Research Article **ISSN: 3079-6059**

The P-Value Conundrum Navigating the Nuances of Statistical Significance

Authors

Temba Munsaka (PhD)* Africa Research University

Corresponding Author:

Temba Munsaka (PhD)* Africa Research University.

Submission:18/11/2024, **Published:** 09/02/2025

Citation: Munsaka, T. (2025), The P-Value Conundrum Navigating the Nuances of Statistical Significance. AJIESS 1 (2):1-10

Abstract

This study seeks to fundamentally question the relevance of p-values in contemporary science, accentuating the current contention as well as potential misinterpretation, which is almost synonymous with their usage. The study takes the approach of a “review of literature,” using a variety of academic sources to provide a comprehensive overview of

the P-Value problem. It delves into history, foundational statistical concepts, and controversies that are still causing hot debate across a multitude of disciplines in relation to p-values.

The analysis demonstrates that the p-values are more complex than we would like to be as a single statistic and lead closely to incorrect conclusions, including the dangers of straying from the use of arbitrary significance thresholds. It also covers different statistical approaches, and a growing interest in effect sizes, confidence intervals, and the need for contextual interpretation.

The implications of this study are profound for the scientific community, policymakers, and researchers. Such evidence highlights the importance of a deeper understanding of P-values and urges a more rigorous approach to statistical analysis to provide valid and reliable evidence from scientific research. We provide an overview of the upcoming states of the P-value debate through most of the current literature, giving proper context to the conversation on the role of statistical significance.

Keywords: statistics, hypothesis, Significance Interval, Type 1 and 2 error, standard error, test statistics.

Introduction

P-value, a staple in modern scientific research, has been a contentious issue for a long time. The p value for statistical significance is widely used as a guide for statistical hypothesis testing. However, despite their popularity, the use of P-values has come under fire, with some expressing concerns about the potential for P-values to be misused or misinterpreted. The idea behind the p-value is simple: it tells you how likely you would be to see a test statistic as extreme as that observed by random chance if the null hypothesis is actually true. Essentially, the P-Value provides a measure for researchers to assess the

chance that their results are spontaneous, given that there is no effect or difference that actually exists. The acceptance of p as statistically significant when it is below the traditional threshold of 0.05 Phd is widely interpreted as strong evidence against the null hypothesis, with the following conclusion: the observed effect is statistically significant.

However, this apparent simplicity hides deep-seated complexities and pitfalls in the p-value paradigm. P-values, sample size, effect size, and statistical versus practical significance have been ongoing concerns for researchers for decades in terms of their interpretation and implications. Furthermore, the common use of a significance level of 0.05, as a de facto standard has also been criticised for its arbitrary nature, as well as for the odds of an inflated false-positive rate.

The P-Value problem has received attention in recent years, with increasing calls for a more critical and nuanced discussion of statistical inference. IMPORTANT: Such abuse generated many criticisms of P-values, and major scientific societies such as the American Statistical Association (ASA)issued warnings urging people to avoid misinterpretation and misuse of P-values. They also supported the introduction of a wider variety of statistical techniques, such as effect sizes, confidence intervals, and Bayesian methods, for reporting experimental data. In this paper, we will try to deliver at the core of the P-Value complexity, exploring the spectrum of reasons and questions for the debate and its impact on science and practice. By examining the historical backdrop, statistical underpinnings, and evolving options, this study attempts to provide a fair and knowledgeable perspective on how P-values fit into the quest for scientific discovery and enhance our understanding of the natural phenomena surrounding us.

Methodology

A comprehensive literature review was performed as part of this study's willingness to conduct a distinct process of multidimensional literature review of the P-value problem. In selecting articles, we were motivated by the common goal of presenting a balanced and nuanced analysis of this complex and

controversial issue, examining its historical background, statistical underpinnings, ongoing debates and criticisms, and proposing alternatives. Description of the search strategy: Articles were retrieved according to the inclusion criteria within a range of academic sources by articulating several descriptors. Key search terms were "P-Value," "statistical significance," "null hypothesis," "effect size," and "Bayesian analysis." These keywords were grouped and fine-tuned to reflect the depth range in the p-value problem.

