Making sense of P-values

P.B. Stark, http://www.stat.berkeley.edu/~stark

Draft: 24 November 2015

URL: http://www.stat.berkeley.edu/~stark/Preprints/pValues.pdf

Science progresses in part by ruling out potential explanations of data.

P-values are a tool for this task.

P-values measure the adequacy of a particular explanation for a particular set of data. The explanation being examined is often called "the null hypothesis."[1]

If the p-value is small, either the explanation is wrong, or the explanation is right but something unlikely happened: something that has a probability equal to the p-value.[2]

Smaller p-values are stronger evidence that the explanation is wrong: the data cast doubt on that explanation. However, larger 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.

Nor is the p-value is the probability that the explanation is right, a common misinterpretation. Indeed, the p-value is computed by *assuming* that the explanation is right.

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

P-values do not measure the size or practical importance of an effect, but they help assess whether an apparent effect is an artifact, rather than real. In this way, they are complementary to estimates of effect size and to confidence intervals.

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

[1] There is some inconsistency in the use of the term "null hypothesis." In general, the null hypothesis asserts that the probability distribution P of the data X is in some set $\mathcal{P}_0$ 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. In the context of testing whether some real-valued parameter θ is equal to zero, some authors write $H_0 : \theta = 0$ as the null hypothesis. This is (perhaps not deliberate) shorthand for $X \sim P_0$, where $\{P_\theta\}_{\theta \in \Theta}$ 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 $\Theta \subset \Re$.

[2] The simplest mathematically rigorous definition of a p-value is as follows. Let P be 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 P-measurable subsets of $\mathcal{X}$ such that (1) $P(R_\alpha) = \alpha$ and (2) If $\alpha' < \alpha$ then $R_{\alpha'} \subset R_\alpha$. Then the p-value of H_0 for data $X = x$ is $\inf_{\alpha \in [0,1]}\{\alpha : x \in R_\alpha\}$.

Appropriate use of p-values can help distinguish scientific discoveries from accidental artifacts. But misuses can create false confidence.

One common misuse is to hunt for explanations that have small p-values, but not take into account or report the hunting, and to report only the p-values that are small. 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 p-value misleading.

Another misuse involves proposing “straw man” explanations that have little hope of explaining the data: null hypotheses that have little to do with how the data were collected or generated. It is then unsurprising to find a small p-value.

In many fields and many journals, there is a practice of comparing p-values to a fixed threshold, such as 0.05, and considering a result to be scientifically established if and only if a p-value is below that threshold. This is poor science and poor statistics, and produces incentives for researchers to “game” their analyses through 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 so-called “crisis of reproducibility.”