International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 03 , March - 2026

# The Use and Misuse of P-Values in Scientific Research: A Systematic Review

Akwasi Oppong
Department of Mathematics and Statistics
Northern Arizona University

**Abstract - This systematic review examines the use and misuse of p-values in scientific research. The study is based on the findings of 31 articles and other relevant studies. P-values are commonly used as a statistical tool for research studies. However, the use of p-values is often misunderstood and misinterpreted. The literature also points out common misunderstandings such as the use of p-values as an estimate of the probability of the hypothesis being true. There are also issues such as p-hacking and the use of arbitrary thresholds. There are also recent criticisms by prominent statisticians and organizations such as the American Statistical Association (ASA), suggesting the use of more informative approaches beyond the conventional p-value less than 0.05. In this review, conceptual discussions, empirical evidence, and reform suggestions are synthesized to highlight the need for urgent improvements in statistical literacy, reporting, and a paradigm shift in scientific research for increased reliability and validity.**

**Keywords: p-Values; Statistical Significance; Hypothesis Testing; P-Hacking; Effect Size; Confidence Intervals; False Positives; Null Hypothesis; Significance Thresholds**

## 1. INTRODUCTION

P-values have been an integral part of hypothesis testing in science since the formal development of the concept in the early 20th century by Ronald Fisher and subsequent development by the Neyman-Pearson framework (Biau et al., 2010). P-values are essentially a way of testing the consistency of the data with a given null hypothesis. Their use has been pervasive in various fields of study in biomedicine, psychology, and the social sciences, among others. P-values have thus become an integral part of statistical decision-making (Ioannidis, 2019).

However, in the past ten years, there has been an increasing trend of criticism of the p-value paradigm. The main argument of the critics is that the p-values have frequently been misinterpreted, overemphasized, and misused, leading to false positives, overstatements, and irreproducibility. The ASA (American Statistical Association), in their formal statement of 2016, cautioned against the dichotomous "significant/nonsignificant" threshold and the need for transparency and effect sizes. Several empirical studies have provided evidence of the widespread misuses of the p-values, including p-hacking, selective reporting, and the overuse of arbitrary thresholds (Wasserstein et al., 2019; Ioannidis, 2019).

This review aims to provide a synthesis of conceptual debates, evidence, and reform in the use of p-values. We will examine the development and use of p-values, some misconceptions and misuses, and alternative methods that may be more reliable in statistical inference.

## 2. HISTORICAL CONTEXT AND THEORETICAL FOUNDATIONS

The concept of the p-value originated from Ronald Fisher’s work in the 1920s and was intended as a way of measuring the probability of the observed data being as extreme as they are under the null hypothesis (Fisher, 1925). Fisher's idea of the p-value was as a continuous measure of evidence against the null, without strict thresholds.

Subsequently, Neyman and Pearson developed frameworks for hypothesis testing that included significance levels, critical regions, and Type I/II error probabilities. Modern methods have mixed these approaches, using p-value tests as a binary decision, "significant" if $p < 0.05$, despite their unique derivations (Biau et al., 2010).

**Key Point:**

The p-value is *not* the probability that the null hypothesis is true, and it is *not* a measure of the size or importance of an effect (Greenland et al., 2016). Misunderstanding this fundamental point underlies many of the widespread misconceptions and misuses.

## 3. CONCEPTUAL CRITIQUES AND COMMON MISCONCEPTIONS

The reliance on p-values as a main measure of statistical significance has resulted in a wide range of conceptual misunderstandings and misapplication in scientific studies. Through an assessment of the literature, including 31 articles, it has been identified that there are several themes and key issues that affect statistical integrity.

### 3.1 Fundamental Misinterpretations of p-Values

The most prominent misconception is the understanding that the p-value is an estimate of the probability that the null hypothesis is correct or that it is a direct estimate of the evidence for or against the null hypothesis. This is refuted by Cohen (2011) and Gagnier & Morgenstern (2017), who state that the only thing that a p-value is an estimate for is the probability that the data as extreme or more extreme than the current data would be observed if the null hypothesis is assumed to be correct. This is further reinforced by Greenland et al. (2016), who state that p-values should not be confused with probabilities of hypotheses or effect sizes.

### 3.2 The Dichotomous Threshold and Its Pitfalls

This arbitrary threshold of $p < 0.05$ has led to dichotomous thinking in which results are classified as "significant" or "not significant." However, this practice has been contested by Wasserstein et al. (2019), who argue that such dichotomous classification is a misrepresentation and leads to over-confidence in results that just manage to meet this criterion and to overlooking results that fail to meet this criterion but may be important in themselves. Hofmann & Meyer-Nieberg (2018) describe how this dichotomous practice leads to misinterpretation and encourages questionable research practices like p-hacking.

