Confusion Over Measures of Evidence (p's) Versus Errors (α's) in Classical Statistical Testing

Raymond HUBBARD and M. J. BAYARRI

Confusion surrounding the reporting and interpretation of results of classical statistical tests is widespread among applied researchers, most of whom erroneously believe that such tests are prescribed by a single coherent theory of statistical inference. This is not the case: Classical statistical testing is an anonymous hybrid of the competing and frequently contradictory approaches formulated by R. A. Fisher on the one hand, and Jerzy Neyman and Egon Pearson on the other. In particular, there is a widespread failure to appreciate the incompatibility of Fisher's evidential p value with the Type I error rate, α, of Neyman–Pearson statistical orthodoxy. The distinction between evidence (p's) and error (α's) is not trivial. Instead, it reflects the fundamental differences between Fisher's ideas on significance testing and inductive inference, and Neyman–Pearson's views on hypothesis testing and inductive behavior. The emphasis of the article is to expose this incompatibility, but we also briefly note a possible reconciliation.

KEY WORDS: Conditional error probabilities; Fisher; Hypothesis test; Neyman–Pearson; p Values; Significance test.

1. INTRODUCTION

Modern textbooks on statistical analysis in the business, social, and biomedical sciences, whether at the undergraduate or graduate levels, typically present the subject matter as if it were gospel: a single, unified, uncontroversial means of statistical inference. Rarely do these texts mention, much less discuss, that classical statistical inference as it is commonly presented is essentially an *anonymous* hybrid consisting of the marriage of the ideas developed by Ronald Fisher on the one hand, and Jerzy Neyman and Egon Pearson on the other (Gigerenzer 1993; Goodman 1993, 1999; Royall 1997). It is a marriage of convenience that neither party would have condoned, for there are important philosophical and methodological differences between them, Lehmann's (1993) attempt at partial reconciliation notwithstanding.

Most applied researchers are unmindful of the historical development of methods of statistical inference, and of the conflation of Fisherian and Neyman–Pearson ideas. Of critical importance, as Goodman (1993) pointed out, is the extensive failure to recognize the *incompatibility* of Fisher's evidential p value with the Type I error rate, α, of Neyman–Pearson statistical orthodoxy. The distinction between *evidence* (p's) and *error* (α's) is no semantic quibble. Instead, it illustrates the fundamental differences between Fisher's ideas on *significance testing* and *inductive inference*, and Neyman–Pearson's views on *hypothesis testing* and *inductive behavior*. Because statistics textbooks tend to anonymously cobble together elements from both schools of thought, however, confusion over the reporting and interpretation of statistical tests is inevitable. Paradoxically, this misunderstanding over measures of evidence versus error is so deeply entrenched that it is not even seen as being a problem by the vast majority of researchers. In particular, the misinterpretation of p values results in an overstatement of the evidence against the null hypothesis. A consequence of this is the number of "statistically significant effects" later found to be negligible, to the embarrassment of the statistical community.

Given the above concerns, this article has two major objectives. First, we outline the marked differences in the conceptual foundations of the Fisherian and Neyman–Pearson statistical testing approaches. Second, we show how the rival ideas from the two schools of thought have been unintentionally mixed together. We illustrate how this mixing has resulted in widespread confusion over the interpretation of p values and α levels. This mass confusion, in turn, has rendered applications of classical statistical testing all but meaningless among applied researchers. In passing, we suggest a possible reconciliation between p's and α's.

