STATISTICS

Aligning statistical and scientific reasoning

Misunderstanding and misuse of statistical significance impede science

By Steven N. Goodman

Imagine the American Physical Society convening a panel of experts to issue a missive to the scientific community on the difference between weight and mass. And imagine that the impetus for such a message was a recognition that engineers and builders had been confusing these concepts for decades, making bridges, buildings, and other components of our physical infrastructure much weaker than previously suspected.

POLICY

That, in a sense, is what happened with the recent release of a statement from the American Statistical Association (ASA), with the deceptively innocuous title, "ASA statement on statistical significance and p-values" (*1*). The scientific measure in need of clarification was the *P* value—perhaps the most ubiquitous statistical index used in scientific research to help decide what is true and what is not. The ASA saw misunderstanding and misuse of statistical significance as a factor in the rise in concern about the credibility of many scientific claims (sometimes called the "reproducibility crisis") and is hoping that its official statement on the matter will help set scientists on the right course.

The formal definition of *P* value is the probability of an observed data summary (e.g., an average) and its more extreme values, given a specified mathematical model and hypothesis (usually the "null"). The problem is that this index by itself is not of particular interest. What scientists want is a measure of the credibility of their conclusions, based on observed data. The *P* value neither measures that nor is it part of a formula that provides it.

This confusion between the index we have and the measure we want produces misconceptions that the *P* value is the probability that the null hypothesis is true or that the observed data occurred by chance—different ways of saying the same thing (*2*, *3*). This pernicious error creates the illusion that the *P* value alone measures the credibility of a conclusion, which opens the door to the mistaken notion that the dividing line between scientifically justified and unjustified claims is set by whether the *P* value has crossed the "bright line" of significance, to the exclusion of external considerations like prior evidence, understanding of mechanism, or experimental design and conduct.

Bright-line thinking, coupled with attendant publication and promotion incentives, is a driver behind selective reporting: cherry-picking which analyses or experiments to report on the basis of their *P* values. This in turn corrupts science and fills the literature with claims likely to be overstated or false. We cannot solve these problems without understanding how we got to this point.

R. A. Fisher revolutionized statistical inference and experimental design in the 1920s and '30s by establishing a comprehensive framework for statistical reasoning and writing the first statistical best-seller for experimenters. He formalized an approach to inference involving *P* values and assessment of significance, based on the frequentist notion of probability, defined in terms of verifiable frequencies of repeatable events. He wanted to avoid the subjectivity of the Bayesian approach, in which the probability of a hypothesis ("inverse probability"), neither repeatable nor observable, was central.

Fisher was a champion of *P* values as one of several tools to aid the fluid, inductive process of scientific reasoning—not to substitute for it. Fisher used "significance" merely to indicate that an observation was worth following up, with refutation of the null hypothesis justified only if further experiments "rarely failed" to achieve significance (*4*). This is in stark contrast to the modern practice of making claims based on a single demonstration of statistical significance.

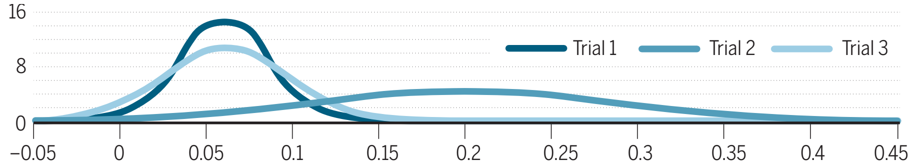
In their development of "hypothesis testing" in the 1930s, Jerzy Neyman and Egon

Departments of Medicine and of Health Research and Policy, Stanford University School of Medicine, Division of Epidemiology, Meta-research Innovation Center at Stanford (METRICS), Stanford, CA 94305, USA. Email: steve.goodman@stanford.edu

