

Evaluating the Use of Confidence Intervals in Quantifying Uncertainty and Improving Statistical Reporting Practices

Ava Rodriguez, Ava Smith, Charlotte Davis

1 Introduction

Statistical inference forms the backbone of scientific research across disciplines, yet the methods for quantifying and communicating uncertainty remain contentious and often misunderstood. Confidence intervals have been widely promoted as superior alternatives to point estimates and null hypothesis significance testing for conveying the precision of estimates and the uncertainty inherent in statistical inference. Despite decades of advocacy from statisticians and methodological experts, the implementation and interpretation of confidence intervals in practice continue to exhibit significant shortcomings. This research addresses the critical gap between statistical theory and practical application by systematically evaluating how confidence intervals are used, interpreted, and reported across scientific domains.

The fundamental concept of confidence intervals dates back to the work of Jerzy Neyman in the 1930s, who introduced them as a method for interval estimation that would contain the true parameter value with a specified long-run frequency. While mathematically elegant, the frequentist interpretation

of confidence intervals has proven challenging for applied researchers to grasp intuitively. The common misconception that a 95

Our investigation extends beyond documenting misinterpretations to propose constructive solutions for improving statistical reporting practices. We examine how confidence intervals can be more effectively integrated into scientific communication to enhance transparency, facilitate interpretation, and support evidence-based decision making. By developing a novel framework for evaluating confidence interval reporting quality and testing modified interval estimation procedures, this research contributes to the ongoing effort to strengthen statistical practice across scientific disciplines. The interdisciplinary nature of our approach allows for identification of both universal challenges and domain-specific considerations in uncertainty communication.

2 Methodology

This research employed a comprehensive multi-method approach to evaluate confidence interval usage and develop improved reporting frameworks. Our methodology integrated three complementary components: systematic literature analysis, experimental surveys with active researchers, and computational simulation studies. The systematic literature review examined 500 recently published articles from five distinct disciplines: psychology, medicine, ecology, economics, and engineering. We developed a novel Confidence Interval Reporting Index (CIRI) comprising 12 criteria assessing various aspects of confidence interval presentation and interpretation. These criteria included whether confidence intervals were reported for primary outcomes, whether they were explicitly interpreted in the text, whether precision was discussed in relation to sample size, and whether clinical or practical significance was considered alongside statistical significance.

The experimental component involved recruiting 300 active researchers across the same five disciplines, balanced for career stage and methodological expertise. Participants completed a survey assessing their understanding and interpretation of confidence intervals through multiple-choice questions, interpretation exercises, and practical scenarios. The survey included items testing basic conceptual understanding, such as the definition of confidence level, as well as more complex interpretive tasks requiring application to realistic research contexts. We employed randomized experimental conditions to test whether different presentation formats (e.g., numerical ranges, graphical displays, or narrative descriptions) influenced interpretation accuracy.

Our simulation studies investigated the performance of conventional and modified confidence interval procedures under various conditions. We compared standard Wald intervals with Bayesian-inspired approaches incorporating shrinkage estimators, which borrow information across related parameters to improve estimation precision. Simulation conditions varied sample size, effect magnitude, distributional characteristics, and dependency structures to assess robustness across realistic research scenarios. Performance metrics included empirical coverage probability, interval width, and estimation bias under both ideal and violation conditions.

3 Results

The systematic literature review revealed substantial variability in confidence interval reporting practices across disciplines. The overall Confidence Interval Reporting Index (CIRI) scores averaged 5.8 out of 12 possible points, indicating moderate reporting quality with considerable room for improvement. Medical research demonstrated the highest average CIRI score (7.2), while psychology journals showed the lowest (4.3). Notably, only 45

Experimental survey results demonstrated significant misunderstandings in confidence interval interpretation across researcher groups. Only 32

Simulation results indicated that conventional confidence interval procedures often exhibit suboptimal performance in small-sample scenarios, with empirical coverage probabilities frequently falling below the nominal level. The Bayesian-inspired shrinkage estimators demonstrated improved performance, particularly for small to moderate sample sizes, achieving coverage probabilities closer to the nominal level while maintaining reasonable interval widths. These modified procedures showed particular advantages in scenarios with multiple related parameters, where borrowing information across estimates enhanced precision without substantial bias introduction. The improvement was most pronounced for correlation parameters and variance components, which are commonly estimated with poor precision in conventional approaches.

4 Conclusion

This research provides comprehensive evidence regarding the current state of confidence interval usage and interpretation across scientific disciplines. Our findings indicate that despite statistical recommendations favoring interval estimation over dichotomous significance testing, confidence intervals remain poorly understood and inconsistently applied in practice. The disciplinary variations in reporting quality and interpretation accuracy suggest that field-specific norms and training approaches significantly influence statistical communication practices. The development of the Confidence Interval Reporting Index offers a practical tool for journals and institutions to assess and improve statistical reporting standards.

The experimental demonstration that presentation format influences interpretation accuracy has important implications for scientific communication. Re-

searchers, journal editors, and educators should consider how statistical results are visually and verbally presented to minimize misinterpretation. Our findings support the use of graphical displays alongside numerical intervals and recommend explicit narrative interpretations that accurately convey the frequentist meaning of confidence intervals. Additionally, the simulation results suggesting advantages of modified interval estimation procedures indicate potential directions for methodological development, particularly for research contexts with limited sample sizes or complex dependency structures.

This research contributes to improved statistical practice by identifying specific areas for intervention in researcher education, journal guidelines, and methodological development. Future work should explore targeted training interventions to address common misconceptions and develop discipline-specific guidelines for confidence interval reporting. The integration of Bayesian-inspired approaches with frequentist interval estimation represents a promising direction for enhancing the precision and interpretability of uncertainty quantification in scientific research.

References

- Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, 25(1), 7-29.
- 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 to misinterpretations. *European Journal of Epidemiology*, 31(4), 337-350.
- Hoekstra, R., Morey, R. D., Rouder, J. N., Wagenmakers, E. J. (2014). Robust misinterpretation of confidence intervals. *Psychonomic Bulletin Review*, 21(5), 1157-1164.

Morey, R. D., Hoekstra, R., Rouder, J. N., Lee, M. D., Wagenmakers, E. J. (2016). The fallacy of placing confidence in confidence intervals. *Psychonomic Bulletin Review*, 23(1), 103-123.

Neyman, J. (1937). Outline of a theory of statistical estimation based on the classical theory of probability. *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 236(767), 333-380.

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

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

McShane, B. B., Gal, D., Gelman, A., Robert, C., Tackett, J. L. (2019). Abandon statistical significance. *The American Statistician*, 73(sup1), 235-245.

Amrhein, V., Greenland, S., McShane, B. (2019). Scientists rise up against statistical significance. *Nature*, 567(7748), 305-307.

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E. J., Berk, R., ... Johnson, V. E. (2018). Redefine statistical significance. *Nature Human Behaviour*, 2(1), 6-10.