2. FISHER'S SIGNIFICANCE TESTING AND INDUCTIVE INFERENCE

Fisher's views on significance testing, presented in his research papers and in various editions of his enormously influential texts, *Statistical Methods for Research Workers* (1925) and *The Design of Experiments* (1935a), took root among applied researchers. Central to his conception of inductive inference is what he called the null hypothesis, H_0. Fisher sought to provide a more "objective" approach to inductive inference. Therefore, he rejected the methods of inverse probability, that is, the probability of a hypothesis (H) given the data (x), or $\Pr(H|x)$, in favor

Raymond Hubbard is Thomas F. Sheehan Distinguished Professor of Marketing, College of Business and Public Administration, Drake University, Des Moines, IA 50311 (E-mail: Raymond.Hubbard@drake.edu). M. J. Bayarri is Professor, Department of Statistics and Operations Research, University of Valencia, Burjassot, Valencia 46100, Spain (E-mail: Susie.bayarri@uv.es). The authors thank Stuart Allen, Scott Armstrong, James Berger, Steven Goodman, Rahul Parsa, and Daniel Vetter for comments on earlier versions of this article. Any remaining errors are our responsibility. This work is supported in part by the Ministry of Science and Technology of Spain under grant SAF2001-2931.

 DOI: 10.1198/0003130031856

of the direct probability, or $\Pr(x|\mathrm{H})$. This was facilitated by his conviction that: "it is possible to argue from consequences to causes, from observations to hypotheses" (Fisher 1966, p. 3). More specifically, Fisher used discrepancies in the data to reject the null hypothesis, that is, the probability of the data given the truth of the null hypothesis, or $\Pr(x|\mathrm{H}_0)$. Thus, a significance test is a procedure for establishing the probability of an outcome, as well as more extreme ones, on a null hypothesis of no effect or relationship.

In Fisher's approach the researcher sets up a null hypothesis that a sample comes from a hypothetical infinite population with a known sampling distribution. The null hypothesis is said to be "disproved," as Fisher called it, or rejected if the sample estimate deviates from the mean of the sampling distribution by more than a specified criterion, the level of significance. According to Fisher (1966, p. 13), "It is usual and convenient for experimenters to take 5 per cent. as a standard level of significance, in the sense that they are prepared to ignore all results which fail to reach this standard. . . ." Consequently, the Fisherian scheme of significance testing centers on the rejection of the null hypothesis at the $p \leq .05$ level. Or as he (Fisher 1966, p. 16) declared: "Every experiment may be said to exist only in order to give the facts a chance of disproving the null hypothesis."

For Fisher (1926, p. 504), then, a phenomenon was considered to be demonstrable when we know how to conduct experiments that will typically yield statistically significant ($p \leq .05$) results: "A scientific fact should be regarded as experimentally established only if a properly designed experiment *rarely fails* to give this level of significance." (Original emphasis.) But it would be wrong, contrary to popular opinion, to conclude that although Fisher (1926, p. 504) endorsed the 5% level, that he was wedded to it: "If one in twenty does not seem high enough odds, we may, if we prefer it, draw the line at one in fifty (the 2 per cent point), or one in a hundred (the 1 per cent point)."

Fisher regarded p values as constituting inductive evidence against the null hypothesis; the smaller the p value, the greater the weight of said evidence (Johnstone 1986, 1987b; Spielman 1974). In terms of his famous disjunction, a p value $\leq .05$ on the null hypothesis indicates that "Either an exceptionally rare chance has occurred or the theory is not true" (Fisher 1959, p. 39). Accordingly, a p value for Fisher represented an "objective" way for researchers to assess the plausibility of the null hypothesis:

> . . . the feeling induced by a test of significance has an objective basis in that the probability statement on which it is based is a fact communicable to and verifiable by other rational minds. The level of significance in such cases fulfils the conditions of a measure of the rational grounds for the disbelief [in the null hypothesis] it engenders (Fisher 1959, p. 43).

In other words, Fisher considered the use of probability values to be more reliable than, say, "eyeballing" results.

Fisher believed that statistics could play an important part in promoting inductive inference, that is, drawing inferences from the particular to the general, from samples to populations. For him, the p value assumes an epistemological role. As he put it, "The conclusions drawn from such [significance] tests constitute the steps by which the research worker gains a better understanding of his experimental material, and of the problems it presents" (Fisher 1959, p. 76). He proclaimed that "The study of inductive reasoning is the study of the embryology of knowledge" (Fisher 1935b, p. 54), and that "Inductive inference is the only process known to us by which essentially new knowledge comes into the world" (Fisher 1966, p. 7). In announcing this, however, he was keenly aware that not everyone shared his inductivist approach, especially

> . . . mathematicians [like Neyman] who have been trained, as most mathematicians are, almost exclusively in the technique of deductive reasoning [and who as a result would] . . . deny at first sight that rigorous inferences from the particular to the general were even possible (Fisher 1935b, p. 39).

