The Pitfalls of the P-Value and Confidence Interval - Why Freqentist Inference Should Remain General And How Bayseian Inference Could Help

Ryan Bergner

The Pitfalls of the P-Value and Confidence Interval - Why Freqentist Inference Should Remain General And How Bayseian Inference Could Help - December 10, 2022

1. Introduction

I argue against the prevalence and overuse of both p-values and traditional frequentist confidence intervals, although confidence intervals are a better method for estimating a parameter. I argue that p-values have several flaws that are not only inherent based on their definition but flawed in the subjectivity and bias it allows from the standpoint of the researcher when conducting a particular experiment. Then, I argue confidence intervals are better. Still, by no means an all-encompassing nor sufficient substitute for providing an accurate definition of the significance of a statistical test *should* adhere to. I concur with the authors of "The Fallacy of Placing Confidence in Confidence Intervals" in the fact they implicitly (and explicitly [Morey et al., 6]) make the argument for Bayesian Inference throughout their critique of frequentist inference by highlighting the pitfalls of confidence intervals and p-values simultaneously using primarily problems arising when an interval measurement lacks a density aspect, as well as the reference class problem - as discussed in the conclusion of the article mentioned above. (7)

2. Problems Regarding P-Values and their Misuse

> *"Probability values (p-values) seem to be the solid foundation on which scientific progress relies."* - ***Osteoarthritis and Cartilage, 20***

The probability value (or p-value) has acted as the cornerstone of scientific papers, clinical trials, and other various types of societally important studies that have depended on for the better of the twentieth century. I claim the frequent use of the p-value has created an over-reliance on the metric and has opened the door for researchers, journalists, and anyone else who benefits from certain, specific statistical claims: an avenue to overconfident and fairly adjustable conclusions, leading to non-realistic understanding of the way the world works at best, and rampant scientific deception and political manipulation worst.

To put it bluntly, the p-value suffers from several issues, mainly stemming from the fact that, when considering its definition contrasted with the standard tests used by the vast majority of practitioners of statistics to make decisions, both serious and trivial impacting all fields of study ranging from business decisions to scientific proofs, are almost overwhelmingly dominated by the p-value. This is not because of its verified robustness under a wide array of inputs — but its simple, almost arbitrary interpretation allowing the ability for practitioners to apply several frequentist methods to any sample of data to obtain the minimum valid qualifying null-hypothesis rejection of below .05, provided a handful of weak conditions compared to the methods derived for other mathematical tools.

For example, consider any valid scientific experiment - all it takes is applying linear regression to a relatively small, fixed number of occurrences of the X and Y variables that the researcher decided to look at the moment he chose to record the data. In contrast, take, for example, yearly fatal airline accidents per year as a proportion of total flights. In the U.S. Aircraft Carrier Dataset, it can be seen that a much lower significance is required, or at least expected, for an airline provider in the United States. All fatal crashes typically happen on an extremely small scale, on average three decimal places smaller than required for traditional statistical significance. The reason this is the case is that people are scared to death about getting into a plane crash, and most have an expectation they are not going to be injured or die on a routine flight. The only reason this is the case is that most people have observed the aftermath of past plane crashes, usually having an extremely low survival rate compared to almost all other vehicular accidents. This is the only information that informs us as members of a society that we don’t tolerate a high number of plane crashes and would boycott a particular airline company failing to stay below the threshold of what most define as “a safe level” even though we may expect to see a plane crash once in awhile.

So, it doesn’t seem obvious that certain unknown processes only require thirty or more observations to assume normality. It may be the case that the level of confidence in an experiment shouldn’t be determined with a non-realistic predefined constant but with a dynamic and flexibly defined distribution that incorporates sampling error, an ever-increasing number of observations that converge to what we can know to be the closest we can get to “the truth.”

