EDITORIAL

Editorials represent the opinions of the authors and *JAMA* and not those of the American Medical Association

The Enduring Evolution of the *P* Value

Demetrios N. Kyriacou, MD, PhD

Related article page 1141

Mathematics and statistical analyses contribute to the language of science and to every scientific discipline. Clinical trials and epidemiologic studies published in biomedical journals are essentially exercises in mathematical measurement.[1] With the extensive contribution of statisticians to the methodological development of clinical trials and epidemiologic theory, it is not surprising that many statistical concepts have dominated scientific inferential processes, especially in research investigating biomedical cause-and-effect relations.[1-6] For example, the comparative point estimate of a risk factor (eg, a risk ratio) is used to mathematically express the strength of the association between the presumed exposure and the outcome of interest.[7-9] Mathematics is also used to express random variation inherent around the point estimate as a range that is termed a *confidence interval*.[1] However, despite the greater degree of information provided by point estimates and confidence intervals, the statistic most frequently used in biomedical research for conveying association is the *P* value.[10-14]

In this issue of *JAMA*, Chavalarias et al[15] describe the evolution of *P* values reported in biomedical literature over the last 25 years. Based on automated text mining of more than 12 million MEDLINE abstracts and more than 800 000 abstracts and full-text articles in PubMed Central, the authors found that a greater percentage of scientific articles reported *P* values in the presentation of study findings over time, with the prevalence of *P* values in abstracts increasing from 7.3% in 1990 to 15.6% in 2014. Among the abstracts and full-text articles with *P* values, 96% reported at least 1 “statistically significant” result, with strong clustering of reported *P* values around .05 and .001. In addition, in an in-depth manual review of 796 abstracts and 99 full-text articles from articles reporting empirical data, the authors found that *P* values were reported in 15.7% and 55%, respectively, whereas confidence intervals were reported in only 2.3% of abstracts and were included for all reported effect sizes in only 4% of the full-text articles. The authors suggested that rather than reporting isolated *P* values, research articles should focus more on reporting effect sizes (eg, absolute and relative risks) and uncertainty metrics (eg, confidence intervals for the effect estimates).

To provide context for the increasing reporting of *P* values in the biomedical literature over the past 25 years, it is important to consider what a *P* value really is, some examples of its frequent misconceptions and inappropriate use, and the evidentiary application of *P* values based on the 3 main schools of statistical inference (ie, Fisherian, Neyman-Pearsonian, and Bayesian philosophies).[10,11]

The prominence of the *P* value in the scientific literature is attributed to Fisher, who did not invent this probability measure but did popularize its extensive use for all forms of statistical research methods starting with his seminal 1925 book, *Statistical Methods for Research Workers*.[16] According to Fisher, the correct definition of the *P* value is “the probability of the observed result, plus more extreme results, if the null hypothesis were true.”[13,14] Fisher’s purpose was not to use the *P* value as a decision-making instrument but to provide researchers with a flexible measure of statistical inference within the complex process of scientific inference. In addition, there are important assumptions associated with proper use of the *P* value.[10,11,13,14] First, there is no relation between the causal factor being investigated and the outcome of interest (ie, the null hypothesis is true). Second, the study design and analyses providing the effect estimate, confidence intervals, and *P* value for the specific study project are completely free of systemic error (ie, there are no misclassification, selection, or confounding biases). Third, the appropriate statistical test is selected for the analysis (eg, the χ^2 test for a comparison of proportions).

Given these assumptions, it is not difficult to see how the concept of the *P* value became so frequently misunderstood and misused.[10,13,14,16,17] Goodman has provided a list of 12 misconceptions of the *P* value.[14] The most common and egregious of these misconceptions, for example, is that the *P* value is the probability of the null hypothesis being true. Another prevalent misconception is that if the *P* value is greater than .05, then the null hypothesis is true and there is no association between the exposure or treatment and outcome of interest.

Within the different philosophies of statistical inference, both the Fisherian and the Neyman-Pearsonian approaches are based on the “frequentist” interpretation of probability, which specifies that an experiment is theoretically considered one of an infinite number of exactly repeated experiments that yield statistically independent results.[10-15,18] Frequentist methods are the basis of almost all biomedical statistical methods taught for clinical trials and epidemiologic studies. Although both the Fisherian and Neyman-Pearsonian approaches have many similarities, they have important philosophical and practical differences.

Fisher’s approach uses a calculated *P* value that is interpreted as evidence against the null hypothesis of a particular research finding.[14] The smaller the *P* value, the stronger the evidence against the null hypothesis. There is no need for a predetermined level of statistical significance for the calculated *P* value. A null hypothesis can be rejected, but this is not necessarily based on a preset level of significance or probability of committing an error in the hypothesis test (eg, $\alpha < .05$). In addition, there is no alternative hypothesis. Inference regarding the hypothesis is preferred over a mechanical decision to accept or reject a hypothesis based on a derived probability.

