J Korean Med Sci. 2025 Nov 17;40(44):e321
https://doi.org/10.3346/jkms.2025.40.e321
eISSN 1598-6357·pISSN 1011-8934

JKMS

Review Article
Editing, Writing & Publishing

The *P* Value: What It Is and What It Is Not

Farrokh Habibzadeh

Independent Research Consultant, Shiraz, Iran
Past President, World Association of Medical Editors (WAME)
Editorial Consultant, *The Lancet*
Associate Editor, *Frontiers in Epidemiology*

Received: Sep 14, 2025
Accepted: Oct 21, 2025
Published online: Nov 6, 2025

Address for Correspondence:
Farrokh Habibzadeh, MD
Independent Research Consultant, Shiraz 7186663745, Iran.
Email: Farrokh.Habibzadeh@gmail.com

ORCID iD
Farrokh Habibzadeh
https://orcid.org/0000-0001-5360-2900

Disclosure
The author has no potential conflicts of interest to disclose.

ABSTRACT

The *P* value remains one of the most frequently reported statistical measures in biomedical literature, yet it is also one of the most widely misunderstood statistics. Introduced by Fisher as a measure of evidence against the null hypothesis, it was subsequently incorporated into the Neyman-Pearson decision framework, which emphasized long-run error rates and decision thresholds. Over the past decades, reliance on the conventional cut-off of $P = 0.05$ has fostered misconceptions, including the belief that the *P* value represents the probability that the null hypothesis is true or that statistical significance implies clinical importance. In this review, I examine the historical evolution of the *P* value, clarify the conceptual distinctions between evidential and decision-theoretic perspectives, and illustrate their implications through a case study. Common misinterpretations and the limitations of threshold-based inference are discussed, together with the consequences for reproducibility, statistical power, and interpretation of results. Recent recommendations from statistical associations and methodologists to move beyond dichotomous significance testing are highlighted. Complementary approaches, such as estimation of effect sizes with confidence intervals (CIs), likelihood ratios, and Bayesian inference, are briefly considered. I conclude that although the *P* value may provide useful information when properly interpreted, it should not be used as a sole criterion for inference. Transparent reporting of effect sizes, CIs, and contextual information offers a more reliable foundation for scientific interpretation and decision making.

Keywords: Statistics as Topic; Confidence Intervals; Likelihood Functions; Biostatistics; Publication Bias; *P* Value

INTRODUCTION

The *P* value is by far the most frequently reported statistic in the scientific literature. It appears across virtually all fields of empirical research, including medicine, biology, psychology, and economics. In over 350,000 PubMed Central articles between 1990 and 2015, each article reported an average of nine *P* values, with the frequency steadily increasing over time.[1] Despite its widespread use, the *P* value is also among the most misunderstood and misinterpreted statistics. Common misconceptions include the belief that the *P* value

represents the probability that the null hypothesis (H_0) is true, that (1 − *P*) is the probability that the alternative hypothesis (H_1) is true, that a $P < 0.05$ indicates results of major importance or clinical relevance, and that a non-significant result ($P \geq 0.05$) implies no effect. These persistent misinterpretations prompted the American Statistical Association issued a landmark statement in 2016 to caution against the inappropriate use and reporting of the *P* value.[2] Other commentaries and reviews have also highlighted the need for more nuanced approaches to statistical evidence.[3,4]

In this review, I expand upon my earlier article[5] by integrating updated methodological and philosophical perspectives and recent literature to outline what the *P* value is and what it is not. Before doing so, however, it is useful to revisit the historical origins of the *P* value and the philosophical debates that have shaped its role in scientific inference.

THE ORIGIN

Fisher's school

Fisher[6] formalized the concept of *P* value in his 1925 seminal book, Statistical Methods for Research Workers. Fisher did not use the term null hypothesis in his first publication; instead, he used the term 'hypothesis under test' (in the same meaning as H_0), and defined the *P* value to denote the probability of obtaining a result at least as extreme as the one actually observed, if the hypothesis under test is true. Fisher viewed the *P* value as a continuous measure of evidence against H_0—a tool for inductive inference. In practice, a smaller *P* value suggested stronger evidence against the null hypothesis. Importantly, Fisher[6,7] did not fix a universal threshold for significance but considered 0.05 a convenient convention, not a rule.

