Why are p-values controversial?

Todd A. Kuffner *
Department of Mathematics, Washington University in St. Louis
St. Louis, MO 63130
e-mail: kuffner@wustl.edu
and
Stephen G. Walker
Department of Mathematics, University of Texas at Austin
Austin, TX 78712-1823
email: s.g.walker@math.utexas.edu

October 16, 2016

Abstract

While it is often argued that a p-value is a probability; see Wasserstein & Lazar (2016), we argue that a p-value is *not* defined as a probability. A p-value is a bijection of the sufficient statistic for a given test which maps to the same scale as the type I error probability. As such, the use of p-values in a test should be no more a source of controversy than the use of a sufficient statistic. It is demonstrated that there is, in fact, no ambiguity about what a p-value is, contrary to what has been claimed in recent public debates in the applied statistics community. We give a simple example to illustrate that rejecting the use of p-values in testing for a normal mean parameter is conceptually no different from rejecting the use of a sample mean. The p-value is innocent; the problem arises from its misuse and misinterpretation. The way that p-values have been informally defined and interpreted appears to have led to tremendous confusion and controversy regarding their place in statistical analysis.

Keywords: sufficient statistic; type I error; decision rule

*The authors are grateful to the Editor and Associate Editor for helpful suggestions which led to improvements in this article.

1 Introduction

The use of the p-value is ubiquitous in science. They are widely used and widely mis-used. Large swathes of decisions and conclusions are being drawn from p-values, which are ever increasingly, it seems, being dragged away from their real interpretation. The gap has been caused by the removal of the idea of a decision rule from a test. It may be convenient to conduct a test of a hypothesis without a formal decision rule, but it leads to a test outcome with no clear idea of the errors to which the experimenter is being exposed to and hence the outcome becomes a heuristic, no better than the implementation of a rule of thumb.

This note specifically is in response to the ongoing public debate in the applied statistics community caused by an editorial decision at *Basic and Applied Social Psychology* (BASP) to ban the use of p-values (Trafimow & Marks, 2015). This headline-grabbing decision elicited responses from the International Society for Bayesian Analysis (ISBA) (Schmidt et al., 2015) and the American Statistical Association (ASA) (Wasserstein & Lazar, 2016), among others. The subsequent fallout has even led some to claim that the p-value has "an ambiguous interpretation" (Demidenko, 2016). We explain why this is not the case.

In statistical decision theory and according to the principles of statistical inference, a testing procedure is not valid unless there is a decision rule specified in advance, independently of the data. Conventionally, the decision rule is chosen to put an upper-bound on the probability of incorrectly rejecting a true null hypothesis, which we call a type I error probability, α. The type I error probability determines the decision rule, whereas a p-value, just like any other test statistic, determines the decision. The type I error probability and the p-value are two different things. The controversy exists because p-values are being used as decision rules, even though they are data-dependent, and hence cannot be formal decision rules. Incorrectly using p-values as decision rules effectively eliminates the idea of a valid decision rule from a test, and therefore invalidates the decision.

The debate about p-values has, in our view, strayed into philosophical discussions which can potentially further complicate and confuse the issue. Moreover, the way in which p-values are often taught and used in practice suggests a tendency to view p-values as something disconnected from their actual definition. We first remind readers of key concepts and the *formal* definition of a p-value, and clarify when the p-value is a valid statistical

inferential quantity. In particular, we distinguish between the mathematical, statistical and philosophical meaning of p-values, ignoring the latter in order to bring clarity to precisely why the use of p-values is not controversial in settings where their usage conforms to basic mathematical and statistical principles.

2 The definition of a p-value as a bijection

Consider testing a hypothesis, H, which is either true or false. The goal is to make a decision about this hypothesis, where the two possible decisions are to reject H or fail to reject H. The decision is to be based on the observed values of a random sample, $X = \{X_1, \ldots, X_n\}$, assumed i.i.d. according to some population distribution. For more on the ideas of testing a hypothesis, see Morris & Larsen (2006) and Lehmann & Romano (2005).

Adhering to statistical principles, the test statistic associated with the hypothesis H should be a function of a sufficient statistic, $S(X)$, which is a function of the random sample. For illustrative purposes, we suppose the dimension of the sufficient statistic is 1, so that $S(X)$ takes values, for example, on the real line. Similar arguments apply in more general settings. The two possible decisions split the real line into two complementary regions, say R and R^C. These regions are such that if $S(X) \in R$, the hypothesis is rejected, while if $S(X) \in R^C$, the hypothesis is not rejected.