### 3.3 P-Hacking, Data Dredging, and Selective Reporting

Many papers, including Taroni et al. (2016) and Wang & Long (2022), have discussed the phenomenon of p-hacking, whereby the researcher tries repeatedly analyzing data until a significant p-value emerges or selectively reporting results to meet significance criteria. In this context, Lesaffre (2008) has cautioned that such practices lead to the distortion of the scientific record and result in the erosion of trust.

### 3.4 Overreliance on Significance and Neglect of Effect Sizes

Authors such as Cohen (2011), Andrade (2019), and Di Leo & Sardanelli (2020) have criticized the emphasis placed on p-value, to the neglect of effect sizes and confidence intervals. There is a risk of overstating the importance of statistically significant results while neglecting their practical significance. Greenland (2019) states that p-value needs to be considered one aspect of inference, not the only one that determines the validity of a study.

### 3.5 Misuse in Various Disciplines and Contexts

According to the literature, the misinterpretation of p-values is a common phenomenon across disciplines. For instance, Ferr (2025) highlights the misconceptions in the field of psychological studies, while Dahiru (2008) cautions against the blind use of p-values in medical studies. In the case of forensic studies, Taroni et al. (2016) pose the question of the use of p-values in the field of forensic science.

### 3.6 The Reproducibility Crisis and P-Value Limitations

The misuse and misunderstanding of p-value, as discussed by Ioannidis (2019) and Wasserstein & Lazar (2016), is also associated with the overall reproducibility crisis. The authors argue that the use of p-value less than 0.05 as a benchmark for publication is a major factor for the production of false positive results, irreproducible results, and a distorted scientific literature. This is a crisis that calls for a re-evaluation of the use of p-value in research methodology.

### 3.7 Critical Perspectives on the Validity and Utility of p-Values

However, some authors, such as Imbens (2021) and Concato & Hartigan (2016), argue that p-values are useful when properly understood but are frequently misunderstood and misinterpreted. On the other hand, authors such as Hofmann & Meyer-Nieberg (2018) propose that p-values should simply be discarded because they are conceptually incorrect and propose alternative measures.

### 3.8 Calls for Reform and Improved Understanding

Another theme that is common in the literature, as seen in the studies by Wasserstein et al. (2019), Kmetz (2019), and Biau et al. (2010), is the need to improve statistical education and reporting and to consider alternative statistical approaches such as effect sizes, confidence intervals, and Bayesian statistics.

In summary, as seen in the extensive literature, the conceptual problems and misconceptions about the use of p-values, from basic misunderstandings to incorrect research approaches, are major challenges to scientific integrity.

## 4. EMPIRICAL EVIDENCE OF P-VALUE MISUSE

The theoretical objections to, as well as misconceptions about, p-values are supported by empirical studies that show how p-values are misused and misinterpreted in scientific practices. An analysis of several studies indicates that the misuse of p-values is not just a theoretical problem but a real concern in terms of their impact on research quality and reproducibility.

### 4.1 P-Value Prevalence and Overuse in Scientific Literature

Ioannidis (2019) undertook a large-scale analysis of published articles in biomedical journals and observed that more than 96% of published articles presented a p-value less than 0.05 in their results. Such a high rate suggests that this practice is well ingrained in the culture of conducting and reporting research.

Wang & Long (2022) also undertook a large-scale analysis of published articles in biomedical journals and observed that researchers often resort to p-hacking techniques such as data dredging and selective reporting in their published articles, which increases the probability of false positives in published results.

Wang & Long (2022) observed that researchers often tend to examine their data until they obtain a statistically significant result, regardless of whether it is a false positive or not.

### 4.2 P-Hacking and Data Dredging

Several studies have also demonstrated the prevalence of p-hacking approaches. Taroni et al. (2016) have pointed out the prevalence of studies in forensic science research that have significant values without proper consideration of context or multiple testing problems. Wang & Long (2022) have also shown that p-hacking is prevalent in biomedical research, where researchers may conduct many tests or manipulate data to meet the criteria for a value of $p < 0.05$.

### 4.3 P-Value Misinterpretation in Practice

Research survey and analysis have shown that there are a number of misconceptions about p-values. According to Cohen (2011), most researchers misunderstand that a p-value less than 0.05 represents the probability that the null hypothesis is true, which is a fundamental misunderstanding. Andrade (2019) states that a large percentage of scientists have misinterpreted that p-values provide a direct estimate of the probability that hypotheses are true.