A. How can these data be interpreted?

APPROACH	QUESTION	STATISTICAL ANALYSIS	INTERPRETATION
Hypothesis test ("bright line")	Should we act as though the observed effect is nonzero (given prespecified error rates)?	1. $P \le 0.05$ 2. $P \le 0.05$ 3. NS	Studies 1 and 2 indicate action based on a nonzero true effect is justified. Study 3 indicates it is not.
Fisherian *P* value	How much evidence is there that the true effect is different from zero?	1. $P = 0.03$ 2. $P = 0.05$ 3. $P = 0.11$	Studies 1 and 2 provide moderate, statistically significant evidence that the new treatment is better. Study 3 supplies weak but insufficient evidence to say the treatment is effective.
Estimation	What range of true effects is statistically consistent with the observed effects?	Effect, 95% confidence interval (%) 1. 6, 0.5 to 12 2. 20, 2.5 to 38 3. 6, –1.4 to 13	Studies 1 and 3 indicate the new treatment had a small to moderate effect. Study 2 is consistent with either small or large effects.
Bayes factor	How strongly do the data support a large, clinically important effect (10 to 25%) versus a small, unimportant one (0 to 10%)?	Bayes factor, large:small effect 1. 1:14 2. 3:1 3. 1:7	Studies 1 and 3 together decrease odds of an important effect 98 fold (1/7 x 1/14 = 1/98), Study 2 increases odds 3 fold, for net 33 fold decrease (3 x 1/98 = 1/33).

B. Likelihood functions



Changing questions, changing answers. Three randomized trials show response rates of 20% in the control arm and rates in the treatment arms of (1) 26% ($n = 900$), (2) 40% ($n = 100$), and (3) 26% ($n = 500$). The effect deemed clinically important is 10%. (**A**) Each statistical approach asks a different question, hence interpretations are different. Scientists must decide which statistical question best matches their scientific question. (**B**) Likelihood functions, proportional to the probability of the observed data (vertical axis) under each possible true effect (horizontal axis), measure how strongly the observed effects support different true effects (which cannot be directly observed).

GRAPHIC: C. SMITH/SCIENCE

Pearson went where Fisher was unwilling to go (*5*). In a hypothesis test, one specifies a null statistical hypothesis and an alternative, and is to "reject" the null and "accept" the alternative—or vice versa—on the basis of whether an estimate falls into a prespecified region defined by two error rates: type I (alpha, false positive) and type II (beta, false negative). Once these error rates are set, scientific reasoning is effectively out of the picture (see the figure). Judgment ideally enters through customization of the alternative hypothesis and the error rates, contingent on the seriousness of each kind of error.

The Neyman-Pearson method did not use *P* values, but was combined with the Fisherian *P*-value approach in textbooks and research articles (*6*, *7*). Without foundational justification, this created the illusion that quantitative inference could be automated, with hypothesis rejection determined by whether the *P* value is less than the type I error, set at 5% in most sciences today. This combination did violence to both approaches, particularly to Fisher's. He vehemently opposed using *P* values for automatic inference, referring to hypothesis tests disparagingly as "decision functions" or "acceptance procedures." His dismay was pointed and prescient:

> [N]o scientific worker has a fixed level of significance at which from year to year, and in all circumstances, he rejects hypotheses; he [examines] each particular case in the light of…evidence and…ideas [p. 42 of (*8*)].
>
> The concept that the scientific worker can regard himself as an inert item in a vast co-operative concern working according to accepted rules, is encouraged by directing attention away from his duty to form correct scientific conclusions,…and by stressing his supposed duty to mechanically make a succession of automatic "decisions"…. The idea that this responsibility can be delegated to a giant computer programmed with Decision Functions belongs to a phantasy of circles, rather remote from scientific research [pp. 104–105 (*8*)].

Sixty years later, we have the ASA expressing the same sentiment:

> Researchers should bring many contextual factors into play to derive scientific inferences, including the design of a study, the quality of the measurements, the external evidence for the phenomenon under study, and the validity of assumptions that underlie the data analysis…. The widespread use of "statistical significance" (generally interpreted as "p ≤ 0.05") as a license for making a claim of a scientific finding (or implied truth) leads to considerable distortion of the scientific process (*1*).

The concordance of these statements, separated by over half a century, underscores lack of progress in approaches to statistical inference in the applied literature, despite advances in statistical methodology. This is due in part to the way statistical inference is taught to scientists; not as a variety of named, competing approaches, each with strengths and weaknesses, but as anonymized procedures, universally applicable, seemingly without controversy or alternatives (*6*, *7*).

Contrast this situation with other sciences. In any high-school physics textbook, one will find theories and models by Copernicus, Galileo, Newton, Einstein, and so on. Students are trained to understand the incomplete explanatory power of each theory, the controversies, why new theories were accepted (or not), and what questions they raised.

