

Decision Rules in Frequentist and Bayesian Hypothesis Testing: P-Value and Bayes Factor

Nicolao Giovanni⁽¹⁾, Fordellone Mario⁽¹⁾, Schiattarella Paola⁽¹⁾, Smimmo Annafrancesca⁽¹⁾, Di Gennaro Piergiacomo⁽¹⁾, Speranza Teresa⁽¹⁾, Simeon Vittorio⁽¹⁾, Signoriello Simona⁽¹⁾, Chiodini Paolo⁽¹⁾

(1) Department of Mental, Physical Health and Preventive Medicine, Medical Statistics Unit, University of Campania "Luigi Vanvitelli"

CORRESPONDING AUTHOR: Nicolao Giovanni, giovanni.nicolao@unicampania.it

INTRODUCTION

The P-value is a widely used tool in inferential statistics and represents the probability of obtaining a value equal to or more extreme than the one observed, assuming that the null hypothesis (H_0) is true [1].

One of its main advantages is its intuitive interpretation: a smaller P-value indicates a lower compatibility of the observed results with the null hypothesis [2].

However, the P-value has important limitations that could lead to significant distortions in the interpretation of the results obtained [3].

The most important limitation is its sensitivity to sample size: as the sample size increases, the power of the test also increases. Consequently, even minor and perhaps clinically irrelevant effects can produce statistically significant P-values, while important effects might not be detected in smaller samples [1].

The use of a fixed significance threshold (typically 0.05) can promote a binary interpretation of the results (significant vs. non-significant), oversimplifying the researcher's decision-making process. This approach risks not fully capturing the degree of statistical evidence, thereby increasing the likelihood of assessment errors [4].

Another limitation is that the P-value does not provide information about the evidence in favor of an alternative hypothesis (H_1): a small P-value may suggest that the data do not support the null hypothesis (H_0), but it does not quantify, through a comparative approach, how much more likely the data are under the alternative hypothesis [5].

The excessive use of the P-value encourages researchers to explore alternative approaches, such as the Bayes Factor (BF) [6].

The BF is a Bayesian tool used to compare the evidence in favor of two hypotheses by comparing the likelihood of the data under the null hypothesis with the likelihood of the data under the alternative hypothesis. Therefore, unlike the p-value, the BF directly measures the probability of the data under each hypothesis, providing a quantitative comparison between H_0 and H_1 [7].

Among the advantages of the BF is its ability to provide a continuous measure of evidence, comparing the alternative hypothesis with the null hypothesis while also allowing the incorporation of prior information into the analyses. Its value can be interpreted using specific scales [8].

OBJECTIVES

The objective of this work is to compare the P-value and the BF as statistical tools for hypothesis testing, in order to highlight their behaviors in different scenarios involving (i) sample size and (ii) effect size.

METHODS

A simulation study was conducted with various scenarios constructed by combining sample size and effect size. The proposed simulation uses a t-test on the difference between the means of two independent groups as the endpoint. Nine distinct scenarios were generated, which include: (i) three levels of effect size, defined as the standardized difference between the means of the two groups, equal to 0.1, 0.2, and 0.5; and (ii) three different sample sizes, equal to 50, 100, and 150. A total of 5000 replications were performed, and the results are expressed in terms of medians of the p-value and BF [9].

The Bayesian results were obtained using the R package "Bayes Factor." The default prior was applied, which is identified as a Cauchy distribution centered on 0 and is moderately informative. In the simulation, the default prior of the package was chosen for illustrative purposes, but the process of selecting a prior is not trivial and requires specific considerations related to the research context.

RESULTS

The results of the study show that the Bayes Factor (BF) is less sensitive to sample size compared to the P-value when effect sizes are small (0.1 and 0.2). It can also be observed that the P-value becomes statistically significant for sample sizes of 100 and 150 units with an effect size of 0.5, and its significance increases at a very high rate, compared to the BF where the evidence in favor of H_1 remains moderate. In other words, the P-value becomes extremely low in the presence of an effect size of 0.5 for a sample size of 150 units, whereas the BF remains more cautious, indicating only moderate evidence in favor of the alternative hypothesis.

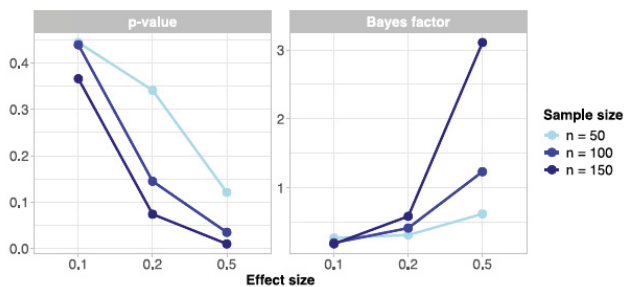


Figure 1. Comparison of Results Between the P-value and the Bayes Factor in the Simulation Study

CONCLUSIONS

The results reveal that the P-value is more sensitive to changes in sample size and effect size compared to the BF. Additionally, the BF provides a more nuanced approach to decision-making, addressing the binary nature of the P-value in rejecting the null hypothesis. The Bayesian alternative can be advantageous for researchers in the healthcare context, as it allows for the incorporation of informative priors that could enhance analysis results and reduce the likelihood of assessment errors. However, a significant challenge of using the BF lies in the choice of the prior distribution, which can significantly impact the final results of the analyses.

REFERENCES

1. Lehmann, E.L. (1993). The Fisher, Neyman–Pearson theories of testing hypotheses: One theory or two? *J. Am. Stat. Assoc.* 88, 1242–1249.
2. Casella, G., Berger, G.L. (2001). *Statistical Inference*, 2nd Ed. (Brooks/Cole, Pacific Grove).
3. Goodman, S. N. (1999). Toward evidence-based medical statistics. 1: The P value fallacy. *Annals of internal medicine*, 130(12), 995-1004.
4. Gardner, M.J., Altman, D.G. (1986). Confidence intervals rather than p values: Estimation rather than hypothesis testing. *BMJ* 292, 746–750.

5. Greenland, S., Senn, S. J., Rothman, J., et al. (2016). Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. *European journal of epidemiology*, 31(4), 337-350.
6. Goodman, S. N. (1999). Toward evidence-based medical statistics. 2: The Bayes factor. *Annals of internal medicine*, 130(12), 1005-1013.
7. Wasserstein, R.L., Lazar, N.A. (2016). The ASA statement on p-values: Context, process, and purpose. *Am. Statistician* 70, 129–133.
8. Held, L., Ott, M. (2018). On p-values and Bayes factors. *Annual Review of Statistics and Its Application*, 5(1), 393-419.
9. Fordellone M, Schiattarella P, Nicolao G, et al. Decision Rules in Frequentist and Bayesian Hypothesis Testing: P-Value and Bayes Factor. *Int J Public Health*. 2025 May 14;70:1608258.