### 4.4 Impact on Reproducibility and Scientific Credibility

The overuse and misapplication of p-values have also been implicated in the reproducibility crisis. Ioannidis (2019) claims that the overuse of $p < 0.05$ is a major factor in the large number of false-positive results that do not replicate in subsequent studies. The misapplication of p-values has also been implicated by Wasserstein & Lazar (2016) and other researchers as a major factor in a scientific system dominated by non-reproducible results.

### 4.5 Disciplinary Variations and Contextual Misuse

However, the misuse of p-values is not limited to the biomedical field. Taroni et al. (2016) discuss the misuse of p-values in the field of forensic science, sometimes ignoring the limitations of p-values in certain contexts. Ferr (2025) shows the misuse of p-values in the field of psychological studies, as p-values are frequently misinterpreted in this field.

### 4.6 Evidence from Meta-Analyses and Systematic Reviews

Meta-analyses by Benjamin & Berger (2019), which show that a significant percentage of published results are false positives and that this is partly caused by incorrect use of p-values. In addition, it has been shown that a significant percentage of published results report a p-value less than 0.05 without sufficiently taking into account effect sizes and other evidence, thereby overemphasizing the importance of the results.

Summary: The evidence from various sources shows that there is a high probability that p-values are misused in practice. Common practices include p-hacking, selective reporting, and misinterpreting significance levels, which lead to a high number of false-positive results and a general decrease in scientific credibility.

## 5. RECOMMENDATIONS AND EMERGING ALTERNATIVES

Considering the misuse and concept flaws of p-values, there has been a growing concern in the scientific community to make reforms in statistical practices. This includes improving existing statistical methodologies as well as exploring new statistical methodologies that can offer more valuable results.

### 5.1 Improving Statistical Reporting and Education

The first step in this process is to improve the statistical literacy of the researchers. Lakens (2021) states that the correct usage of p-values is to present them in a transparent manner with the presentation of effect sizes and confidence intervals. In addition, researchers need to be educated on the correct interpretation of p-values. They need to understand that p-values actually indicate the degree of compliance of the data with the null hypothesis and not the probability of the null hypothesis being true.

Journals and funding bodies need to impose stricter reporting guidelines, and researchers need to be transparent about the results. This will include the presentation of effect sizes, confidence intervals, and the entire methodology used (Biau et al., 2010).

### 5.2 Moving Beyond the Binary Significance Paradigm

The American Statistical Association (2016) and Wasserstein et al. (2019) recommend that the overemphasis of the significance threshold of $p < 0.05$ should be removed. Instead, the p-values should be part of a broader context, including effect sizes and measures of uncertainty.

Benjamin & Berger (2019) recommend a reduction of the significance threshold to a $p < 0.005$ in order to minimize false positives. In addition, the significance of replication and robustness tests is highlighted. The culture of scientific research should be changed from a binary decision to a continuum.

### 5.3 Embracing Effect Sizes and Confidence Intervals

Researchers such as Andrade (2019) and Di Leo & Sardanelli (2020) suggest that it is more important to use effect size measures than to use p-value measures alone. This is so since effect size measures allow a researcher to understand the actual size of a given effect, which can then be interpreted as being practically important or unimportant.

### 5.4 Adoption of Bayesian Methods

However, Bayesian inference presents a promising alternative that measures evidence for hypotheses directly. Zhang (2022) and Quatto et al. (2020) have suggested that Bayesian methods, such as the use of Bayes factors, enable researchers to use prior knowledge and update their beliefs based on observed evidence.

### 5.5 Development of New Methodological Frameworks

New approaches are being developed as alternatives or supplements to p-values:

- S-values (Greenland, 2019) are an interpretation of p-values as measures of evidence, helping to resolve some interpretability problems.
- False Discovery Rate (FDR) adjustments are useful for controlling the proportion of false positives in multiple test problems (Hochberg & Benjamini, 1995).
- Pre-registering studies and sharing data help reduce data dredging, p-hacking, and selective reporting.

### 5.6 Cultural and Policy Changes

Another area of major need is the culture of research. The culture of research must be changed. Incentives must be provided in journals and institutions to encourage transparent reporting, replication of results, and the use of different inferential tools. Educational programs must be provided to understand the concepts of statistics beyond p-values. A culture of critical interpretation must be promoted (Kmetz, 2019).