The literature search was performed using several electronic databases, including the Web of Science, Scopus, and Google Scholar. Issuing this cross-database approach was critical to ensuring holistic coverage of the topic, as various databases tend to have their own means of indexing and cataloguing scholarly materials. By searching across these diverse platforms, the review identified a range of peer-reviewed journal articles, statistical textbooks, and position papers from leading scientific organisations. In the first stage of the procedure, the titles and abstracts of the recovered literature were filtered to evaluate their eligibility for the research question. This screening cascade included applying pre-determined eligibility criteria, focusing on the concept of P-value, discussion of statistical significance, hypothesis testing, and coverage of alternative statistical approaches. Studies that failed to meet these criteria were excluded.

We then conducted a detailed and structured analysis of the selected literature, targeting the extraction and synthesis of salient themes, debates, and trends evident in the use and interpretation of P-values within scientific research. The focus of the analysis also included the historical evolution of the P-value concept, including the initial contribution of statisticians such as Ronald Fisher and Jerzy Neyman.

Furthermore, the review took a deep dive into the statistical principles that underpin the calculation and interpretation of p-values, covering the details of hypothesis testing, significance levels, and

reasoning behind statistical inference in a very basic way. To appreciate the minutia and depth of the p-value problem, we needed to go through this deep dive from statistical theory to develop an understanding of where its pitfalls live. We spend a great deal of our analysis on the ongoing debates and criticisms that p-values face within the scientific community. They reviewed common misinterpretations and misuses of P-values (e.g. a P-value that is low must mean that there is a high effect size, or that a P-value that is not significant means that there is no effect). These fallacies have even been pointed out as reasons why people unnecessarily rely on p-values to make decisions.

Moreover, the review examined the increasing awareness regarding the drawbacks associated with the widespread use of 0.05 significance level as a de facto standard. We examined the arbitrariness of this threshold and the risks of inflated false-positive rates, highlighting the importance of a more context-sensitive approach to statistical significance. Alongside the scrutiny of the P-value issues, this review also examined some of the alternative statistical methods that use the spotlight in the scientific community. This includes the use of effect sizes, confidence intervals, and Bayesian analysis, which present a larger and more contextual understanding of the results of research instead of the simple binary that we see in the "significant/not significant" dichotomy.

To maintain the strength and rigor of the review, the author critically evaluated the reviewed studies, taking into account the research design, methodological rigor, and credulence of the sources. The critical appraisal process entailed a self-assessment of the overall methodological quality of the studies, the suitability of the statistical analyses used, and transparency and clarity of reporting.

The findings were synthesised with the goal of providing a balanced and nuanced view of the P-value conundrum. This process involved critically scrutinising the different perspectives and arguments within the research literature, recognising the complexities and nuances of the subject matter, and remaining a thoughtful and

dispassionate observer throughout the review process.

This review also explores the possible implications of the P-value conundrum for researchers, policymakers, and the broader scientific community. The synthesis of these paradigms examined how a more critical, nuanced understanding of P-Values could influence research practices, reporting, and evidence for decision-making in the scientific community, thus advancing scientific progress and positively impacting society as a whole.

Through this thorough and methodical review of the literature, the authors aimed to create an invaluable guide for those looking to state their thirst to understand the complexities of the p-value problem. The synthesis of current knowledge is invaluable and can drive further work, discussion, and the establishment of better statistical practices within the scientific community.

Results

The literature review on the P-value conundrum is as follows.

The Rise of P-Values

The history of P-Values goes back to the 1920s and 1930s when Ronald Fisher, in his work, laid the foundation for their extensive use in scientific research. Fisher first introduced the p-value as a quantitative measure of evidence against the null hypothesis in tests (Fisher, 1935). P-values, such as the probability of achieving data as extreme as it was observed, assuming that the null hypothesis is true.

Fisher's p-value and significance testing ideas were key foundational developments leading to modern statistical techniques, such as those used in scientific research. With its appealing interpretation and the widely adopted 0.05 threshold for “statistical significance,” the P-Value became the primary currency for drawing conclusions from empirical research.

A world built on historical context and statistical foundations

The P-Value in the hypothesis testing framework has a rich history, dating back

to the work done by statisticians, including Ronald Fisher and Jerzy Neyman, who laid the groundwork for the hypothesis testing framework and significance levels. Nevertheless, the concept of P-Values was in practice and usage way before any formal definition cut off back in October 2023, so the interpretation and application of P-Values evolved along with it, giving rise to heated debates and misconceptions.