Associated with any such test is a type I error, which occurs when H is rejected yet H is true. It is conventional to control the probability of a type I error. This type I error probability is some number, say α, such that $0 < \alpha < 1$.

The idea of a p-value is to transform the sufficient statistic $S(X)$ to be on the same scale as the type I error probability. Such a transformation would be done for convenience. It neither adds to nor takes away from the information provided by the value of the sufficient statistic.

As a function of a sufficient statistic for a test of some hypothesis, H, a p-value is also a sufficient statistic for that same test. It is instructive to now examine the p-value as a mathematical object. Namely, the p-value is a bijection from $\mathbb{R}$ to $(0, 1)$. See Gerstein (2012) for more on the notion of a bijection.

Let α be a number such that $0 < \alpha < 1$, and let $R_\alpha \equiv R(\alpha)$. Though not strictly necessary, it may also be convenient to imagine the simple setting in which R_α has the form $[c_\alpha, \infty)$, i.e. the test rejects H if $S(X) \geq c_\alpha$. The p-value is defined in settings where the rejection regions are nested sets in the sense that

$$\alpha < \alpha' \Rightarrow R_\alpha \subset R_{\alpha'}.$$

Define a p-value as (Lehmann & Romano, 2005, §3.3)

$$\hat{\alpha} \equiv \hat{\alpha}_{S(X)} = \inf_{0<\alpha<1} \{\alpha : S(X) \in R_\alpha\}.$$

We want to show that the p-value, $\hat{\alpha} = f(S(X))$ for suitable choice of map $f : S(X) \mapsto \hat{\alpha}$, is a bijection from $\mathbb{R}$ to $(0, 1)$. The actual form of $\hat{\alpha} = f(S(X))$ is specific to the model, hypothesis and test.

Write $S \equiv S(X)$ for simplicity. First, we note that the function $\hat{\alpha}$ is well-defined. That is, given two values S_1, S_2 such that $S_1 = S_2$, we have that $\hat{\alpha}_{S_1} = \hat{\alpha}_{S_2}$.

Next, we require that $\hat{\alpha}$ is injective (one-to-one). To show this, we need that if $\hat{\alpha}_{S_1} = \hat{\alpha}_{S_2}$, then $S_1 = S_2$. Suppose that $S_1 \neq S_2$. If we can show that this implies $\hat{\alpha}_{S_1} \neq \hat{\alpha}_{S_2}$, this will establish injectivity. Without loss of generality, suppose $S_2 < S_1$. Then there exists an α' such that $S_1 \in R_{\alpha'}$ but $S_2 \notin R_{\alpha'}$. Therefore, if $S_2 < S_1$, it cannot be the case that $\hat{\alpha}_{S_1} = \hat{\alpha}_{S_2}$.

Finally, we must have that $\hat{\alpha}$ is surjective (onto). For surjectivity, we require that for every $\beta \in (0, 1)$, there exists an $\tilde{S} \in \mathbb{R}$ such that

$$\hat{\alpha} = \inf_{0<\alpha<1} \{\alpha : \tilde{S} \in R_\alpha\} = \beta,$$

which is seen by choosing $\tilde{S} = \inf\{S : S \in R_\beta\}$.

Crucially, the p-value is not itself defined as a probability, but rather it takes values on the same scale as something which is formally defined as a probability. This subtle point is not irrelevant.

What do we gain from this way of thinking? We argue that stripping away the informal interpretation of the formal definition of a p-value clarifies that using p-values in testing is conceptually no different from using tests based on sufficient statistics. In broader generality, the point is that the p-value is equal in information to the sufficient statistic for the

test in question. The use of sufficient statistics in testing is an uncontroversial practice, firmly rooted in statistical principles. Thus, any viewpoint implying that the p-value is controversial must stem from a misunderstanding of it. Just because the p-value is on the same scale as the type I error probability does not mean that using p-values is any different from using a t-test, z-score or likelihood ratio test. The p-value carries the same information about the hypothesis in question as does the sufficient statistic. The p-value is not the reason why tests are fallible; all tests can be wrong, and this is why the type I error is desired to be small.

3 Example

The point is easily demonstrated in the following simple, commonly-encountered example. Suppose we have a sample, $X = X_1, \ldots, X_n \overset{i.i.d.}{\sim} \mathcal{N}(\theta, 1)$, and interest is centered on testing a hypothesis of the form $H : \theta = 0$. Assuming that a type I error α is set in advance, a sufficient statistic for this test is the sample mean $S(X) = \bar{X} = n^{-1} \sum_{i=1}^{n} X_i$. The p-value is simply a bijection of the sufficient statistic $\bar{X}$ to the same scale as the type I error probability, i.e. the p-value assigns a value between 0 and 1 to each value in the sample space of $\bar{X}$. As such, a p-value is merely a transformation of the sufficient statistic for the test, $\bar{X}$, and therefore is also a sufficient statistic for the test. In this example, rejecting the existence and interpretation of a p-value is conceptually equivalent to rejecting the use and interpretation of $\bar{X}$.