## 6. DISCUSSION

This extensive body of literature, which is reviewed here, points to a critical issue that is prevalent in scientific research, which is related to the misuse of p-value. P-value is one of the most fundamental aspects of scientific research, yet it is often misused, misinterpreted, and even abused.

### 6.1 The Centrality and Flaws of P-Values in Scientific Practice

However, the practice of relying on p-values as the main indicator of significance has become deeply ingrained in the scientific culture, with the arbitrary threshold of $p < 0.05$ as the main indicator of significance. The dichotomous approach creates a false sense of security, oversimplifying the complexities of the data, while the significance of the effects, the size of the effects, and reproducibility are ignored.

Misconceptions, such as the probability that the null hypothesis is true, are clearly evident in the literature, which shows that these practices lead to unethical practices in research, including p-hacking, which increases the probability of Type I error.

### 6.2 Empirical Evidence Supporting Widespread Misuse

Research has shown that a vast majority of published research reports have a significant level of p values less than 0.05 without proper contextualization and transparency. The high incidence of p-hacking, data dredging, and publication bias leads to a high false positive rate, which in turn contributes to the reproducibility crisis.

The implications of this crisis are significant and include wasted resources and a loss of public trust in science.

However, understanding these facts and figures makes it imperative that a change in statistical practices is urgently called for.

### 6.3 Pathways Toward Reform and Better Practices

The literature suggests several avenues for improvement. For instance, statistical literacy and reporting practices should be enhanced. There are also calls for a change in the culture of the research community. For example, researchers are advised to move beyond the sole reliance on the p-value < 0.05. Instead, they are encouraged to use other measures of evidence such as effect size and confidence intervals. The use of Bayesian statistics and other measures such as the Bayes factor and s-value is also proposed as an alternative.

### 6.4 Challenges and Future Directions

However, there are challenges to implementing such reforms, including cultural factors, lack of statistical literacy, and institutional pressures to obtain statistically significant results. To achieve such reforms, there is a need to make concerted efforts as researchers, journals, funding agencies, and educators.

There is a need to conduct more research on how to develop effective statistical tools, how to promote statistical literacy, and how to create policies to encourage transparency. Emphasizing the significance of scientific context as opposed to statistical significance is essential to move credible science forward.

### 6.5 Limitations of the Current Literature

Even though this review represents a wide spectrum of views and perspectives, it must be noted that this review is limited by the differences in discipline-specific standards and methods, and the degree of statistical literacy in each researcher. More empirical studies must be conducted to assess the efficacy of proposed reforms and methods.

## 7. CONCLUSION

This systematic review emphasizes various issues related to the use and misuse of p-value, which is a fundamental aspect of statistical inference. Empirical evidence shows that the use of arbitrary significance levels leads to questionable research practices such as "p-hacking" or selective reporting. Conceptual issues show that there is a fundamental misunderstanding of what p-value really is, which further emphasizes the importance of improving statistical literacy.

With these issues in mind, the scientific community is slowly adopting reforms to overcome these issues, shifting focus from the binary "significant / nonsignificant" framework to more sophisticated methods such as using effect sizes, confidence intervals, or even Bayesian methods. These methods provide richer, more accurate, and more reliable results, creating a culture of transparency.

To overcome these issues, there is a need to address the problem of p-value misuse, which requires a collective effort from all aspects of the scientific community, including researchers, journals, policymakers, and educators. Improving statistical literacy, adopting transparent reporting, and using methodological innovations would greatly help advance credible, impactful, and influential science in the future.

## REFERENCES

[1] Wasserstein, R. L., Schirm, A. L. and Lazar, N. A. (2019) ‘Moving to a World Beyond “p < 0.05”’, The American Statistician, 73(sup1), pp. 1–19. doi: 10.1080/00031305.2019.1583913.

[2] Lakens, D., 2021. The practical alternative to the p value is the correctly used p value. Perspectives on psychological science, 16(3), pp.639-648.

[3] Zhang, W., 2022. p-value based statistical significance tests: Concepts, misuses, critiques, solutions and beyond. Computational Ecology and Software, 12(3), p.80.

[4] Gagnier, J.J. and Morgenstern, H., 2017. Misconceptions, misuses, and misinterpretations of p values and significance testing. JBJS, 99(18), pp.1598-1603.

[5] Lesaffre, E., 2008. Use and misuse of the p-value. Bulletin of the Hospital for Joint Diseases, 66(2), pp.146-149.

