The Value of p-Values

P.B. Stark

I agree with the spirit of the ASA p-value statement, but I disagree with some of the content, for instance:

- The informal definition of a p-value at the beginning of the document is vague and unhelpful.[1]
- The statement draws a distinction between "the null hypothesis" and "the underlying assumptions" under which the p-value is calculated. But the null hypothesis *is* the complete set of assumptions under which the p-value is calculated.
- The "other approaches" section ignores the fact that the assumptions of some of those methods are identical to those of p-values. Indeed, some of the methods use p-values as input (e.g., the False Discovery Rate).
- The statement ignores the fact that hypothesis tests apply in many situations in which there is no parameter or notion of an "effect," and hence nothing to estimate or to calculate an uncertainty for.
- The statement ignores the crucial distinction between frequentist and Bayesian inference.[2]

I offer the following plainer-language alternative:

Science progresses in part by ruling out potential explanations of data. p-values help assess whether a given explanation is adequate. The explanation being assessed is often called "the null hypothesis."[3]

If the p-value is small, either the explanation is wrong, or the explanation is right but something unlikely happened—something that had a probability equal to the p-value.[4] Small p-values are stronger evidence that the explanation is wrong: the data cast doubt on that explanation.

If the p-value is large, the explanation accounts for the data adequately—although the explanation might still be wrong.[5] Large p-values are not evidence that the explanation is right: lack of evidence that an explanation is wrong is not evidence that the explanation is right. If the data are few or low quality, they might not provide much evidence, period.

There is no bright line for whether an explanation is adequate: scientific context matters.

A p-value is computed by *assuming* that the explanation is right. The p-value is *not* the probability that the explanation is right.[6]

p-values do not measure the size or importance of an effect, but they help distinguish real effects from artifacts. In this way, they complement estimates of effect size and confidence intervals.

Moreover, p-values can be used in some contexts in which the notion of "effect size" does not make sense. Hence, p-values may be useful in situations in which estimates of effect size and confidence intervals are not.

Like all tools, p-values can be misused. One common misuse is to hunt for explanations that have small p-values, and report only those, without taking into account or reporting the hunting. Such "p-hacking," "significance hunting," selective reporting, and failing to account for the fact that more than one explanation was examined ("multiplicity") can make the reported p-values misleading.

Another misuse involves testing "straw man" explanations that have no hope of explaining the data: null hypotheses that have little connection to how the data were collected or generated. If the explanation is unrealistic, a small p-value is not surprising. Nor is it illuminating.

Many fields and many journals consider a result to be scientifically established if and only if a p-value is below some threshold, such as 0.05. This is poor science and poor statistics, and creates incentives for researchers to "game" their analyses by p-hacking, selective reporting, ignoring multiplicity, and using inappropriate or contrived null hypotheses.

Such misuses can result in scientific "discoveries" that turn out to be false or that cannot be replicated. This has contributed to the current "crisis of reproducibility" in science.

Online discussion of the ASA Statement on Statistical Significance and P-Values, *The American Statistician*, 70. P.B. Stark, University of California, Berkeley–Statistics, Berkeley, CA 94720-3860 (Email: pbstark@berkeley.edu).

[1] See footnote 4 below. The reference to "extreme" values of "a statistical summary" limits the scope to tests based on a test statistic. It is an inaccurate and confusing substitute for a simpler statement about monotonicity (i.e., nesting) of rejection regions.

[2] The document has other problems, among them: It characterizes a p-value of 0.05 as "weak" evidence against the null hypothesis, but strength of evidence depends crucially on context. It categorically recommends using multiple numerical and graphical summaries of data, but there are situations in which these would be gratuitous distractions—if not an invitation to p-hacking!

[3] The use of the term "null hypothesis" is not entirely consistent, but in general, the null hypothesis asserts that the probability distribution $\mathbb{P}$ of the data X is in some specified set $\mathcal{P}$ of probability distributions on a measurable space $\mathcal{X}$. A "point null hypothesis" or "simple null hypothesis" completely specifies the probability distribution of the data, i.e., $\mathcal{P}$ is a singleton set. In the context of testing whether some parameter θ is equal to θ_0, some authors write $H_0 : \theta = \theta_0$ as the null hypothesis. This is (perhaps not deliberate) shorthand for the hypothesis $X \sim \mathbb{P}_{\theta_0}$, where $\{\mathbb{P}_\theta\}_{\theta\in\Theta}$ is a pre-specified family of probability distributions on $\mathcal{X}$ that depends on a parameter θ known a priori to be in the set Θ, which contains θ_0.

[4] The simplest general definition of a p-value of a point null hypothesis I know of is as follows. Suppose the null hypothesis is that $\mathbb{P}$ is the probability distribution of the data X, which takes values in the measurable space $\mathcal{X}$. Let $\{R_\alpha\}_{\alpha\in[0,1]}$ be a collection of $\mathbb{P}$-measurable subsets of $\mathcal{X}$ such that (1) $\mathbb{P}(R_\alpha) \le \alpha$ and (2) If $\alpha' < \alpha$ then $R_{\alpha'} \subset R_\alpha$. Then the p-value of H_0 for data $X = x$ is $\inf_{\alpha\in[0,1]}\{\alpha : x \in R_\alpha\}$.

[5] Here, "adequately" is with respect to the chosen test.

[6] This is a common misinterpretation. Other misinterpretations are that 1 minus the p-value is the probability that the *alternative hypothesis* (a different explanation of the data) is true, and that the p-value is the probability of observing the data "by chance."