Theories of statistical inference are no less nuanced or contested, as evidenced by the 23 commentaries that followed the ASA statement (*1*). But such controversy, rarely taught in applied courses or texts, is unappreciated by most who use statistical tools. This seeming absence of controversy about the foundations of these methods has fostered growth of social-scientific structures reifying those values—enshrined in journal practices, promotion, and funding criteria, as well as in the standard discourse of science—which makes them extraordinarily difficult to change.

"Theories of statistical inference are…nuanced [and] contested…."

Another reason these practices persist is that, until the recent rise of concern about research reproducibility (*9*), the scientific community has perceived few adverse consequences from their use. Many papers over the past century have issued cautions similar to those of the ASA, but have largely been ignored by the general scientific community. Benefits of having seemingly objective rules have outweighed theoretical cavils (*6*, *10*).

The ASA suggested several ways to improve statistical interpretation, including more complete reporting of all analyses performed, and a number of alternative inferential approaches. One of these, Bayes factors (*11*, *12*), is a measure derived from Bayes theorem indicating how strongly the data should shift belief toward one hypothesis versus another. If we were told that the experimental results lowered the prestudy odds of the null hypothesis by a factor of 4, this would lead to a far different reasoning process than does "$P = 0.03$," which is difficult to combine with external knowledge (*11*) (see the chart).

Bayes factors and fully Bayesian analyses are not without their own complications (*10*, *12–15*), as are all other recommended approaches. But, if they were more widely used, rules would evolve. That said, no *P*-value alternatives will solve the problems noted by the ASA if they are used in bright-line fashion, such as applying a confidence interval only to see if it includes the null value.

P values are unlikely to disappear, and the ASA did not recommend their elimination—rather, a change in how they are interpreted and used. But how can scientists follow the ASA (and Fisher's) dictates to combine them with contextual factors? There are few examples in the scientific literature. How many papers explain why, in one context, a finding with $P = 0.006$ is insufficient to make a claim, whereas, in another, $P = 0.08$ might be all that is needed (*11*)? Any attempt to do that in an individual research paper would likely meet resistance from reviewers or editors.

The field of genomics has shown us that evidential thresholds are changeable within disciplines, with $P \leq 10^{-8}$ now sought for claiming relations derived from genome-wide scanning. Thresholds in physics are far lower than the $P \leq 0.05$ level used in biomedicine and the social sciences. Whether such thresholds can or should be modified by design, by discipline, or by individual study are rich areas for future exploration (*16*).

Science has progressed dramatically over the past 90 years, despite these issues. How much faster and more efficiently can it proceed if new statistical approaches to inference are adopted, and if optimal statistical and scientific practices are aligned with incentive structures? The ASA has posed a challenge to all who use statistical measures to justify their claims. Let us hope the next century will see as much progress in the inferential methods of science as in its substance. ■

REFERENCES

1. R. L. Wasserstein, N. A. Lazar, *Am. Stat.* 10.1080/00031305.2016.1154108 (2016).
2. S. Goodman, *Semin. Hematol.* **45**, 135 (2008).
3. D. R. Cox, *Br. J. Clin. Pharmacol.* **14**, 325 (1982).
4. R. A. Fisher, *J. Min. Agric. Great Britain* **33**, 503 (1926).
5. J. Neyman, E. S. Pearson, *Philos. Trans. R. Soc. Lond. A* **231**, 289 (1933).
6. G. Gigerenzer *et al.*, *The Empire of Chance* (Cambridge Univ. Press, Cambridge, 1989).
7. G. Gigerenzer, J. N. Marewski, *J. Manage.* **41**, 421 (2015).
8. R. A. Fisher, *Statistical Methods and Scientific Inference* (Hafner, New York, ed. 1, 1956).
9. F. S. Collins, L. A. Tabak, *Nature* **505**, 612 (2014).
10. B. Efron, *Am. Stat.* **40**, 1 (1986).
11. S. N. Goodman, *Ann. Intern. Med.* **130**, 1005 (1999).
12. R. E. Kass, A. E. Raftery, *J. Am. Stat. Assoc.* **90**, 773 (1995).
13. S. Greenland, C. Poole, *Epidemiology* **24**, 62 (2013).
14. H. Hoijtink, P. van Kooten, K. Hulsker, *Multivariate Behav. Res.* **51**, 2 (2016).
15. R. D. Morey, E. J. Wagenmakers, J. N. Rouder, *Multivariate Behav. Res.* **51**, 11 (2016).
16. V. E. Johnson, *Proc. Natl. Acad. Sci. U.S.A.* **110**, 19313 (2013).

10.1126/science.aaf5406