This concession aside, Fisher steadfastly argued that inductive reasoning is the primary means of knowledge acquisition, and he saw the p values from significance tests as being evidential.

3. NEYMAN–PEARSON HYPOTHESIS TESTING AND INDUCTIVE BEHAVIOR

Neyman–Pearson (1928a, 1928b, 1933) statistical methodology, originally viewed as an attempt to "improve" on Fisher's approach, gained in popularity after Word War II. It is widely thought of as constituting the basis of classical statistical testing (Royall 1997; Spielman 1974). Their work on hypothesis testing, terminology they employed to contrast with Fisher's "significance testing," differed markedly, however, from the latter's paradigm of inductive inference. (We keep the traditional name "Neyman–Pearson" to denote this school of thought, although Lehmann (1993) mentioned that Pearson apparently did not participate in the confrontations with Fisher.) The Neyman–Pearson approach formulates two competing hypotheses, the null hypothesis (H_0) and the alternative hypothesis (H_A). In a not so oblique reference to Fisher, Neyman commented on the rationale for an alternative hypothesis:

> . . . when selecting a criterion to test a particular hypothesis H, should we consider only the hypothesis H, or something more? It is known that some statisticians are of the opinion that good tests can be devised by taking into consideration only the [null] hypothesis tested. But my opinion is that this is impossible and that, if satisfactory tests are actually devised without explicit consideration of anything beyond the hypothesis tested, it is because the respective authors *subconsciously* take into consideration certain relevant circumstances, namely, the alternative hypothesis that may be true if the hypothesis tested is wrong (Neyman 1952, p. 44; original emphasis).

Specification of an alternative hypothesis critically distinguishes between the Fisherian and Neyman–Pearson methodologies, and this was one of the topics that both camps vehemently disagreed about over the years.

In a sense, Fisher used some kind of casual, generic, unspecified, alternative when computing p values, somehow implicit when identifying the test statistic and "more extreme outcomes" to compute p values, or when talking about the sensitivity of an experiment. But he never explicitly defined nor used specific alternative hypotheses. In the merging of the two schools of thought, it is often taken that Fisher's significance testing implies an alternative hypothesis which is simply the complement of the null, but this is difficult to formalize in general. For example, what is the complement of a $N(0, 1)$ model? Is it the mean differing from 0, the variance differing from 1, the model not being Normal? Formally, Fisher only had the null model in mind and wanted to check if the data were compatible with it.

In Neyman–Pearson theory, therefore, the researcher chooses a (usually point) null hypothesis and tests it against the alternative hypothesis. Their framework introduced the probabilities of committing two kinds of errors based on considerations regarding the decision criterion, sample size, and effect size. These errors were false rejection (Type I error) and false acceptance (Type II error) of the null hypothesis. The former probability is called α, while the latter probability is designated β.

The Neyman–Pearson theory of hypothesis testing introduced the completely new concept of the power of a statistical test. The power of a test, defined as $(1-\beta)$, is the probability of rejecting a false null hypothesis. Because Fisher's statistical testing procedure admits of no alternative hypothesis ($\mathrm{H_A}$), the concepts of Type II error and the power of the test are not relevant. Fisher made this clear: "The notion of an error of the so-called 'second kind,' due to accepting the null hypothesis 'when it is false' . . . has no meaning with respect to simple tests of significance, in which the only available expectations are those which flow from the null hypothesis being true" (Fisher 1966, p. 17). Fisher never saw the need for an alternative hypothesis (but see our comments above).