Neyman-Pearson's school

In 1928, Neyman and Pearson[8,9] introduced a new framework for statistical inference, the so-called 'statistical inference tests of hypothesis,' to improve upon Fisher's significance testing.

Unlike Fisher, who considered only a single hypothesis (H_0), they proposed two competing hypotheses—the null (H_0) and the alternative (H_1). A test statistic (e.g., z, χ^2, or t) is computed and compared with a predefined critical value (e.g., $|z| > 1.96$). If the statistic exceeds this threshold, H_0 is rejected; otherwise, it is retained.[10,11]

This approach does not interpret the magnitude of the test statistic as evidence but uses it as a decision rule, making the process deductive, a behavioral choice to reject or retain H_0. Each statistic corresponds to a *P* value, but its explicit calculation is unnecessary in this framework. Neyman and Pearson also introduced the concepts of type I (α) and type II (β) errors, representing false rejection or false retention of H_0, respectively, the foundational ideas still central to hypothesis testing today.[10,11]

In clinical research, Fisher's view aligns more with studies exploring whether data are compatible with a null hypothesis, while the Neyman-Pearson approach fits more with confirmatory trials requiring explicit decision rules.

https://doi.org/10.3346/jkms.2025.40.e321

CASE STUDY

Suppose that in a randomized clinical trial, we examined the effect of a certain drug on reducing the serum cholesterol and found that the measured serum cholesterol concentration had a mean of 205 (standard deviation [SD], 16) mg/dL in the treatment group, and 216 (SD, 11) in the control group. Suppose that each group consisted of 20 individuals. Using the Student's *t*-test to compare the means of the two groups, revealed a calculated *t* statistic of 2.53—corresponding to a *P* value of 0.016—which exceeded the critical *t* value of 2.03.

REPORTING

Fisherain view

A researcher who follows Fisher would state that the hypothesis under test is that "the drug has no effect and that the mean cholesterol concentration does not differ between the treatment and control groups." The researcher then would state that given the calculated *P* value of 0.016, there is moderate evidence against the hypothesis under test that the drug is ineffective. The researcher neither rejects nor accepts the hypothesis.

Neyman-Pearson view

A researcher who follows Neyman-Pearson school would say that the H_0 is that mean cholesterol concentration is equal in the treatment and control group; the alternative hypothesis, H_1, would then be that the means are not equal (the mean in one group is lower than that in another group). Then, the researcher needs to define a statistical significance level, e.g., $\alpha = 0.05$. This is the highest acceptable probability of a type I error, to reject the H_0, when it is true. As the calculated *t* statistic of 2.53 exceeds the critical value of the statistic, 2.03, the researcher decides to reject the null and states that the means are different and because the mean cholesterol in the treatment group is less than that in the control group, the researcher would decide that the drug is effective.

The way we report the results today

We would state that the drug significantly ($P = 0.016$) decreased the serum cholesterol by 11 (95% confidence interval [CI], 2.2–19.8) mg/dL. But, given the CI, we concluded that the drug effectiveness is inconclusive.

DISCUSSION

The *P* value is the most frequently reported (and arguably the most misunderstood) statistic in biomedical research.[5] It is not the probability that the null hypothesis is true, nor that the alternative is false. It also does not indicate the probability of obtaining similar results in another study. A small *P* value (e.g., $P < 0.05$) does not necessarily reflect an important or clinically relevant effect, while a non-significant one does not imply no effect. Instead, the *P* value represents the probability of observing a result equal to or more extreme than that observed, assuming the null hypothesis is true.[12]

The difference between the schools originates from their points of views. While Fisher saw the above-mentioned case study as a single study and using inductive rules tried to make the conclusion about the level of evidence against the null hypothesis, Neyman and Pearson

https://doi.org/10.3346/jkms.2025.40.e321

saw the study as one in a series of replicas and tried to make a deductive conclusion based on the findings of this study, as an instance of replicas of this study design. Suppose that we are going to perform the above-mentioned case study 10,000 replicas under two conditions: if H_0 is true, and if H_0 is false. We then have 10,000 *P* values for each condition. The values are distributed evenly when the H_0 is true and the *P* value is < 0.05 in only 5% of conditions, the α level, which is a type I error; rejecting the H_0 when it is really true (**Fig. 1A**). On the other hand, if H_0 is really false, in most replicas (here, around 70%) the *P* value is < 0.05 (**Fig. 1B**). In remaining 30% of cases, we retained the H_0 (as the $P \geq 0.05$), the β level in this study design, which is a type II error, retaining H_0, when it is really false. With the current design ($n = 20$ in each arm, SD ≈ 16, and effect size of 11 mg/dL), we were able to identify the effect in 70% of cases (1 − β), which is technically termed the 'study power.'

