International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Statistical Significance is Desirable but not Absolute: In Search of the True Study

Arnabjyoti De

MBBS, Mahatma Gandhi Medical College and Research Institute

Abstract

In the field of scientific research, the pursuit of robust and reliable results is paramount. Traditionally, the p-value has been the cornerstone of statistical analysis, used to determine the significance of study findings. However, there is a growing consensus that the calculation of effect size and statistical power analysis offer a more comprehensive and superior approach to data interpretation. While the p-value has its place in statistical analysis, the calculation of effect size and statistical power analysis provides a more nuanced and informative perspective on research findings. Embracing these methods will enhance the quality and impact of scientific research, leading to more robust and meaningful conclusions.

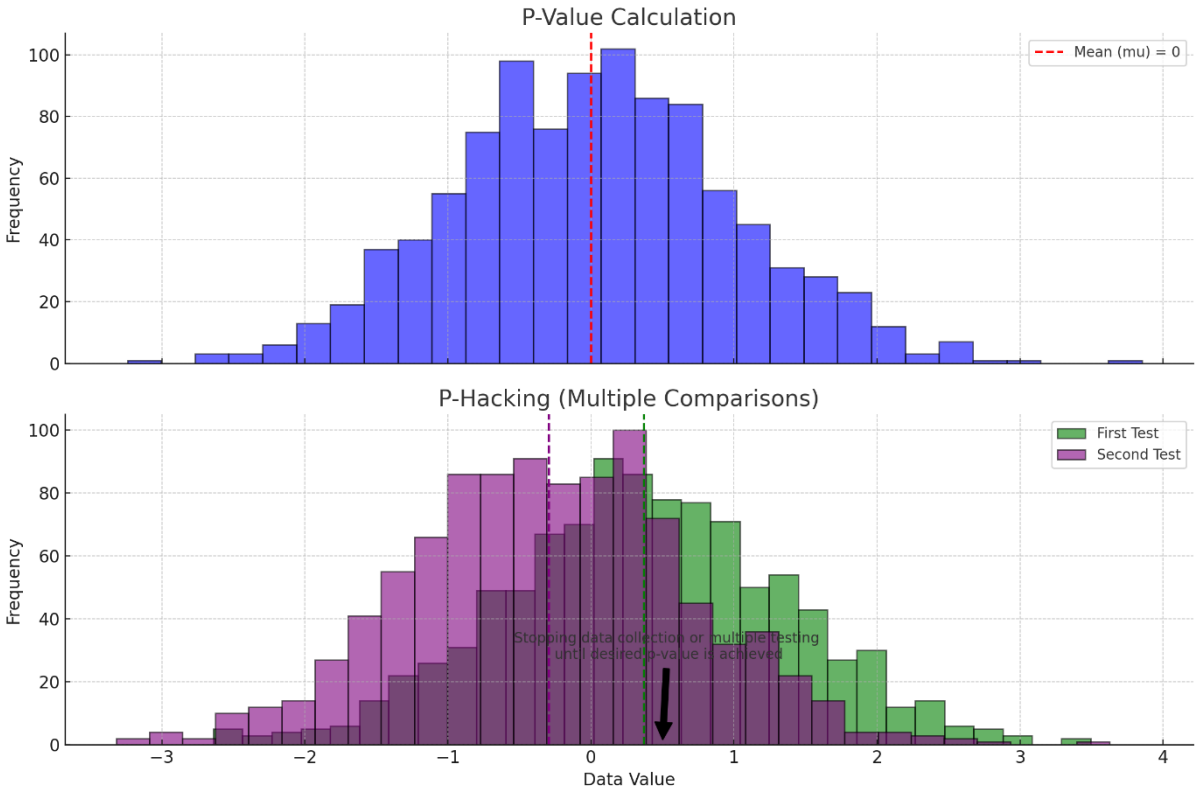
In the field of scientific research, the pursuit of robust and reliable results is paramount. Traditionally, the p-value has been the cornerstone of statistical analysis, used to determine the significance of study findings. However, there is a growing consensus that the calculation of effect size and statistical power analysis offer a more comprehensive and superior approach to data interpretation[1]

The p-value, while useful, has significant limitations. It merely indicates the probability that the observed data would occur by chance under the null hypothesis, without providing any information about the magnitude or practical significance of the effect. Consequently, a statistically significant p-value does not necessarily imply that the findings are meaningful or impactful in a real-world context. This has led to widespread misuse and misinterpretation of p-values, contributing to the reproducibility crisis in scientific research.