Fisher (1966, p. 21) nevertheless hints at the notion of the power of a test when he referred to how "sensitive" an experiment might be in detecting departures from the null. As Neyman (1967, p. 1459) later expressed, "The consideration of power is occasionally implicit in Fisher's writings, but I would have liked to see it treated explicitly." Essentially, however, Fisher's "sensitivity" and Neyman–Pearson's "power" refer to the same concept. But here ends the, purely conceptual, agreement: power has no methodological role in Fisher's approach whereas it has a crucial one in Neyman–Pearson's.

Although Fisher's view of inductive inference focused on the rejection of the null hypothesis, Neyman and Pearson dismissed the entire idea of inductive reasoning out of hand. Instead, their concept of *inductive behavior* sought to establish rules for making *decisions* between two hypotheses, irrespective of the researcher's belief in either one. Neyman explained:

> Thus, to accept a hypothesis H means only to decide to take action A rather than action B. This does not mean that we necessarily believe that the hypothesis H is true . . . [while rejecting H] . . . means only that the rule prescribes action B and does not imply that we believe that H is false (Neyman 1950, pp. 259-260).

Neyman–Pearson theory, then, replaces the idea of inductive reasoning with that of inductive behavior. In defending his preference for inductive behavior over inductive inference, Neyman wrote:

> . . . the term "inductive reasoning" remains obscure and it is uncertain whether or not the term can be conveniently used to denote any clearly defined concept. On the other hand . . . there seems to be room for the term "inductive behavior." This may be used to denote the adjustment of our behavior to limited amounts of information. The adjustment is partly conscious and partly subconscious. The conscious part is based on certain rules (if I see this happening, then I do that) which we call rules of inductive behavior. In establishing these rules, the theory of probability and statistics both play an important role, and there is a considerable amount of reasoning involved. *As usual, however, the reasoning is all deductive* (Neyman 1950, p. 1; our emphasis).

The Neyman–Pearson approach is deductive in nature and argues from the general to the particular. They formulated a "rule of behavior" for choosing between two alternative courses of action, accepting or rejecting the null hypothesis, such that ". . . in the long run of experience, we shall not be too often wrong" (Neyman and Pearson 1933, p. 291).

The decision to accept or reject the hypothesis in their framework depends on the costs associated with committing a Type I or Type II error. These costs have nothing to do with statistical theory, but are based instead on context-dependent pragmatic considerations where informed personal judgment plays a vital role. Thus, the researcher would design an experiment to control the probabilities of the α and β error rates. The "best" test is one that minimizes β subject to a bound on α (Lehmann 1993). And in an act that Fisher, as we shall see, could never countenance, Neyman referred to α as the significance level of a test:

> The error that a practicing statistician would consider the more important to avoid (which is a subjective judgment) is called the error of the first kind. The first demand of the mathematical theory is to deduce such test criteria as would ensure that the probability of committing an error of the first kind would equal (or approximately equal, or not exceed) a preassigned number α, such as $\alpha = 0.05$ or 0.01, etc. *This number is called the level of significance* (Neyman 1976, p. 161; our emphasis).

Since α is specified or fixed *prior* to the collection of the data, the Neyman–Pearson procedure is sometimes referred to as the fixed α/fixed level approach (Lehmann 1993). This is in sharp contrast to the data-based p value, which is a *random variable* whose distribution is uniform over the interval $[0, 1]$ under the null hypothesis. Thus, the α and β error rates define a "critical" or "rejection" region for the test statistic, say z or $t > 1.96$. If the test statistic falls in the critical region $\mathrm{H_0}$ is rejected in favor of $\mathrm{H_A}$, otherwise $\mathrm{H_0}$ is retained.

Moreover, while Fisher claimed that his significance tests were applicable to single experiments (Johnstone 1987a; Kyburg 1974; Seidenfeld 1979), Neyman–Pearson hypothesis tests do not allow an inference to be made about the outcome of any *specific* hypothesis that the researcher happens to be investigating. The latter were quite explicit about this: "We are inclined to think that as far as a particular hypothesis is concerned, no test based upon the theory of probability can by itself provide any valuable evidence of the truth or falsehood of that hypothesis" (Neyman and Pearson 1933, pp. 290–291).

