# Problematic probabilities: Reassessing the p-value in public health research

Gilbert Jr De Los Santos Bernardino[1]

AFFILIATION
1 College of Nursing, University of the Cordilleras, Baguio City, Philippines

CORRESPONDENCE TO
Gilbert Jr De Los Santos Bernardino. College of Nursing, University of the Cordilleras, Unit 3 Brentwood Breeze Homes, Manuel Roxas Street, Baguio City, 2600 Philippines.
E-mail: gdbernardino@uc-bcf.edu.ph

ORCID iD: https://orcid.org/0000-0002-8272-8518

KEYWORDS
confidence intervals, p-values, public health research, Bayes' theorem

Received: 26 August 2023, Revised: 13 February 2024
Accepted: 22 February 2024

Popul. Med. 2024;6(April):11

https://doi.org/10.18332/popmed/185251

**Dear Editor,**

P-values are a product of hypothesis testing, indicating the 'probability that no effect of an intervention (null hypothesis) has occurred within a population'[1]. A high p-value implies a 'higher probability for no effect', while a low p-value suggests a potential intervention effect[1]. The smaller the p-value in a particular research, the greater the statistical significance of its assertions. Moreover, the p-value signifies the likelihood, within a given statistical model, that the statistical summary would be equal to or more extreme than the observed results if the null hypothesis is true[2]. According to Gibson[3], the presence of a significantly noteworthy effect, unlikely under the assumption of no actual effect, invalidates the null hypothesis. A low p-value indicates the observation of something highly uncommon, casting doubt on the accuracy of the null hypothesis.

As public health practitioners, there exists an expectation that quality care must be evidence-based, i.e. supported by p-values in empirical studies. According to Cohen[4], there have been instances where p-values in medical literature have justified the million-dollar development of various pharmacological drugs and secured the tenure of some academic researchers who have made statistically significant claims in various fields. In public health research, the balance between avoiding false positives and false negatives depends on the consequences of errors and the research stage. For example, during the exploratory phase, an excess of false positives may waste resources, while an abundance of false negatives can lead to overlooking meaningful discoveries. However, it is crucial to adjust the p-value for both multiplicity and selection bias to ensure accurate results[3].

In recent times, the stellar position of p-values has come to the critical scrutiny of scholars. Statistical significance may not always imply scientific, human, or economic importance. Lower p-values do not always mean more substantial effects and higher p-values do not necessarily indicate a lack of importance[2]. A minute effect can generate a small p-value if the sample size or measurement precision is sufficiently high, and conversely, substantial effects may yield unremarkable p-values with a small sample size or imprecise measurements[2]. According to Altman and Bland[5], in randomized controlled trials, a p-value exceeding 5% is conventionally considered non-significant, but labeling a study as non-significant does not necessarily mean there are no clinically relevant findings. Controlled trials with small sample sizes often lack the statistical power needed to identify significant differences in treatment outcomes[5].

In this article, existing limitations of the p-value will be reviewed. In addressing such limitations, possible solutions, such as considering confidence intervals and Bayesian reasoning, will be explored.

## Limitations of the p-value

One reason why p-values must not be regarded as sole criteria lies in the fact that chance errors can happen at any point in empirical studies[4]. Benjamin et al.[6] highlight that, despite studies reporting statistically significant results with $p<0.005$, the replicability of such studies remains notably low. This underscores that the strength of evidence conveyed by a p-value is contingent on the nature of the research and its position within the research continuum. Differences in the level of study participants may also be difficult to control owing to the multitude of characteristics such as genetic composition and variations in environmental exposures or experiences of study participants, particularly in studies with large sample sizes[4]. Even if the possibility of confounding variables may be addressed by statistical models, it is essential to note that statistical models and various study designs may also have limitations[7]. Cross-sectional studies, for example, may only be able to provide a snapshot of a particular event or phenomenon without necessarily being true in the future[8]. The voting preferences of a group of individuals at one point may not be the same after a few months or after several years. Similarly, the findings of a non-experimental study may not be as rigorous as those of a randomized controlled trial.