In contrast to Fisher, Neyman and Pearson in the 1930s formalized the hypothesis testing process with a priori assertions and declarations. For example, they added the concept of a formal alternative hypothesis that is mutually exclusive of the null hypothesis.[10,11] In addition, a value is preselected to merit the rejection of the null hypothesis known as the significance level.[13,14] The goal of the statistical calculations in the Neyman-Pearsonian approach is decision and not inference. By convention, the cutoff for determining statistical significance usually was selected to be a *P* value below .05. A calculated *P* value below the preselected level of significance is conclusively determined to be "statistically significant," and the null hypothesis is rejected in favor of the alternate hypothesis. If the *P* value is above the level of significance, the null hypothesis is conclusively not rejected and assumed to be true.

Inevitably, this process leads to 2 potential errors. The first is rejecting the null hypothesis when it is actually true. This is known as a type I error and will occur with a frequency based on the level selected for determining significance (α). If α is selected to be .05, then a type I error will occur 5% of the time. The second potential error is accepting the null hypothesis when it is actually false. This is known as a type II error. The complement of a type II error is to reject the null hypothesis when it is truly false. This is termed the *statistical power* of a study and is the probability that a significance test will detect an effect that truly exists. It is also the basis for calculating sample sizes needed for clinical trials. The objective is to design an experiment to control or minimize both types of errors.[10,11]

The main criticism of the Neyman-Pearsonian approach is the extreme rigidity of thinking and arriving at a conclusion. The researcher must either accept or reject a proposed hypothesis and make a dichotomous scientific decision accordingly based on a predetermined accepted level of statistical significance (eg, $\alpha < .05$). Making decisions with such limited flexibility is usually neither realistic nor prudent. For example, it would be unreasonable to decide that a new cancer medication was ineffective because the calculated *P* value from a phase 2 trial was .051 and the predetermined level of statistical significance was considered to be less than .05.

Statistical and scientific inference need not be constricted by such rigid thinking. A form of inductive inference can be used to assess causal relations with degrees of certainty characterized as spectrums of probabilities.[19] This form of scientific reasoning, known as *Bayesian induction*, is especially useful for both statistical and scientific inferences by which effects are observed and the cause must be inferred. For example, if an investigator finds an association between a particular exposure and a specific health-related outcome, the investigator will infer the possibility of a causal relation based on the findings in conjunction with prior studies that evaluated the same possible causal effect. The degree of inference can be quantified using prior estimations of the effect estimate being evaluated.

The main advantage of Bayesian inductive reasoning is the ability to quantify the amount of certainty in terms of known or estimated conditional probabilities. Prior probabilities are transformed into posterior probabilities based on information obtained and included in Bayesian calculations. The main limitations of Bayesian method is that prior information is often unknown or not precisely quantified, making the calculation of posterior probabilities potentially inaccurate. In addition, calculating Bayes factors (a statistical measure for quantifying evidence for a hypothesis based on Bayesian calculations) as an alternative to *P* values requires additional computational steps.[20,21] In addition, Bayesian methods are often not taught in classical statistics courses. For these reasons, Bayesian methods are not frequently used in most biomedical research analyses.[22] However, scientific inferences based on using both *P* values and Bayesian methods are not necessarily mutually exclusive. Greenland and Poole[22] have suggested incorporating *P* values into modern Bayesian analysis frameworks.

Fundamentally, statistical inference using *P* values involves mathematical attempts to facilitate the development of explanatory theory in the context of random error. However, *P* values provide only a particular mathematical description of a specific data set and not a comprehensive scientific explanation of cause-and-effect relationships in a target population. Each step in the biomedical scientific process should be guided by investigators and biostatisticians who understand and incorporate subject matter knowledge into the research process from prior epidemiologic studies, clinical research, basic science, and biological theory.

With the increasing use of *P* values in the biomedical literature as reported by Chavalarias et al, it becomes critically important to understand the true meaning of the *P* value, including its strengths, limitations, and most appropriate application for statistical inference. Despite more teaching of methods and statistics in clinical medicine and for investigators, the authors' findings that such a small proportion of abstracts reported effect sizes or measures of uncertainly are disappointing. There is nothing inherently wrong when *P* values are correctly used and interpreted. However, the automatic application of dichotomized hypothesis testing based on prearranged levels of statistical significance should be substituted with a more complex process using effect estimates, confidence intervals, and even *P* values, thereby permitting scientists, statisticians, and clinicians to use their own inferential capabilities to assign scientific significance.

ARTICLE INFORMATION

Author Affiliations: Senior Editor, *JAMA*; Department of Emergency Medicine, Northwestern University Feinberg School of Medicine, Chicago, Illinois.

Corresponding Author: Demetrios N. Kyriacou, MD, PhD, *JAMA*, 330 N Wabash, Chicago, IL 60611 (demetrios.kyriacou@jamanetwork.org).

Conflict of Interest Disclosures: The author has completed and submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest and none were reported.

REFERENCES

1. Rothman KJ, Greenland S, Lash T, eds. *Modern Epidemiology*. 3rd ed. Philadelphia, PA: Lippincott Williams & Wilkins; 2008.