Neyman–Pearson theory is *nonevidential*. Fisher recognized this when agreeing that their "acceptance procedures" approach could play a part in repeated sampling, quality control decisions. This admission notwithstanding, Fisher was adamant that Neyman–Pearson's cost-benefit, decision making, orientation to statistics was an inappropriate model for the conduct of science, reminding us that there exists a:

> . . . deep-seated difference in point of view which arises when Tests of Significance are reinterpreted on the analogy of Acceptance Decisions. It is indeed not only numerically erroneous conclusions, serious as these are, that are to be feared from an uncritical acceptance of this analogy. An important difference is that decisions are final, while the state of opinion derived from a test of significance is provisional, and capable, not only of confirmation, but of revision (Fisher 1959, p. 100).

Clearly, Fisher and Neyman were at odds over the role played by statistical testing in scientific investigations, and over the nature of the scientific enterprise itself. In fact, the dogged insistence on the correctness of their respective conceptions of statistical testing and the scientific method resulted in ongoing

acrimonious exchanges, at both the professional and personal levels, between them.

4. CONFUSION OVER THE INTERPRETATION OF p's AND α's

Most users of statistical tests in the applied sciences are unaware of the above distinctions between the Fisherian and Neyman–Pearson camps (Gigerenzer 1993; Goodman 1993; Royall 1997). This is because many statistics textbooks combine (sometimes incongruous) ideas from both schools of thought, usually without acknowledging, or worse yet, recognizing, this. Johnstone (1986) remarked that statistical testing usually follows Neyman–Pearson formally, but Fisher philosophically. For instance, Fisher's idea of disproving the null hypothesis is taught in tandem with the Neyman–Pearson concepts of alternative hypotheses, Type II errors, and the power of a statistical test.

As a prime example of the bewilderment arising from the mixing of Fisher's views on inductive inference with the Neyman–Pearson principle of inductive behavior, consider the widely unappreciated fact that the former's p value is *incompatible* with the Neyman–Pearson hypothesis test in which it has become embedded (Goodman 1993). Despite this incompatibility, the upshot of this merger is that the p value is now inextricably entangled with the Type I error rate, α. As a result, most empirical work in the applied sciences is conducted along the following approximate lines: The researcher states the null (H_0) and alternative (H_A) hypotheses, the Type I error rate/significance level, α, and supposedly—but very rarely—calculates the statistical power of the test (e.g., t). These procedural steps are entirely consistent with Neyman–Pearson convention. Next, the test statistic is computed for the sample data, and in an attempt to have one's cake and eat it too, an associated p value is determined. The p value is then mistakenly interpreted as a frequency-based Type I error rate, and simultaneously as an incorrect (i.e., $p < \alpha$) measure of evidence against H_0.

Confusion over the meaning and interpretation of p's and α's is close to total. It is almost guaranteed by the fact that, Fisher's efforts to distinguish between them to the contrary, this same confusion exists among some statisticians. These themes are addressed below.

4.1 Fisher—The Significance Level (p) of a Test is Not a Type I Error Rate (α)

Fisher was insistent that the significance level of a test had no ongoing sampling interpretation. With respect to the .05 level, for example, he emphasized that this does not indicate that the researcher "allows himself to be deceived once in every twenty experiments. The test of significance only tells him what to ignore, namely all experiments in which significant results are not obtained" (Fisher 1929, p. 191). For Fisher, the significance level provided a measure of evidence for the "objective" disbelief in the null hypothesis; it had no long-run frequentist characteristics.

Indeed, interpreting the significance level of a test in terms of a Neyman–Pearson Type I error rate, α, rather than via a p value, infuriated Fisher who complained:

> In recent times one often-repeated exposition of the tests of significance, by J. Neyman, a writer not closely associated with the development of these tests, seems liable to lead mathematical readers astray, through laying down axiomatically, what is not agreed or generally true, that the level of significance must be equal to the frequency with which the hypothesis is rejected in repeated sampling of any fixed population allowed by hypothesis. This intrusive axiom, which is foreign to the reasoning on which the tests of significance were in fact based seems to be a real bar to progress. . . " (Fisher 1945, p. 130).