Ioannidis[9] also highlighted that as more scientific teams and conflicts of interests and prejudices gravitate towards a scientific field, it can be less likely that its findings will be true, thus rendering the reliance on p-values alone as dubious. Conflicts of interest pertain to circumstances that may cloud the judgment of a researcher while conducting a study[10]. One example of conflict of interest can come in the form of accepting favors or financial remuneration from companies or organizations that are engaged in business[11]. Another example can be evident in the act of coming up with a manuscript that may render strong socio-political support to an individual, organization, or ideology[10]. Furthermore, several other significant factors can impact the p-value in research. Multiplicity, which involves making multiple comparisons, can inflate the likelihood of obtaining false-positive results[12,13]. Selection bias, where certain groups are disproportionately included or excluded from the study, can also influence the p-value[14]. Additionally, small and noisy studies with low statistical power may produce unreliable or inconclusive p-values. In public health research, for example, a study investigating the effectiveness of multiple interventions simultaneously may face multiplicity issues, and selection bias may arise if certain demographic groups are overrepresented or excluded. Similarly, a small-scale study on a rare disease may encounter challenges related to both small sample size and noise, affecting the reliability of its p-values.

## Potential strategies

Instead of relying only on p-values, Cohen[4] suggests that researchers may consider the importance of confidence intervals because of their capacity in 'estimating the range within which the true value can be expected'. Large studies in individual studies can have narrower confidence intervals, thereby indicating a more precise estimate[15]. In meta-analysis studies, the precision of estimates decreases as the heterogeneity of the studies included increases[15].

Taking into account prior pieces of evidence drawn from Bayes' theorem may also prove to be helpful because, in a way, other equally credible sources of information are taken into account[4], and the p-value is given more context[7]. Bayesianism, by virtue of its being inductive, considers data to be partial. An example can be depicted in the study of Bland and Altman[16,17] that aimed to determine the prevalence of diabetes in a region in the UK. In the original inquiry, the researchers relied on available survey results from the UK as a whole and from nearby areas. Given the information that diabetes in the UK has a prevalence of 2% and that some areas have a prevalence between 1 and 3%, it can be questionable to come up with a claim that the prevalence of diabetes in the locale of the study is 0 or exceeding 10%[17]. Bayesianism can be a good alternative because while it draws from multiple sources of data, it is open to the idea of changing a belief in the light of new evidence[17]. One specific example of Bayesianism presented in the above example pertains to disease prevalence estimation. Bayesian methods allow researchers to incorporate prior knowledge or beliefs about the prevalence of disease into their analysis, updating these beliefs based on observed data. This is particularly valuable when dealing with limited or imperfect data, as Bayesian approaches provide a flexible framework to combine prior information with current evidence, resulting in more robust estimates of disease prevalence.

## Future challenges

P-values can provide an indispensable tool in making sense of various statistical tests. However, caution must be observed in integrating these statistical findings in actual settings. Even if the p-value has reached an elevated status in scientific scholarship, Ou[18] warns that we must examine our tendency to have 'uncritical adherence' to its real-world implications. As such, we must be receptive to new forms of knowledge that call for re-examining existing frameworks[18]. Public health practitioners, in this regard, must assert their capacity to be actively involved in producing empirical evidence while being critical of the tools and methods that aim to provide generalizations in the name of empirical knowing. Policymakers in the realm of public health should be cautious of making decisions guided by p-values alone. Instead, components of empirical reports should be considered, such as confidence intervals and the possibility of selection bias in some research. Given the widespread misuse and misunderstanding of p-values, some statisticians advocate for alternative methods in public health research. These approaches prioritize estimation over testing and involve tools like confidence intervals, credibility intervals, or prediction intervals[2]. In public health situations, researchers may opt for Bayesian methods, likelihood ratios, or Bayes Factors to evaluate evidence more effectively. Decision-theoretic modeling and false discovery rates are additional strategies that can directly assess the magnitude of an effect and its uncertainty or scrutinize the correctness of a hypothesis in public health studies[2].

Benjamin et al.[6] also advise that instead of focusing merely on statistically significant results, readers also need to pay attention to how transparency is reflected so that other scholars may be adequately informed. In some instances, publication bias becomes apparent when only studies with statistically significant positive findings are published, while those with statistically insignificant or negative results are not[19]. In a similar vein, this may run parallel to the notion of publishing research findings that may not necessarily contain the desired p-values reflective of statistical significance, as in the case of the Journal of Negative Results[20] that gives a platform for empirical studies in the fields of evolutionary biology and ecology without giving sole reliance on significance thresholds. Moreover, registering research protocols will increase trust in the reported findings by ensuring that the methodology was established beforehand and verifying that it was not manipulated to align with the authors' preferences[21].