The P-Value is Often Misinterpreted and Misapprehended

Despite their limitations, p-values are often misinterpreted, leading to incorrect implications. One of the most frequent misinterpretations is the equivalent statistical significance ($p < 0.05$) of the probability of the null hypothesis being true. As Wasserstein et al. As Lee and Hsu, (2019) explain, this is a false dichotomy – a P-Value does not tell us in any way the probability the null hypothesis is true. The second common manner in which they are misconstrued is the "dichotomous" interpretation of P-Values, where they are used to classify results as either "significant" or "non-significant", based on arbitrary cut-offs. This approach is blind to the fact that P-values are continuous, and all statistics are uncertain (Trafimow & Marks, 2015).

Additionally, the common practice of p-hacking, emphasising positive results, exacerbates the problem, as shown by Ioannidis (2005). With publication bias, the tendency for positive findings to be published more often than negative or null results can lead to scientific literature that paints a skewed picture of reality.

The Limitations of P-Values

Despite their ubiquitous use, p-values have (many) well-documented limitations that put into question their utility as the sole basis for scientific inferences. These limitations are important for assessing the research outcomes and making proper judgments regarding the findings. P-values do not indicate the probability of the null hypothesis being true. According to Ioannidis (2005), the probability that a research finding is true can vary with the prior probability of the effect, statistical power of the study, and extent of bias. The p-values provide an imperfect and limited perspective.

P-values are highly sensitive to sample size. Using sufficiently large samples, we can always find "statistical significance" whenever we want, even for trivial effects ($p < 0.05$). However, small studies (Nuzzo, 2014) may have overlooked important effects. However, this mismatch between statistical and practical significance leads to the over-interpretation of positive results and exclusion of potentially meaningful negative findings. P-values did not describe the size or clinical importance of an effect. If the effect size is smaller than the statistically significant result, it may have little practical importance. p-values can lead to attention away from the fact that the observed effect might be the most relevant question (McGough and Faraone, 2009).

P-values can indicate the significance, but they do not indicate the direction of the effect. They do not speak of the nature of the relationship; they simply tell us that the observed effect is unlikely to be attributable to chance under the null hypothesis. P-Values implicated in the "file drawer problem" and publication bias. A potential bias could arise if researchers were more likely to publish studies with statistically significant results, which has been shown (Ioannidis, 2005; Young et al., 2008) and often negative or inconclusive findings remain unpublished. There will be an overrepresented number of positive results in the published literature.

However, there are many questionable research practices that can manipulate p-values, such as p-hacking (selective reporting analyses that yield significant results) and HARKing (hypothesising after the results are known) (Gelman & Loken, 2014). Such practices can lead to an exaggerated rate of false positive results. The P-Value does not indicate whether a finding is replicable/robust. Moreover, a statistically significant finding in one study is not necessarily replicated in subsequent studies, particularly in domains with small effect sizes and low statistical power (Ioannidis 2005).

Novel alternatives including statistical approaches

Several alternatives to p-values have been proposed because of their drawbacks and misuse. These alternatives were intended to be more

informative and nuanced than the binary "significant/non-significant" dichotomy. An alternative proposed in guidance documents from the American Educational Research Association (2008) and the American Psychological Association (1990) emphasises effect sizes and their confidence intervals: estimates of effect size that quantify within-sample uncertainty in the observed effect (McGough & Faraone, 2009).

Confidence intervals should be provided as they can deliver useful information about the range of values for the effect size, which is consistent with the data and facilitates the interpretation of the practical significance of the findings. This guides the consideration of large versus small effects, rather than statistical significance alone. An alternative solution is to use Bayesian methods, which offer a more straightforward and easily interpretable assessment of one hypothesis being more likely than another given the data (Gelman & Loken, 2014). Bayesian methods allow for the inclusion of prior knowledge, and often provide better insights into the uncertainty associated with research findings. Bayesian analysis can generate probabilities for the hypotheses of interest (rather than P-Values), which can help researchers and readers to better understand the relative support for the drift hypotheses.