2. Yule GU. Notes on the theory of association of attributes in statistics. *Biometrika*. 1903;2:121-134.

3. Berkson J. Limitations of the application of fourfold table analysis to hospital data. *Biometrics*. 1946;2(3):47-53.

4. Mainland D. The rise of experimental statistics and the problems of a medical statistician. *Yale J Biol Med*. 1954;27(1):1-10.

5. Woolson RF, Kleinman JC. Perspectives on statistical significance testing. *Annu Rev Public Health*. 1989;10:423-440.

6. van Houwelingen HC. The future of biostatistics: expecting the unexpected. *Stat Med*. 1997;16(24):2773-2784.

7. Greenland S. Interpretation and choice of effect measures in epidemiologic analyses. *Am J Epidemiol*. 1987;125(5):761-768.

8. Sinclair JC, Bracken MB. Clinically useful measures of effect in binary analyses of randomized trials. *J Clin Epidemiol*. 1994;47(8):881-889.

9. Walter SD. Choice of effect measure for epidemiological data. *J Clin Epidemiol*. 2000;53(9):931-939.

10. Oakes M. *Statistical Inference*. Chestnut Hill, MA: Epidemiology Resources Inc; 1990.

11. Howson C, Urbach P. *Scientific Reasoning: The Bayesian Approach*. La Salle, IL: Open Court; 1989.

12. Goodman SN, Royall R. Evidence and scientific research. *Am J Public Health*. 1988;78(12):1568-1574.

13. Goodman SN. p values, hypothesis tests, and likelihood: implications for epidemiology of a neglected historical debate. *Am J Epidemiol*. 1993;137(5):485-496.

14. Goodman S. A dirty dozen: twelve *p*-value misconceptions. *Semin Hematol*. 2008;45(3):135-140.

15. Chavalarias D, Wallach JD, Li AHT, Ioannidis JPA. Evolution of reporting *P* values in the biomedical literature, 1990-2015. *JAMA*. doi:10.1001/jama.2016.1952.

16. Fisher RA. *Statistical Methods for Research Workers*. Edinburgh, United Kingdom: Oliver & Boyd; 1925.

17. Stang A, Poole C, Kuss O. The ongoing tyranny of statistical significance testing in biomedical research. *Eur J Epidemiol*. 2010;25(4):225-230.

18. Perezgonzalez JD. Fisher, Neyman-Pearson or NHST? a tutorial for teaching data testing. *Front Psychol*. 2015;6:223.

19. Greenland S. Bayesian interpretation and analysis of research results. *Semin Hematol*. 2008;45(3):141-149.

20. Greenland S, Poole C. Living with p values: resurrecting a Bayesian perspective on frequentist statistics. *Epidemiology*. 2013;24(1):62-68.

21. Greenland S. Bayesian perspectives for epidemiological research, I: foundations and basic methods. *Int J Epidemiol*. 2006;35(3):765-775.

22. Bland JM, Altman DG. Bayesians and frequentists. *BMJ*. 1998;317(7166):1151-1160.

Toward High-Reliability Vaccination Efforts in the United States

Matthew M. Davis, MD, MAPP

Infectious disease eradication is a major public health achievement. Smallpox is the only human infectious disease that has been eliminated by deliberate intervention, and this was accomplished using strategic global immunization efforts. The public generally understands that effective vaccination is a key component of infectious disease prevention and eradication; for example, there has been substantial recent public interest in vaccination for emerging global health threats such as avian influenza, Ebola, and Zika virus.

Related article page 1149

Despite public enthusiasm for childhood vaccination,[1] vaccine-preventable diseases such as measles and pertussis are resurgent in the United States. The number of measles cases has increased from no indigenous spread in 2000 to 189 cases in 2015.[2] Pertussis had a nadir of 2900 cases per year in the early 1980s, but has subsequently increased to 18 166 cases in 2015.[2] Although these incidence rates are low compared with the incidence before vaccines were developed for these diseases, each case of vaccine-preventable disease—particularly if it results in death or disability—represents a failed opportunity to prevent disease.

In this issue of *JAMA*, Phadke and coauthors[3] examine the role of vaccine refusals, measured as nonmedical exemptions, in outbreaks of measles and pertussis since the nadir of these diseases in the United States. Their review highlights vaccine exemptions, in the context of recent widely publicized outbreaks such as one originating at Disneyland in December 2014, from which measles spread across the United States.[4,5] However, nonmedical exemptions do not entirely explain recent outbreaks of measles and pertussis. Rather, this review presents 2 key findings that inform future policy considerations regarding disease eradication.

First, Phadke et al reviewed 18 published studies that included 1416 cases of measles since 2000, and found that most cases of measles with known information about exemptions occurred among children whose parents refused measles vaccination. The association of vaccine refusal with the Disneyland measles outbreak led to a change in California law that makes it more difficult for parents to obtain nonmedical exemptions to childhood vaccinations required for daycare and school entry.[6] However, the authors also reviewed 32 reports of pertussis outbreaks that included 10 609 cases that had pertussis vaccination status reported