In such an example, if one is concerned that an investigator is only publishing a p-value and from its value claiming that the hypothesis needs to be rejected, the statement in full should be that he/she was performing the test with a *pre-specified* type I error α which turned out to be larger than the p-value. It is instructive to understand that no test, or decision, can be performed without such an α - it is the probability of rejecting H when it is true - always a possibility.

4 The source of the p-value controversy

So why exactly is there still a debate about this? To understand the real source of confusion, we must acknowledge that there are three types of 'testers' in the applied statistics community.

Tester 1 Sets a type I error probability α before seeing the data, then computes the observed value of p based on the realization of the random sample, and subsequently rejects H if $p < \alpha$.

Tester 2 First computes the observed value of p, and then claims that his/her α would have been bigger had he/she actually chosen one beforehand.

Tester 3 First computes the observed value of p, believes it is small, and subsequently rejects H. He/she believes the type I error is actually the observed value of p, since continuing with Tester 2's approach, any $\alpha > p$ will work. Therefore, he/she argues, why not choose an α just above p and view that as the type I error probability?

Testers 2 and 3 are not adhering to statistical principles; namely, these approaches are tantamount to selecting the type I error probability based on the data. This invalidates the test.

The idea of a test is that a valid type I error is available. Now, Tester 1 has a type I error set at α. On the other hand, Tester 2 does not have a type I error set, but if he/she believes it could be anything above the observed p value, the smallest value of which is p itself, then the existence of Tester 3 follows. However, the type I error is not p.

In fact, for Testers 2 and 3 there is no *decision rule* for the decision. For more on decision theory, see Raiffa & Schlaifer (1961). There is in its place a *heuristic*, which is that suspicion of a small p-value is sufficient to reject the hypothesis. Decisions require decision rules, and the rule is to reject H if $S(X) \in R_\alpha$, with R_α set in advance. This is the decision rule. Then the type I error is the probability, i.e. α, that $S(X)$ lies in R_α, and yet the hypothesis is true. This procedure exists and is well defined even without the notion of a p-value. As we have mentioned, the p-value is a version of the rule which is to reject H if $p < \alpha$.

5 Discussion

The p-value does not, in fact, have an ambiguous mathematical meaning. The claim that the p-value has an ambiguous interpretation is likely a symptom of widespread misunderstanding of statistical inference and decision theory. If there is a decision rule for the test in question, and a valid type I error exists, then the p-value is a valid inferential tool, which neither adds to nor dilutes from the information provided by the sufficient statistic of the test. The p-value is innocent, but woefully misunderstood. To properly interpret p-values, one must understand their actual role in a valid testing procedure. The p-value is not a valid decision rule, as it is not a type I error probability. Rather, the p-value is a conveniently-scaled test statistic, which is computed after a decision rule is specified, and which determines the decision. As tempting as it is, especially when teaching non-mathematical undergraduates, to immediately jump to an intuitive explanation of a p-value as a probability, we must recognize that this simplification is undermining the legitimacy of statistical methods in the sciences.

References

Demidenko, E. (2016), "The p-value you can't buy," *American Statistician* **70**(1), 33–38.

Gerstein, L. (2012), *Introduction to Mathematical Structures and Proofs*, Springer Science & Business Media.

Lehmann, E. L., and Romano, J. P. (2005), *Testing Statistical Hypotheses*, Third Edition, New York: Springer.

Morris, M., and Larsen, R. J. (2006), *Introduction to Mathematical Statistics and its Applications*, New Jersey: Pearson/Prentice Hall.

Raiffa, H., and Schlaifer, R. (1961), *Applied Statistical Decision Theory*, Cambridge: Harvard Business School.

Schmidt, A. M., Berger, J. O., Dawid, A. P., Kadane, J. B., O'Hagan, T., Pericchi, L. R.,

Robert, C. P., and Szucs, D. (2015), "Banning null hypothesis significance testing," *ISBA Bulletin* **22**(1), 5–9.

Trafimow, D., and Marks, M. (2015), Editorial, *Basic and Applied Social Psychology*, **37**(1), 1–2.

Wasserstein, R. L., and Lazar, N. A. (2016), "The ASA's statement on p-values: context, process and purpose," *American Statistician* **70**(2), 129–133.