Instead of the outcome of a single study, there have also been calls for a more inclusive assessment of the strength of evidence, including the replication of findings, effect sizes, and plausibility of models (Wasserstein and Lazar, 2016; Nuzzo, 2014). This means that no one study should be relied upon exclusively, even a study with a statistically significant P-value, and the evidence provided by multiple studies should be considered to reach conclusions and make decisions.

Some researchers have called for abandoning the significant/non-significant dichotomization of p-values, a more continuous, uncertain perspective on research results (Trafimow & Marks, 2015), and the need for a more sensible approach. Given that statistical analysis always comes with some level of uncertainty, this would enable us to represent findings and interpret results with more detail

concerning the surrounding level of uncertainty, the size of the effect, and the overall strength of the evidence through how they are presented. They represent progressive methods addressing the weaknesses and abuse of P-values, providing a clearer and finer analysis of results and prompting a more contextual and nuanced look at the evidence at hand.

Conclusions: This review highlights the critical need for researchers to consider the social dynamics of scientific practice, engage in open dialogue with the public, and prioritise transparent practices. They highlight the importance of developing a more critical understanding of P-Values and choosing to adopt better statistical practices to ensure that research findings are valid and reproducible. [2] transparent reporting, open data sharing, collaborative work to address the P-Value problem

Bailey and Oppenheimer (2021) Recommendations for Improving Statistical Practices and Communication. This should involve a multipronged approach to deal with P-value issues and enhance the reliability and integrity of scientific research. First, it should move from excessive dependence on statistical significance to a broader estimation and interpretation of effect sizes and their related uncertainty (Wasserstein et al., 2019). This entails encouraging the reporting of effect sizes, confidence intervals, and other indices that offer more refined and informative measures of strength of evidence.

In addition, Bayesian methods, rather than frequentist approaches, should be promoted because they offer a more intuitive and direct interpretation in terms of the probabilities of hypotheses, given what we observe in the data (Gelman & Loken, 2014). Bayesian methods have the advantage of integrating prior knowledge and can help researchers and consumers to better understand the uncertainty underlying research results. Furthermore, the totality of evidence should be considered rather than the results from only one study (Ioannidis, 2005). Which means taking into account things like replication, consistency of findings across studies, plausibility of the proposed hypotheses (which is certainly not the same as saying that the p-values

of individual results aren't statistically significant).

Second, improving statistical education and training for researchers and readers of scientific literature is also fundamental, as misuse and misinterpretation by multiple players in our scientific environment is well documented (Gagnier & Morgenstern, 2017). Improving the knowledge of suitable statistical ideas and principles can reduce the basic gaps in science, leading to issues with P-value misuse. An important part of catalysing these changes lies with journal editors and reviewers, who should stop demanding focus on 'statistical significance alone" and instead adopt a comprehensive evaluation of study results, including important factors such as effect sizes, confidence intervals, and evidence strength (Wasserstein et al., 2019).

The registration of study designs, opening data, and coding to the public also need to be promoted in scientific research (Nosek et al., 2018). This approach is a useful way to reduce the impact of dubious research practices, such as p-hacking and selective outcome reporting, and to increase the generalisability and integrity of the scientific process.

Developing guidelines and standards for reporting research findings, such as the use and interpretation of p-values, effect sizes, and other statistics, can also help with more consistent and transparent presentations (Wasserstein and Lazar, 2016). Replication is of utmost importance, and it must be emphasised that research findings need to be evaluated for their robustness over time and not be treated as conclusive based on a single statistically significant result. Instead, scientific conclusions should be based on evidence from multiple high-quality studies.

Finally, having a culture of scientific humility or researchers' acknowledgement of the limitations of their methods, the uncertainty of their findings, and the scientific investigation and understanding need to continue adjusting can help curb researchers' overstatement of the implications of their findings (Gelman, 2016). A system-wide strategy to eliminate the shortcomings of P-values and prevent their abuse, the

ultimate outcome of which will be to scale the degree of scientific enquiry, data integrity, and clinical applicability of the findings.

The implementation of these recommendations will require tireless efforts by researchers, journal editors, funding agencies, and the scientific community. It will take a realignment of incentive structures and norms underlying scientific practice to make this happen, which will require an emphasis on quality, rigor, and transparency in the scientific literature over the importance of statistically significant results.

