17 July 2025 ·

Qeios

Peer Review

Review of: "A Redemption Song for Statistical Significance"

Jonas Michel Wolf[1]

1. Hospital Moinhos de Vento, Porto Alegre, Brazil

This paper offers a timely and focused examination of statistical significance within the context of Ronald Fisher's null hypothesis model, positioning it as a crucial tool for discerning true effects from noise. In an era dominated by complex statistical methods and machine learning, the authors' decision to revisit fundamental concepts like Fisher's statistical significance is commendable and provides a valuable perspective for researchers at all levels of expertise. The clear objective—to demonstrate the utility of statistical significance in filtering out erroneous effect sizes—is well-articulated from the outset, setting a precise scope for the investigation.

A significant strength of this work lies in its methodological approach. By utilizing computer-simulated datasets with varying sample sizes, the authors create a controlled environment to rigorously test a true null hypothesis. This simulation-based strategy effectively isolates the impact of statistical significance, allowing for a direct assessment of its ability to identify situations where a true difference between population means is absent. The explicit mention of employing independent samples t-tests, a widely understood statistical procedure, further enhances the clarity and reproducibility of the study.

The choice of metrics for evaluating the results is also a strong point. The dual assessment using a 5% cut score for p-values for statistical significance and Cohen's "effect size index d" for substantive significance provides a comprehensive view. This combined approach acknowledges that statistical significance alone may not convey the practical importance of a finding, while simultaneously emphasizing its role in validating the existence of an effect. This nuanced perspective on significance is crucial for encouraging a more holistic interpretation of research outcomes.

However, the paper's primary focus on Fisher's statistical significance and the null hypothesis model, while a stated intention, also presents an opportunity for further discussion. While the authors explicitly state they do not attempt to unify statistical theories or claim superiority, a brief acknowledgment of the

doi.org/10.32388/2JF14U

historical disagreements and the motivations behind alternative frameworks (e.g., Neyman-Pearson's inclusion of alternative hypotheses, Bayesian approaches) could enrich the introductory context without detracting from the paper's core message.

One area for potential enhancement lies in elaborating on the "false effect sizes (effect size errors)" and how statistical significance actively filters them. While the conclusion states this outcome, a more detailed explanation or illustrative examples within the results or discussion section would provide a deeper understanding for readers. This could involve showing scenarios where a "substantively significant" effect size might arise purely by chance in the absence of statistical significance, thereby reinforcing the filtering mechanism.

Furthermore, while the paper highlights the historical evolution of statistical thought, from Fisher to Neyman-Pearson and Bayesians, and touches upon modern machine learning, the connection between these historical developments and the contemporary relevance of Fisher's specific contribution could be explored in greater depth. How does understanding this foundational concept help researchers navigate the complexities of modern, computer-intensive methods and massive datasets?

The paper's humanized tone, particularly in its introductory acknowledgment of "disagreement" in statistics, makes complex ideas more accessible. Maintaining this approach throughout the discussion, perhaps by relating the practical implications of misinterpreting effect sizes to real-world research scenarios, could further resonate with a broader audience of researchers and practitioners.

Considering the rapid advancements in statistical science, especially with machine learning algorithms, the paper's assertion that classical methodologies have been "surpassed" could be framed with greater nuance. While machine learning offers powerful tools, it often complements, rather than entirely replaces, classical statistical inference. A brief discussion on the symbiotic relationship between these paradigms would offer a more balanced view.

The paper’s strong conclusion—that statistical significance is a viable tool for filtering out false effect sizes—is well-supported by its simulated data. This finding reinforces the enduring value of foundational statistical principles, even in an increasingly data-driven research landscape. It serves as a vital reminder that while effect sizes indicate magnitude, statistical significance provides a necessary check on their reliability.

To further strengthen the paper, a dedicated section or subsection discussing the limitations of relying solely on p-values, even when used as a filtering tool, could be beneficial. While the paper uses Cohen's *d* for substantive significance, a brief acknowledgment of concerns like p-hacking or the arbitrary nature of alpha levels would demonstrate a comprehensive understanding of the ongoing debates in statistical inference.

In essence, this paper makes a valuable contribution by re-emphasizing the critical role of statistical significance in ensuring the robustness of research findings. Its clear methodology, well-defined objective, and accessible writing style lay a solid foundation. By considering opportunities to further contextualize its arguments within the broader statistical discourse and offer more explicit illustrations of its filtering mechanism, the paper can enhance its impact and provide even richer insights for the scientific community.

Declarations

Potential competing interests: No potential competing interests to declare.