And he periodically reinforced these sentiments:

> The attempts that have been made to explain the cogency of tests of significance in scientific research, by reference to supposed frequencies of possible statements, based on them, being right or wrong, thus seem to miss the essential nature of such tests (Fisher 1959, p. 41).

Here, Fisher is categorically denying the equivalence of p values and Neyman–Pearson α levels, that is, long-run frequencies of rejecting H_0 when it is true. Fisher captured a major distinction between his and Neyman–Pearson's notions of statistical tests when he pronounced:

> This [Neyman–Pearson] doctrine, which has been very dogmatically asserted, makes a truly marvellous mystery of the tests of significance. On the earlier view, held by all those to whom we owe the first examples of these tests, such a test was logically elementary. It presented the logical disjunction: Either the hypothesis is not true, or an exceptionally rare outcome has occurred (Fisher 1960, p. 8).

Seidenfeld (1979) and Rao (1992) agreed that the correct reading of a Fisherian significance test is through this disjunction, as opposed to some long-run frequency interpretation. In direct opposition, however, "the essential point [of Neyman–Pearson theory] is that the solution reached is always unambiguously interpretable in terms of long range relative frequencies" (Neyman 1955, p. 19). Hence the impasse.

4.2 Confusion Over p's and α's Among Some Statisticians

4.2.1 Misinterpreting the p Value as a Type I Error Rate

Despite the admonitions about the p value not being an error rate, Casella and Berger (1987, p. 133) voiced their concern that "there are a great many statistically naïve users who are interpreting p values as probabilities of Type I error. . . ." Unfortunately, such misinterpretations are confined not only to the naïve users of statistical tests. For example, Gibbons and Pratt (1975, p. 21), in an article titled "P-Values: Interpretation and Methodology," erroneously stated: "Reporting a P-value, whether exact or within an interval, in effect permits each individual to choose his own level of significance as the maximum tolerable probability of a Type I error." Again, Hung, O'Neill, Bauer, and Köhne (1997, p. 12) noted that the p value is a measure of evidence against the null hypothesis, but then go on to confuse p values with Type I error rates: "The α level is a preexperiment Type I error rate used to control the probability that the observed P value in the experiment of making an error rejection of H_0 when in fact H_0 is true is α or less."

Or consider Berger and Sellke's response to Hinkley's (1987) comments on their article:

> Hinkley defends the P value as an "unambiguously objective error rate." The use of the term "error rate" suggests that the [Neyman–Pearson] frequentist justifications . . . for confidence intervals and fixed α-level hypothesis tests carry over to P values. *This is not true.* Hinkley's interpretation of the P value as an error rate is presumably as follows: the P value is the Type I error rate that would result if this observed P value were used as the critical significance level in a long sequence

of hypothesis tests . . . This hypothetical error rate does not conform to the usual classical notion of "repeated-use" error rate, since the P value is determined only once in this sequence of tests. The frequentist justifications of significance tests and confidence intervals are in terms of how these procedures perform when used repeatedly.

Can P values be justified on the basis of how they perform in repeated use? We doubt it. For one thing, how would one measure the performance of P values? (Berger and Sellke 1987, p. 136; our emphasis).

Berger and Delampady (1987, p. 329) correctly insisted that the interpretation of the p value as an error rate is strictly prohibited: "P-values are *not* a repetitive error rate. . . A Neyman–Pearson error probability, α, has the actual frequentist interpretation that a long series of α level tests will reject no more than $100\alpha\%$ of true H_0, but the data-dependent-P-values have no such interpretation." (Original emphasis). In sum, although p's and α's have very different meanings, Bayarri and Berger (2000) nevertheless contended that among statisticians there is a near ubiquitous misinterpretation of p values as frequentist error probabilities.