3. Why The Confidence Interval Is An Insufficient Substitute

While confidence intervals ameliorate a handful of the pitfalls associated with p-values, considering them to be a solution misses the point in terms of finding a solution to the main problem with the ideological building blocks that support the credibility of frequentist tools. At the end of the day, problems will always arise when trying to make claims about the variance of a general process based on the variability of a sample, as articulated by Jonas Ranstam in his critique of probability values. (Osteoarthritis and Cartilage 20, 807)

The main problem with confidence intervals (or CIs) is they just expand the bounds of possible values for a theoretical distribution of data already based on a predefined, arbitrary p-value which would almost certainly not be uniform or gaussian if the practitioner of a theoretical experiment had been able to observe a sufficient amount of observations that, ultimately cannot allow a said practitioner to even say anything definite regarding an unknown, unrealized, or unseen characteristic of a process. Yet, the confidence interval assumes all values within it to be equally likely, and its size is inversely related to its precision. Therefore, theoretically, the confidence interval doesn't converge due to an increase in confidence and instead widens. But what is the confidence of a particular process based on? Some seem obvious and need a small about of exterior information to be considered valid, while others are vastly complicated random, and hard to observe. So, for a particular process, we should theoretically have a different required number of observations for different processes. Still, popular frequentist statistics claim thirty is a satisfactory value on no basis.

In reality, the only manual thing that a human could do to become more and more confident of an event would be to view more data — to make up for a lack of information gained from the statistically significant frequentist results; the confidence interval tries to approximate what the distribution would be given an imaginary expansion of new data based on few already observed data to give the bounds of and inherent theoretical property. This conclusion, while useful in some cases, can not be used to "prove" or determine an inherent property of an unknown process because a statistician simply shortcuts the required sampling of data to obtain the desired result through a theoretical "approximation" of what the observed data would be. And, while more difficult - the confidence interval can be manipulated to show statistical significance when existence cannot be determined due to problems with the p-value and the reference class problem.

4. Conclusion

In conclusion, both frequentist methods of the p-value and the confidence interval are insufficient for determining an inherent property of an unknown process under any condition. Researchers and all statistical practitioners should be open to using bayesian methods in some cases in which data can be constantly collected and updated to output a distribution of the continuously-improving posterior probability. While they don't inherently determine a point estimate, the bayesian output provides both an expected value and what could be thought of as an asymptotic variance of a theoretical distribution as the number of observations increases to infinity. Scientists and other conductors of statistical experiments should become comfortable in explaining the ambiguity of distribution. If a certain distribution failed to converge quickly for a certain amount of data, it would motivate the researcher to acquire more observations to eventually converge to a near-point estimate with minimal variance. This would be a better method for widespread statistics for modern-day problems: data is widely available, and statisticians no longer have the excuse to rely on greater than or equal to thirty observations to pretend someone has proved something. In reality, the frequentist experiment is insufficient as it is far too dependent on a small, insufficient set of data that can't possibly be the basis of any inherent proof. To allow a statistician to say they can ameliorate the problem with a confidence interval simply allows said statistical researcher to tell the small minority of the non-statistically literate population who have the wherewithal to question the arbitrary, inaccurate nature of the mechanisms of the studies impacting their lives. In no way does the confidence interval ameliorate the issues — ease of use does not imply accuracy and simply leaves room for less than required accountability in statistical conclusions.

Sources

4.1. Texts

- Morey, R.D., Hoekstra, R., Rouder, J.N. et al. The fallacy of placing confidence in confidence intervals. Psychon Bull Rev 23, 103–123 (2016). https://doi.org/10.3758/s13423-015-0947-8
-
- Ranstam, J. "Why the P-Value Culture Is Bad and Confidence Intervals a Better Alternative." Osteoarthritis and Cartilage, vol. 20, no. 8, 2012, pp. 805–808. https://doi.org/10.1016/j.joca.2012.04.001.

4.2. Airline Dataset

- 1960: National Transportation Safety Board, Annual Review of Aircraft Accident Data: U.S. Air Carrier Operations, Calendar Year 1967 (Washington, DC: December 1968).