)15 for observational studies, and Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD)[16] for risk prediction from a multivariable model.

As mentioned in NHST approach, researchers commonly classify results as statistically "significant" or "not-significant," based on whether the *p*-value is smaller than some prespecified cut point value (usually 0.05). However, this practice is becoming obsolete, and exact *p*-values are preferred by leading medical journals such as the *British Medical Journal* (BMJ), *Journal of the American Medical Association* (JAMA), and the *Lancet*. Guidance should be taken from Fisher's[1] belief about *p*-values from his 1925 book and Efron's[17] interpretation on observed *p*-values (or achieved significance level) as presented in ►**Table 1**. The American Statistical Association (ASA) took this matter to their board in 2015 and discussed it with renowned statisticians in multiple rounds for several

Table 1 Fisher's belief and Efron's interpretation of observed *p*-values

Fisher's beliefs regarding *p*-values		Efron's interpretations of achieved significance levels	
p-value	Fisher's statements	Achieved significance levels (ASL)	Interpretation
0.1–0.9	Certainly, no reason to suspect the hypothesis tested	ASL < 0.10	Borderline evidence against the null hypothesis
0.02–0.05	Judged significant, though barely so … these data do not, however, demonstrate the point beyond the possibility of doubt	ASL < 0.05	Reasonably strong evidence against the null hypothesis
< 0.02	Strongly indicated that the hypothesis fails to account for the whole of the facts	ASL < 0.025	Strong evidence against the null hypothesis
< 0.01	No practical importance of whether *p*-value is 0.01 or 0.000001	ASL < 0.01	Very strong evidence against the null hypothesis

months. Further, a "statement on statistical significance and *p*-values" with six principles has been released by ASA at the beginning of 2016 to guide researchers and avoid any misuse and misinterpretation.[18] In its statement, the ASA advised researchers to avoid drawing scientific conclusions or making policy decisions purely based on *p*-values. Additionally, they recommend describing the data analysis approach that produces statistically significant results, including all statistical tests and choices made in calculations. Otherwise, results may appear misleadingly robust.

Conclusion

In conclusion, we advocate that statistics should be used as a science rather than a recipe for the desired flavor. While researchers want certainty, they must understand that statistics is a science of uncertainty. Thus, solely relying on a single *p*-value to describe the scientific value of a study is a misuse of the *p*-value and when evaluating the strength of any evidence, researchers need to explain their results in the light of theoretical considerations such as scope, explanatory extent, and predictive power.

Conflict of Interest
None declared.

References

1 Fisher RA. Statistical Methods for Research Workers 1st edition. London: Oliver and Boyd; 1925
2 Ioannidis JPA. Why most published research findings are false. PLoS Med 2005;2(8):e124
3 Ziliak ST, McCloskey DN. The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives. Ann Arbor, MI: University of Michigan Press; 2008:1–22
4 Cohen J. The earth is round (p<.05) Am Psychol 1994;49(12): 997–1003
5 Halsey LG, Curran-Everett D, Vowler SL, Drummond GB. The fickle P value generates irreproducible results. Nat Methods 2015;12(3):179–185
6 Trafimow D, Marks M. Editorial. Basic Appl Soc Psych 2015; 37(1):1–2
7 Perneger TV, Combescure C. The distribution of P-values in medical research articles suggested selective reporting associated with statistical significance. J Clin Epidemiol 2017;87:70–77
8 Hassan S, Yellur R, Subramani P, et al. Research design and statistical methods in Indian medical journals: a retrospective survey. PLoS One 2015;10(4):e0121268
9 Lu Y, Belitskaya-Levy I. The debate about p-values. Shanghai Jingshen Yixue 2015;27(6):381–385
10 Benjamin DJ, Berger JO, Johannesson M, et al. Redefine statistical significance. Nat Hum Behav 2018;2(1):6–10
11 Ioannidis JPA. The proposal to lower p value thresholds to. 005. JAMA 2018;319(14):1429–1430
12 Lakens D, Adolfi FG, Albers CJ, et al. Justify your alpha. Nat Hum Behav 2018;2(3):168–171
13 Schulz KF, Altman DG, Moher D; CONSORT Group. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. BMJ 2010;340:c332
14 Moher D, Liberati A, Tetzlaff J, Altman DG; PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. PLoS Med 2009;6(7):e1000097
15 von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP; STROBE Initiative. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. Lancet 2007;370(9596) :1453–1457
16 Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): the TRIPOD statement. Br J Surg 2015;102(3):148–158
17 Efron B, Tibshirani RJ, An Introduction to the Bootstrap 1st edition. New York: Chapman & Hall; 1993
18 Wasserstein RL, Lazar NA. The ASA statement on p-values: context, process, and purpose. Am Stat 2016;70(2):129–133