4.2.2 Using the $p < \alpha$ Criterion as a Measure of Evidence Against H_0

At the same time that the p value is being incorrectly reported as a Neyman–Pearson Type I error rate, it will also be incorrectly interpreted in a quasi-Fisherian sense as evidence against H_0. This is accomplished in an unusual manner by examining the inequality between a measure of evidence and a long-term error rate, or $p < \alpha$. If $p < \alpha$, a statistically significant finding is reported, and the null hypothesis is disproved, or at least discredited. Statisticians also commit this mistake. In an article published in the *Encyclopedia of Statistical Sciences* intended to clarify the meaning of p values, for example, Gibbons (1986, p. 367) falsely concluded that: "Hence the relationship between P values and the classical [Neyman–Pearson] method is that if $P \leq \alpha$, we should reject H_0, and if $P > \alpha$, we should accept H_0." But Gibbons is by no means alone among statisticians regarding this confusion over the evidential content (and mixing) of p's and α's. For instance, Donahue (1999, p. 305) stated: "Obviously, with respect to rejecting the null hypothesis and small values of P, we proceed *as tradition dictates* by rejecting H if $P < \alpha$." (Our emphasis.)

But the p value plays no role in Neyman–Pearson theory. Instead, their framework focuses on decision rules with a priori stated error rates, α and β, which are limiting frequencies based on long-run repeated sampling. If a result falls into the critical region H_0 is rejected and H_A is accepted, otherwise H_0 is accepted and H_A is rejected.

Of course, for a fixed, prespecified α, the Neyman–Pearson decision rule is fully determined by the critical region of the sample, which in turn can be characterized in terms of many different statistics (in particular, of any one-to-one transformation of the original test statistic). Therefore, it could be defined equivalently in terms of the p value, and stated as saying that the null hypothesis should be rejected if the observed $p < \alpha$, and accepted otherwise. But in this manner, only the Neyman–Pearson interpretation is valid, and no matter how small the p value is, the appropriate report is that the procedure guarantees a $100\alpha\%$ false rejections of the null on repeated use. Otherwise stated, only the fact that $p < \alpha$ is of any interest, not the specific value of p itself.

A related issue is whether one can carry out both testing procedures in parallel. We have seen from a philosophical perspective that this is extremely problematic. We do not recommend it from a pragmatic point of view either, because the danger in interpreting the p value as a data-dependent adjustable Type I error is too great, no matter the warnings to the contrary. Indeed, if a researcher is interested in the "measure of evidence" provided by the p value, we see no use in also reporting the error probabilities, since they do not refer to any property that the p value has. (In addition, the appropriate interpretation of p values as a measure of evidence against the null is not clear. We delay this discussion until Section 5.) Likewise, if the researcher is concerned with error probabilities the specific p value is irrelevant.

Despite the above statements, Goodman (1993, 1999) and Royall (1997) noted that because of its superficial resemblance to the Neyman–Pearson Type I error rate, α, the p value has been absorbed into the former's hypothesis testing method. In doing so, the p value has been interpreted as both a measure of evidence and an "observed" error rate. This has led to widespread confusion over the meaning of p values and α levels. Unfortunately, as Goodman pointed out:

. . . because p-values and the critical regions of hypothesis tests are both tail area probabilities, they are easy to confuse. This confusion blurs the division between concepts of evidence and error for the statistician, and obscures it completely for nearly everyone else (Goodman 1992, p. 879).

4.3 *p*'s, α's and the .05 Level

It is ironic that the confusion surrounding the distinction between p's and α's was unwittingly exacerbated by Neyman and Pearson themselves. This occurred when, despite their insistence on flexibility over the balancing of α and β errors, they adopted as a matter of expediency Fisher's 5% and 1% significance levels to help define their Type I error rates (Pearson 1962). Consequently, it is small wonder that many researchers confuse Fisher's evidential p values with Neyman–Pearson behavioral error rates when both concepts are commonly employed at the 5% and 1% levels.