The role of journal editors and reviewers

The journal editor and peer reviewers shaped the scientific literature and communication of the research findings. To solve the P-Value problem, and these gatekeepers of publication are a solution instead of an issue, there are several changes that need to be made directly by editors, reviewers, and journals. These senders of scientific papers can practice a number of guidelines to prevent these P-values from becoming gatekeepers of scientific publications.

First, editors and reviewers must fight against the overuse of p-values and the significant/non-significant binary. The authors will compare the presented results with previous results, suggesting that they should present and discuss their findings in terms of standardised effect sizes, confidence intervals, and the number and nature of included studies, rather than simply emphasising P-values. This will allow for a more comprehensive and nuanced understanding of research findings. At the same time, editors and reviewers should encourage the reporting of effect sizes and confidence intervals, in addition to P-values. This can assist readers in appreciating the practicality of the findings, which go far beyond the simplistic reading of statistical significance.

To this end, editors and reviewers should be careful when considering the publication of studies that only report statistically significant observations given that they are both frequently observed, particularly in fields with small effect sizes and low statistical power. They

need to account for the possibility of publication bias and assess the totality of evidence, including negative or null results.

To promote transparency and accountability, study designs should be preregistered and data and analysis codes shared (6–9 9) whenever possible, and editors and reviewers should encourage this. This may help lessen the voice of negative research practices (e.g. p-hacking, HARKing). Editors and reviewers should also formulate and implement guidelines for reporting statistical methods and results based, for example, on CONSORT and STROBE statements. These guidelines can assist in providing consistency and clarity to communicate the research results.

Editors and reviewers should also be trained and educated on the proper use and interpretation of p-values, effect sizes, and other statistical concepts to better assess the quality and robustness of the research they are charged with appraising. Finally, we need editors and reviewers who are willing to publish articles that question the status quo and highlight places as P-values, and other statistical practices fall short of good science. One important piece of their reform program is ensuring that they drive changes in how science strengthens and shares its methods of data analysis and interpretation. Thus, journal editors and reviewers can contribute to changing the reliance of the scientific community on p-values and a more sophisticated and thorough approach to the evaluation and communication of research findings.

All Are Necessary and Important: Improving Statistical Education and Training

Many of the P-Value myths and the widespread abuse and misinterpretation of P-Values arise due to the lack of proper statistical education of researchers. Despite a growing need, many scientists receive little initial training in statistical methods and concepts, and often retain this knowledge gap well in their careers.

A concerted effort to focus on better statistical education and training in the sciences could help mitigate the limitations and misuse of P-values. Incorporating statistical thinking and data

analysis skills into the core curriculum for undergraduate and graduate programs is vital for imparting students a strong foundation in statistical principles and their appropriate application (Garfield, 1995; Hoekstra et al., 2014). Similarly, dedicated classes and workshops on specific statistical techniques, data analysis, and the interpretation of research findings can also be helpful. Alternatively, these offerings must be open to researchers' needs in diverse disciplines and challenges (Tishkovskaya and Lancaster 2012). rather than the one-size-fit approach used in most classical statistical studies.

Integrating active learning techniques, such as case studies, hands-on data analysis exercises, and group discussions, can also facilitate taking on another level of understanding of statistical concepts and their practical applications (Chance, 2002; Nolan & Speed, 1999).

By focusing on the reasoning needed to interpret research findings, not merely on the arithmetic of statistical tests, researchers can gain skills needed to judge quality, limitations, and implications for politics and practice, including the proper use and interpretation of P-Values (Rumsey, 2002). However, to be effective in building skills and confidence, researchers need professional development support via regular webinars, workshops, and mentoring programs to keep them current with new statistical techniques, analyses, and interpretation of data (Horton & Hardin, 2015). Simultaneously, collaboration with professional societies, funding agents, and others to produce and disseminate educational resources can also help to advance this (Garfield, 2002).

In conclusion, a systematic review of their validity and limitations through educational training will encourage a change in culture towards statistical literacy and critical thinking in the scientific community, where researchers are aware of their responsibility for the proper implementation of statistical methods and interpretation of results (Gal, 2002). With these educational efforts for the scientific community, double checks can be performed to strengthen the trust and authenticity of

scientific research. The solution in this regard is that the scientific community should invest in better statistical education and training, which can bridge the fundamental knowledge gaps that allow for the misuse and misinterpretation of P-values, thus enhancing the reliability and integrity of scientific research.