[6] Cohen, H.W., 2011. P values: use and misuse in medical literature. American journal of hypertension, 24(1), pp.18-23.

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

[8] Andrade, C., 2019. The P value and statistical significance: misunderstandings, explanations, challenges, and alternatives. Indian journal of psychological medicine, 41(3), pp.210-215.

[9] Taroni, F., Biedermann, A. and Bozza, S., 2016. Statistical hypothesis testing and common misinterpretations: Should we abandon p-value in forensic science applications?. Forensic science international, 259, pp.e32-e36.

[10] Vidgen, B. and Yasseri, T., 2016. P-values: misunderstood and misused. Frontiers in Physics, 4, p.6.

[11] Kim, J. and Bang, H., 2016. Three common misuses of P values. Dental hypotheses, 7(3), pp.73-80.

[12] Wang, M. and Long, Q., 2022. Addressing common misuses and pitfalls of P values in biomedical research. Cancer research, 82(15), pp.2674-2677.

[13] Kmetz, J.L., 2019. Correcting corrupt research: Recommendations for the profession to stop misuse of p-values. The American Statistician, 73(sup1), pp.36-45.

[14] Benjamin, D.J. and Berger, J.O., 2019. Three recommendations for improving the use of p-values. The American Statistician, 73(sup1), pp.186-191.

[15] Greenland, S., 2019. Valid p-values behave exactly as they should: Some misleading criticisms of p-values and their resolution with s-values. The American Statistician, 73(sup1), pp.106-114.

[16] Berselli, N., Filippini, T., Adani, G. and Vinceti, M., 2021. Dismissing the use of P-values and statistical significance testing in scientific research: new methodological perspectives in toxicology and risk assessment. In Toxicological risk assessment and multi-system health impacts from exposure (pp. 309-321). Academic Press.

[17] Chen, O.Y., Bodelet, J.S., Saraiva, R.G., Phan, H., Di, J., Nagels, G., Schwantje, T., Cao, H., Gou, J., Reinen, J.M. and Xiong, B., 2023. The roles, challenges, and merits of the p value. Patterns, 4(12).

[18] Di Leo, G. and Sardanelli, F., 2020. Statistical significance: p value, 0.05 threshold, and applications to radiomics—reasons for a conservative approach. European radiology experimental, 4(1), p.18.

[19] Concato, J. and Hartigan, J.A., 2016. P values: from suggestion to superstition. Journal of Investigative Medicine, 64(7), pp.1166-1171.

[20] Imbens, G.W., 2021. Statistical significance, p-values, and the reporting of uncertainty. Journal of Economic Perspectives, 35(3), pp.157-174.

[21] Motulsky, H.J., 2014. Common misconceptions about data analysis and statistics. The Journal of pharmacology and experimental therapeutics, 351(1), pp.200-205.

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

[23] Ferr, H., 2025. Misinterpretations of the p-value in psychological research: Implications for mental health and psychological science. PLOS Mental Health, 2(2), p.e0000242.

[24] Biau, D.J., Jolles, B.M. and Porcher, R., 2010. P value and the theory of hypothesis testing: an explanation for new researchers. Clinical Orthopaedics and Related Research®, 468(3), pp.885-892.

[25] Avdović, A. and Vidović, Z., 2025. Reevaluating p-value and its impact on conventional results display and interpretation. Filomat, 39(34), pp.12327-12343.

[26] Goodman, S., 2008, July. A dirty dozen: twelve p-value misconceptions. In Seminars in hematology (Vol. 45, No. 3, pp. 135-140). WB Saunders.

[27] Hofmann, M. and Meyer-Nieberg, S., 2018. Time to dispense with the p-value in OR? Rationale and implications of the statement of the American Statistical Association (ASA) on p-values. Central European Journal of Operations Research, 26(1), pp.193-214.

[28] Dahiru, T., 2008. P–value, a true test of statistical significance? A cautionary note. Annals of Ibadan postgraduate medicine, 6(1), p.21.

[29] Ioannidis, J.P., 2019. What have we (not) learnt from millions of scientific papers with P values?. The American Statistician, 73(sup1), pp.20-25.

[30] Lu, Y. and Belitskaya-Levy, I., 2015. The debate about p-values. Shanghai Archives of Psychiatry, 27(6), pp.381-385.

[31] Quatto, P., Ripamonti, E. and Marasini, D., 2020. Best uses of p-values and complementary measures in medical research: Recent developments in the frequentist and Bayesian frameworks. Journal of Biopharmaceutical Statistics, 30(1), pp.121-142.