In the case study, a Fisherian researcher would interpret a *P* value near 0.016 as moderate evidence against H_0, without explicitly accepting or rejecting it. By contrast, a Neyman-Pearsonian analyst, using a pre-set α = 0.05 and β = 0.30, would reject H_0 if the test statistic

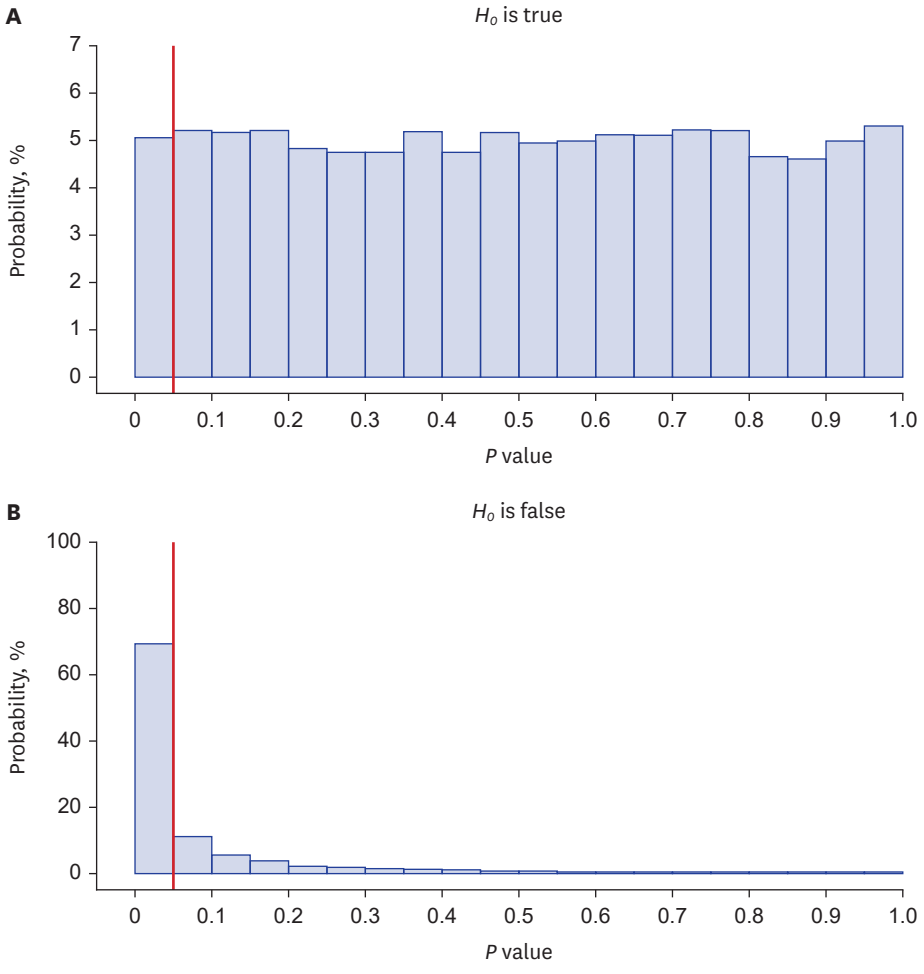

Fig. 1. Distribution of *P* values from 10,000 simulated replications of the case study under two scenarios: (**A**) when H_0 is true; and (**B**) when H_0 is false. The red vertical line marks the conventional significance threshold ($P = 0.05$). The simulations illustrate type I error (α), type II error (β), and statistical power (1 − β).

https://doi.org/10.3346/jkms.2025.40.e321

exceeded the critical value, yielding 70% power (1 − β). The two frameworks thus differ epistemologically: Fisher viewed the *P* value as an inductive measure of evidence, whereas Neyman and Pearson treated it as a deductive decision rule prone to type I (α) and type II (β) errors.[10,11]