P-hacking, or data dredging, involves manipulating data analysis until nonsignificant results become statistically significant. This practice, although often subtle and unintentional, poses a significant threat to the credibility and reliability of scientific findings. The allure of statistical significance is powerful in the research community. The p-value, typically set at 0.05, serves as a gatekeeper for publication, funding, and academic recognition. Consequently, researchers may be tempted to engage in p-hacking by selectively reporting data, conducting multiple analyses without proper corrections, or even excluding outliers to achieve the coveted $p < 0.05$ threshold. This not only distorts the scientific record but also contributes to the reproducibility crisis, where many studies fail to replicate.

P-hacking undermines the fundamental principles of scientific inquiry, which are grounded in transparency, objectivity, and reproducibility. When researchers manipulate data to achieve statistical significance, they compromise the validity of their findings and mislead the scientific community and the public. This can have far-reaching consequences, particularly in fields such as medicine and public health, where policy and practice are influenced by research outcomes[2]

Here is a visual representation of the relationship between p-value calculation and p-hacking:



1. P-Value Calculation:

- The top histogram shows a normal distribution of data centred around the mean (mu = 0). The p-value is calculated based on the deviation from this mean.
- The red dashed line indicates the mean of the data.

2. P-Hacking:

- The bottom histogram shows two sets of data (green and purple) representing multiple comparisons or tests.
- The green and purple dashed lines indicate the means of these data sets.
- The annotation explains how p-hacking can involve stopping data collection or performing multiple tests until the desired p-value is achieved, leading to misleading results.

This visual highlights how p-hacking can artificially create significant p-values by exploiting statistical methods.

In contrast, effect size measures the strength of the relationship between variables, offering a clear indication of the practical significance of the results. Whether using Cohen's d, Pearson's r, or other metrics, effect size provides a quantifiable measure of how much of an effect is present, facilitating better comparison across studies and enhancing the interpretability of findings· By focusing on the magnitude of the effect, researchers can make more informed conclusions about the real-world implications of their work[3]

Moreover, statistical power analysis is crucial in study design and interpretation. Power analysis calculates the probability that a study will detect an effect of a given size, assuming that the effect exists. This helps researchers determine the necessary sample size to achieve reliable results, reducing the likelihood of false negatives. High statistical power ensures that studies are adequately equipped to detect true effects, thereby improving the validity and reliability of scientific findings[4]

Integrating effect size and power analysis into research practices addresses many shortcomings associated with p-value reliance. It promotes transparency, rigor, and reproducibility, fostering a deeper understanding of the data and its implications. Furthermore, this approach aligns with the increasing emphasis on practical significance and the reproducibility of research findings in the scientific community[5] While the p-value has its place in statistical analysis, the calculation of effect size and statistical power analysis provides a more nuanced and informative perspective on research findings. Embracing these methods will enhance the quality and impact of scientific research, leading to more robust and meaningful conclusions.

Conclusion

P-values seem like the accepted standard for publishing.The use of effect-sizes with corresponding confidence intervals is more useful and should be the goal in the first place. Furthermore, in most cases p-values do not answer any of the questions the authors asks. Yet, stating that p <.05 seems to be the holy threshold. The term “significant” a misleading term since it simply does not tell anything about the effect-size at all or how likely the author is wrong, which can be exactly similar as for smaller sample sizes. It seems that including more samples is a form of "p-hacking". Hence, marginal effect-sizes and "significant" p-values do not give any information.Critical looking at authors own results and the results of others should be considered as an epitome of idealistic scientific study so that author could understand the context of their own hypothesis and not just follow desirable conclusions to get a “significant” study

Reference

1. Haas, J. P. (2012). Sample size and power. *American journal of infection control*, *40*(8), 766-767.
2. Halpern, S. D., Karlawish, J. H., & Berlin, J. A. (2002). The continuing unethical conduct of underpowered clinical trials. *Jama*, *288*(3), 358-362.
3. Maier, M., & Lakens, D. (2022). Justify your alpha: A primer on two practical approaches. *Advances in Methods and Practices in Psychological Science*, *5*(2), 25152459221080396
4. Vandenbroucke, J. P., Elm, E. V., Altman, D. G., Gøtzsche, P. C., Mulrow, C. D., Pocock, S. J., ... & Strobe Initiative. (2007). Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): explanation and elaboration. *Annals of internal medicine*, *147*(8), W-163
5. Xu, Mengran & Wegener, Duane. (2023). Persuasive Benefits of Self-Generated Arguments: Moderation and Mechanism. Social Psychological and Personality Science. 15. 194855062211466. 10.1177/19485506221146612.