Conclusion

The P-Value conundrum: A hotly debated and heavily criticised Catch-22 with layers and intricacies that underpin scientific research. Explaining controversies behind the use and interpretation of P-Values: A literature review.

The findings of this study underscore the nuances of p-values, their vulnerability to misinterpretation, and the dangers associated with excessive dependence on arbitrary significance thresholds. It also covers an increasing focus on alternative statistical practices, including effect sizes, confidence intervals, and Bayesian analysis, each of which provides a more accurate and comprehensive context for research findings. However, the ramifications of this work are broad; they highlight the importance of a more careful and less-promiscuous approach to using P-values in scientific studies.

The P-value problem requires attention from researchers, policymakers, and the broader scientific community to promote the further adoption of more cue measures, improved statistical practices, and general collaborative and collective efforts to encourage transparent reporting. This can further help to promote a more nuanced understanding of the significance of P-Values, depth in understanding, as well as offer a valuable reference point for researchers and individuals who wish to understand the intricacies of statistical significance in their scientific work and to enhance their knowledge about the world we live in.

Acknowledgements

No human or animal studies have been conducted in this study. This conceptual paper discusses the existing literature on the P-value conundrum. Although no primary data collection was performed, ethical approval was not required. All sources were accurately

cited and referenced based on APA guidelines.

Competing Interests

The authors declare that they have no financial or personal relationships that may have inappropriately influenced the writing of this paper.

Funding

No financial support was received for the research, authorship, or publication of this article.

Data Availability

No new data were created or analysed in this study; therefore, data sharing was not applicable to this article.

Disclaimer

This article reflects the author's own opinion, not that of any institution or funder.

References

American Statistical Association. (2016). ASA statement on statistical significance and p-values. The American Statistician, 70(2), 129-133.

Cumming, G. (2014). New Statistics: Why and How Psychological science, 25(1), 7-29.

Gellman, A. & Carlin, J. (2014). Beyond power calculations: Assessing type S (sign) and type M (magnitude) errors. Perspectives on Psychological Science, 9(6), 641-651.

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., & Altman, D. G. (2016). Statistical tests, P values, confidence intervals, and power: a guide for misinterpretation. European journal of epidemiology, 31(4), 337-350.

Ioannidis, J. P. (2005). Why are most published research findings false? PLoS medicine, 2(8), e124.

Kline, R. B. (2013). Beyond significance testing: Statistics reform in behavioural sciences. American Psychological Association [Internet].

Nuzzo, R. (2014). Scientific method: statistical errors. Nature, 506(7487), 150-152.

Wasserstein, R. L., & Lazar, N. A. (2016). ASA statement on p-values: context, process, and purpose. The American Statistician, 70(2), 129-133.

McGough JJ and Faraone SV. Estimating the size of treatment effects: moving beyond p-values. Psychiatry (Edgmont) 2009;6(10):21–29.

Hubbard R, Lindsay RM. Why are P-values not a useful measure of evidence in statistical significance testing? Theory Psychol 2008;18(1):69–88.

Trafimow D, Marks M. Editorial. Basic and Applied Social Psychology 2015;37;1–2.

Wasserstein RL Lazar NA. ASA Statement on P-Values: Context, Process, and Purpose. Am Stat 2016;70:129–133.

Nuzzo R. Scientific method: Statistical errors. Nature 2014;506:152–156.

Ioannidis JP. Why are most published research findings false? PLoS Med. 2005;2(8):e124. Wasserstein R.L, Schirm AL, Lazar NA. Moving to a world beyond "p<0.05'. The AmStat2019;73(Sup1):1–19.

.Amrhein V, Greenland S, Mc Shane B. Scientists rise up against statistical significance. Nature 2019; 567:305–307.

Ioannidis JPA. What have we (not) learned from millions of scientific papers with values? Am Stat 2019;73(Sup1):20–25.

Gagnier J, Morgenstern H. Misconception, misuse, and misinterpretation of P-Value and significance testing. J Bone Joint Surg 2017;99(18):1598–1603.

Young NS, Ioannidis JPA, and Al-Ubaydli O. Why do the current publication practices distort science? PLoS Med2008;5(10):e201.