Modern statistical reporting reflects a hybrid of both philosophies. Many journals require the exact *P* value (Fisherian), alongside explicit α levels and study power (Neyman-Pearsonian).[13-15] However, this formalism can mask deeper problems: many health-care providers lack the training to critically appraise statistical reporting and must therefore rely on the authors' interpretations.[16] Such dependence heightens the risk of misinterpretation and, when combined with incentive structures that reward significant results, fuels *P*-hacking and publication bias. These systemic pressures contribute to the reproducibility crisis noted by Ioannidis[17] and Smaldino and McElreath.[18]

Statistical significance merely indicates that an observed effect is unlikely under the null hypothesis, whereas clinical significance concerns whether the effect is meaningful for patients or practice. The two are not synonymous. Clinical relevance can be quantified using concepts such as the minimal clinically important difference, which should complement *P* values when interpreting results.

To counter these issues, many methodologists now recommend shifting focus from dichotomous significance to estimation and evidence synthesis.[2,4,19] Results should emphasize the magnitude and precision of effect estimates using 95% CIs and, where relevant, likelihood ratios (LRs), particularly in diagnostic studies.[20] For example, the drug in the case study decreased serum cholesterol by 11 (95% CI, 2.2–19.8) mg/dL. Although the upper CI limit suggests potential clinical benefit, the lower bound indicates uncertainty, implying that larger, adequately powered studies are needed for firm conclusions.

Ultimately, reliance on arbitrary thresholds has contributed to the replication crisis. Reporting effect sizes with CIs, complementing them with LRs, and fostering open-science practices, e.g., preregistration and data sharing, can mitigate bias and restore trust in research outcomes.[21-23]

FUTURE LINES OF RESEARCH

Many articles published over the past several years have shown that in research studies, a $P < 0.05$ can easily be found merely by chance.[21-23] This would in turn lead to low replication rates of the studies.[17,18,24] Some authors have proposed a lower threshold for the α.[25-28] For instance, Ioannidis[27] has proposed to set the cut-off to 0.005, rather than the conventional well-accepted value of 0.05, to lower the potential type I error rate of the results. McCloskey and Michaillat[29] proposed an α level 0.01 to correct for *P*-hacking. I, along with other researchers, have suggested using a flexible context-dependent α threshold.[12,24,30,31] This would result in more reproducible studies with lower type I and type II errors,[24] but I found that this approach, although theoretically feasible would surface an internal conflict within the frequentist statistical framework.[12,24,30] Probably, it would be better to employ other available statistical methods, e.g., Bayesian statistics, although it is not flawless too.

https://doi.org/10.3346/jkms.2025.40.e321

CONCLUSION

If one wants to report the *P* value, he needs to be aware of its meaning and employ it wisely, as it might be misleading, particularly for those readers who think they know the meaning of the statistics. Probably, it would be better to communicate the results of our study in a more comprehensible way, through the magnitude of the observed effect and its 95% CI or through the LRs, to name only a few. All journal editors should pay attention that doing correct research is not enough. Putting the results across the audience so that they can correctly construe the findings is also of paramount importance. Ultimately, replacing mechanical reliance on *P* values with careful reporting of effect sizes, CIs, and complementary approaches will help restore clarity and trust in scientific research.