## REFERENCES

1. Stratton SJ. Significance: statistical or clinical?. Prehosp Disaster Med. 2018;33(4):347-348. doi:10.1017/S1049023X18000663
2. Wasserstein RL, Lazar NA. The ASA's statement on p-values: context, process, and purpose. Am Stat. 2016;70(2):129-133. doi:10.1080/00031305.2016.1154108
3. Gibson EW. The role of p-values in judging the strength of evidence and realistic replication expectations. Statistics in Biopharmaceutical Research. 2021;13(1):6-18. doi:10.1080/19466315.2020.1724560
4. Cohen HW. P values: use and misuse in medical literature. Am J Hypertens. 2011;24(1):18-23. doi:10.1038/ajh.2010.205
5. Altman DG, Bland JM. Absence of evidence is not evidence of absence. BMJ. 1995;311(7003):485. doi:10.1136/bmj.311.7003.485
6. Benjamin DJ, Berger JO, Johannesson M, et al. Redefine statistical significance. Nat Hum Behav. 2018;2(1):6-10. doi:10.1038/s41562-017-0189-z
7. Gao J. P-values - a chronic conundrum. BMC Med Res Methodol. 2020;20(1):167. doi:10.1186/s12874-020-01051-6
8. Dawson B, Trapp RG, Dawson-Saunders B. Basic & clinical biostatistics. 2nd ed. Appleton & Lange; 1994. Accessed February 16, 2024. https://accessmedicine.mhmedical.com/content.aspx?bookid=2724&sectionid=226990388
9. Ioannidis JPA. Why most published research findings are false. PLOS Medicine. 2005;2(8):e1004085. doi:10.1371/journal.pmed.0020124
10. Romain PL. Conflicts of interest in research: looking out for number one means keeping the primary interest front and center. Current Reviews in Musculoskeletal Medicine. 2015;8(2):122-127. doi:10.1007/s12178-015-9270-2
11. University of Pittsburgh. Examples of COIs. Accessed February 16, 2024. https://www.coi.pitt.edu/about-coi/examples-cois
12. Powell K, Prasad V. Multiplicity: when many analytic plans are applied or many redundant studies are run, false-positive results are ensured. Eur J Clin Invest. 2022;52(8):e13802. doi:10.1111/eci.13802
13. Benjamini Y. It's not the p-values' fault. The American Statistician. Accessed February 16, 2024. https://www.tandfonline.com/action/downloadSupplement?doi=10.1080%2F00031305.2016.1154108&file=utas_a_1154108_sm5354.pdf
14. Institute for Work & Health. Selection bias. IWH; 2014. Accessed February 16, 2024. https://www.iwh.on.ca/what-researchers-mean-by/selection-bias#:~:text=Selection%20bias%20is%20a%20kind,and%20cross%2Dsectional%20studies
15. The Cochrane Collaboration. Confidence intervals. Cochrane; 2011. Accessed February 16, 2024. https://handbook-5-1.cochrane.org/chapter_12/12_4_1_confidence_intervals.htm
16. Bland JM, Altman DG. Statistics notes. The odds ratio. BMJ. 2000;320(7247):1468. doi:10.1136/bmj.320.7247.1468
17. Allmark P. Bayes and health care research. Med Health Care Philos. 2004;7(3):321-332. doi:10.1007/s11019-004-0804-4
18. Ou CHK, Hall WA, Thorne SE. Can nursing epistemology embrace p-values?. Nurs Philos. 2017;18(4):10.1111/nup.12173. doi:10.1111/nup.12173
19. Nair AS. Publication bias - Importance of studies with negative results! Indian Journal of Anaesthesia. 2019;63(6):505-507. doi:10.4103/ija.IJA_142_19
20. Journal of Negative Results. About the journal. JNS. Accessed February 16, 2024. http://www.jnr-eeb.org/index.php/jnr.
21. Tawfik GM, Giang HTN, Ghozy S, et al. Protocol registration issues of systematic review and meta-analysis studies: a survey of global researchers. BMC Med Res Methodol. 2020;20(1):213. doi:10.1186/s12874-020-01094-9

### CONFLICTS OF INTEREST

The author has completed and submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest and none was reported.

### FUNDING

There was no source of funding for this research.

### ETHICAL APPROVAL AND INFORMED CONSENT

Ethical approval and informed consent were not required for this study.

### DATA AVAILABILITY

Data sharing is not applicable to this article as no new data were created.

### PROVENANCE AND PEER REVIEW

Not commissioned; externally peer reviewed.

### DISCLAIMER

The views and opinions expressed in this article are those of the author.