The proof is not in the $P$ value

To the Editor: No doubt you have received a flood of letters concerning your new policies on reporting statistics in manuscripts submitted for publication (2). The new guidelines calling for the reporting of confidence intervals and the interpretation of results in light of them are to be cheered. Ending the reliance on simple null hypothesis tests and requiring researchers more latitude based on effect size have been major goals in the behavioral sciences (1, 6). To see these recommendations adopted in the biological sciences is heartening.

However, despite the laudable inclusion of confidence intervals, there are some apparent misunderstandings in the new policy. Guideline 2 (2) calls for all authors to define $\alpha$ before the study. This recommendation is based on the following assertion “For any statistical test, if the achieved significance level $P$ is less than the critical significance level $\alpha$, defined before any data are collected, then the experimental effect is likely to be real” (2). Although there have been calls for more flexible levels for $\alpha$ (4), the reason cited for setting a given level for $\alpha$ is incorrect.

It is well and good to define $\alpha$ before the conduct of research, but neither $\alpha$ nor the particular $P$ value calculated during a study carry any information about whether “the experimental effect is likely to be real.” Rather, the true purpose for setting a given level for $\alpha$ is to fix the rate of type I errors, while the $P$ value is an index of the likelihood of the experimental data, assuming that the null hypothesis is true (1, 3).

The distinction between the type I error rate, the $P$ value, and the truth of a given alternative hypothesis is slippery. Oakes (5), for example, conducted an informal survey of 70 academic psychologists who used statistical tests on a regular basis. Only one correctly identified the $P$ value as a statement of conditional probability without further implications. This common misunderstanding is further exemplified in Guideline 10 of the Editorial (2).

Although basing interpretation of results on confidence intervals is progressive relative to using null hypothesis testing, the advice given in Table 1 of the Editorial (2) regarding the interpretation of $P$ values is incorrect. For example, Table 1 states that for $P > 0.10$, the “data are consistent with a true zero effect” (2). If the alternative hypothesis is true, then the $P$ value calculated in a null test is more likely to be small; however, it does not follow from this that a large $P$ value indicates an absence of an effect. Without knowing the statistical power of a design, $P > \alpha$ could be consistent with both the null and the alternative hypotheses. Indeed, the $P$ value is not an index of the truth of the null or the proposed alternative. Nor is it an index of whether a particular result will be replicated (5). “Good evidence” that a hypothesized effect is real comes from replication across multiple studies and cannot be inferred from the result of a single statistical test.

REFERENCES


Thomas Koehnle
Department of Neuroscience
University of Pittsburgh
446 Crawford Hall
Pittsburgh, PA 15260
E-mail: koehnle@bns.pitt.edu

REPLY

To the Editor: We appreciate the letter from Thomas Koehnle, but we must emphasize that we published guidelines, not formal policies, for reporting statistics in journals published by the American Physiological Society (1). To be honest, we have received just a trickle of letters about our guidelines. We are delighted that many of the people who have congratulated us are statisticians.

We agree that the proof is not in the $P$ value, but we are confused about why Koehnle chose that pithy phrase as the title of his letter. His title implies that the guidelines encourage researchers to focus on $P$ values when they interpret and report their results. The guidelines do nothing of the sort: in fact, they reiterate that $P$ values have a limited role in data analysis (Guideline 10, paragraph 2). Koehnle cites a review (2) written by one of us in which limitations of $P$ values are discussed in detail.

Koehnle is concerned about our discussion of concepts embedded in two guidelines: the critical significance level $\alpha$ (Guideline 2) and the achieved significance level $P$ (Guideline 10). We are confused about the grounds for each of these concerns.

Koehnle appears to believe that we justify Guideline 2 with a summary of the comparison of the achieved significance level $P$—the result—to the critical significance level $\alpha$—the benchmark. If so, he is mistaken. The summary has nothing to do with the actual choice of the critical significance level $\alpha$.

In Guideline 2 (paragraphs 2 and 3), we illustrate situations where the appropriate $\alpha$ can differ from the traditional value of 0.05. As Koehnle writes, one factor that impacts the choice of $\alpha$ is how often you are willing to declare an effect exists when it does not: that is, how often you are willing to reject a null hypothesis when it is true (a type I error). Koehnle neglects to mention that the second factor that impacts the choice of $\alpha$ is how often you are willing to fail to declare an effect exists when it does: that is, how often you are willing to fail to reject a null hypothesis when it is false (a type II error). Guideline 2 illustrates the choice of the critical significance level $\alpha$ by discussing the relative importance of both types of errors.

Koehnle’s second concern is that Table 1 in Guideline 10, adapted from suggestions made by the eminent statistician Sir David Cox, provides incorrect guidance about the interpretation of $P$ values. On what grounds does Koehnle base this claim? On the valid grounds of statistical power. Guideline 10
states, however, that the interpretations listed in Table 1 are useful only if the power of the study was large enough to detect the experimental effect. This applies primarily to those \( P \) values for which \( P > \alpha \). We stand by the interpretation of \( P \) values provided in Table 1.

We believe the guidelines harbor none of the misunderstandings about which Koehnle is justifiably concerned. Instead, we believe the guidelines offer a concise, accurate framework that we hope will help improve the caliber of statistical information reported in articles published by the American Physiological Society.

REFERENCES


Douglas Curran-Everett
Division of Biostatistics
National Jewish Medical and Research Center
and Depts. of Preventive Medicine and Biometrics
and of Physiology and Biophysics
University of Colorado Health Sciences Center
Denver, CO 80262

Dale J. Benos
Department of Physiology and Biophysics
University of Alabama at Birmingham
Birmingham, AL 35294