REFERENCES

1. Chavalarias D, Wallach JD, Li AH, Ioannidis JP. Evolution of reporting *P* values in the biomedical literature, 1990-2015. *JAMA* 2016;315(11):1141-8. PUBMED | CROSSREF
2. Wasserstein RL, Lazar NA. The ASA statement on *p*-values: context, process, and purpose. *Am Stat* 2016;70(2):129-33. CROSSREF
3. Goodman SN. Toward evidence-based medical statistics. 1: the *P* value fallacy. *Ann Intern Med* 1999;130(12):995-1004. PUBMED | CROSSREF
4. 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(4):337-50. PUBMED | CROSSREF
5. Habibzadeh F. Credibility of the *P* value. *J Korean Med Sci* 2024;39(21):e177. PUBMED | CROSSREF
6. Fisher RA. *Statistical Methods for Research Workers*. Edinburgh, Scotland: Oliver & Boyd; 1925.
7. Fisher RA. *Statistical Methods and Scientific Inference*. 2nd ed. Edinburgh, Scotland: Oliver and Boyd; 1959.
8. Neyman J, Pearson ES. On the use and interpretation of certain test criteria for purposes of statistical inference part I. *Biometrika* 1928;20A(1-2):175-240. CROSSREF
9. Neyman J, Pearson ES. On the use and interpretation of certain test criteria for purposes of statistical inference. *Biometrika* 1928;20A(3-4):263-94. CROSSREF
10. Goodman SN. p values, hypothesis tests, and likelihood: implications for epidemiology of a neglected historical debate. *Am J Epidemiol* 1993;137(5):485-96. PUBMED | CROSSREF
11. Hubbard R, Bayarri MJ. Confusion over measures of evidence (p's) versus errors (α's) in classical statistical testing. *Am Stat* 2003;57(3):171-8. CROSSREF
12. Habibzadeh F. On the use of receiver operating characteristic curve analysis to determine the most appropriate p value significance threshold. *J Transl Med* 2024;22(1):16. PUBMED | CROSSREF
13. Habibzadeh F. Statistical data editing in scientific articles. *J Korean Med Sci* 2017;32(7):1072-6. PUBMED | CROSSREF
14. Habibzadeh F. How to report the results of public health research. *J Public Health Emerg* 2017;1:90. CROSSREF
15. Lang TA, Altman DG. Basic statistical reporting for articles published in biomedical journals: the "Statistical Analyses and Methods in the Published Literature" or the SAMPL guidelines. *Int J Nurs Stud* 2015;52(1):5-9. PUBMED | CROSSREF
16. Misra DP, Zimba O, Gasparyan AY. Statistical data presentation: a primer for rheumatology researchers. *Rheumatol Int* 2021;41(1):43-55. PUBMED | CROSSREF
17. Ioannidis JP. Why most published research findings are false. *PLoS Med* 2005;2(8):e124. PUBMED | CROSSREF
18. Smaldino PE, McElreath R. The natural selection of bad science. *R Soc Open Sci* 2016;3(9):160384. PUBMED | CROSSREF
19. Aguinis H, Vassar M, Wayant C. On reporting and interpreting statistical significance and p values in medical research. *BMJ Evid Based Med* 2021;26(2):39-42. PUBMED | CROSSREF

20. Habibzadeh F, Habibzadeh P. The likelihood ratio and its graphical representation. *Biochem Med (Zagreb)* 2019;29(2):020101. PUBMED | CROSSREF
21. Bem DJ. Feeling the future: experimental evidence for anomalous retroactive influences on cognition and affect. *J Pers Soc Psychol* 2011;100(3):407-25. PUBMED | CROSSREF
22. Carney DR, Cuddy AJ, Yap AJ. Power posing: brief nonverbal displays affect neuroendocrine levels and risk tolerance. *Psychol Sci* 2010;21(10):1363-8. PUBMED | CROSSREF
23. McShane BB, Gal D, Gelman A, Robert C, Tackett JL. Abandon statistical significance. *Am Stat* 2019;73(Suppl 1):235-45. CROSSREF
24. Habibzadeh F. On the effect of flexible adjustment of the p value significance threshold on the reproducibility of randomized clinical trials. *PLoS One* 2025;20(6):e0325920. PUBMED | CROSSREF
25. Benjamin DJ, Berger JO, Johannesson M, Nosek BA, Wagenmakers EJ, Berk R, et al. Redefine statistical significance. *Nat Hum Behav* 2018;2(1):6-10. PUBMED | CROSSREF
26. Greenwald AG, Gonzalez R, Harris RJ, Guthrie D. Effect sizes and p values: what should be reported and what should be replicated? *Psychophysiology* 1996;33(2):175-83. PUBMED | CROSSREF
27. Ioannidis JPA. The proposal to lower *P* value thresholds to .005. *JAMA* 2018;319(14):1429-30. PUBMED | CROSSREF
28. Johnson VE. Revised standards for statistical evidence. *Proc Natl Acad Sci U S A* 2013;110(48):19313-7. PUBMED | CROSSREF
29. McCloskey A, Michaillat P. Critical values robust to p-hacking. *Rev Econ Stat* 2024:1-35. CROSSREF
30. Habibzadeh F. Reinterpretation of the results of randomized clinical trials. *PLoS One* 2024;19(6):e0305575. PUBMED | CROSSREF
31. Lakens D, Adolfi FG, Albers CJ, Anvari F, Apps MAJ, Argamon SE, et al. Justify your alpha. *Nat Hum Behav* 2018;2(3):168-71. CROSSREF