Many researchers will no doubt be surprised by the statisticians' confusion over the correct meaning and interpretation of p values and α levels. After all, one might anticipate that the properties of these commonly used statistical measures would be completely understood. But this is not the case. To underscore this point, in commenting on various issues surrounding the interpretation of p values, Berger and Sellke (1987, p. 135) unequivocally spelled out that: "These are not dead issues, in the sense of being well known and thoroughly aired long ago; although the issues are not new, *we have found the vast majority of statisticians to be largely unaware of them.*" (Our emphasis.) Schervish's (1996) article almost a decade later, tellingly entitled "P Values: What They Are and What They Are Not," suggests that confusion remains in this regard within the statistics community.

The near-universal confusion among researchers over the meaning of p values and α levels becomes easier to appreciate when it is formally acknowledged that both expressions are

used to indicate the "significance level" of a test. But note their completely different interpretations. The level of significance shown by a p value in a Fisherian significance test refers to the probability of observing data this extreme (or more so) under a null hypothesis. This data-dependent p value plays an epistemic role by providing a measure of inductive evidence against H_0 in single experiments. This is very different from the significance level denoted by α in a Neyman–Pearson hypothesis test. With Neyman–Pearson, the focus is on minimizing Type II, or β, errors (i.e., false acceptance of a null hypothesis) subject to a bound on Type I, or α, errors (i.e., false rejections of a null hypothesis). Moreover, this error minimization applies only to long-run repeated sampling situations, not to individual experiments, and is a prescription for behaviors, not a means of collecting evidence. When seen from this vantage, the two concepts of statistical significance could scarcely be further apart in meaning.

The problem is that these distinctions between p's and α's are seldom made explicit in the literature. Instead, they tend to be used interchangeably, especially in statistics textbooks aimed at practitioners. Thus, we have a nameless amalgamation of the Fisherian and Neyman–Pearson paradigms, with the p value serving as the conduit, that has created the potent illusion of a uniform statistical methodology somehow capable of generating evidence from single experiments, while at the same time minimizing the occurrence of errors in both the short and long hauls (Goodman 1993). It is now ensconced in college curricula, textbooks, and journals.

5. WHERE DO WE GO FROM HERE?

If researchers are confused over the meaning of p values and Type I error probabilities, and the Fisher and Neyman–Pearson theories seemingly cannot be combined, what should we do? The answer is not obvious since both schools have important merits and drawbacks. In the following account we no longer address the philosophical issues concerning the distinctions between p's and α's that have been the main themes of previous sections, in the hope that these are clear enough. Instead, we concentrate on the implications for statistical practice: Is it better to report p values or error probabilities from a test of hypothesis? We follow this with a discussion of how we can, in fact, reconcile the Fisherian and Neyman–Pearsonian statistical testing frameworks.

5.1 Some Practical Problems with p's and α's

Neyman–Pearson theory has the advantage of its clear interpretation: Of all the tests being carried out around the world at the .05 level, at most 5% of them result in a false rejection of the null. [The frequentist argument does *not* require repetition of the exact same experiment. See, for instance, Berger (1985, p. 23) and references there.] Its main drawback is that the performance of the procedure is always the prespecified level. Reporting the same "error," .05 say, no matter how incompatible the data seem to be with the null hypothesis is clearly worrisome in applied situations, and hence the appeal of the data-dependent p values in research papers. On the other hand, for quality control problems, a strict Neyman–Pearson analysis is appropriate.

The chief methodological advantage of the p value is that it may be taken as a quantitative measure of the "strength of evidence" against the null. However, while p values are very good as *relative* measures of evidence, they are extremely difficult to interpret as *absolute* measures. What exactly "evidence" of around, say, .05 (as measured by a p value) means is not clear. Moreover, the various misinterpretations of p values all result, as we shall see, in an exaggeration of the actual evidence against the null. This is very disconcerting on practical grounds. Indeed, many "effects" found in statistical analyses have later been shown to be mere flukes. For examples of these, visit the Web pages mentioned in www.stat.duke.edu/~berger under "p values." Such results undermine the credibility of the profession.

A common mistake by users of statistical tests is to misinterpret the p value as the probability of the null hypothesis being true. This is not only wrong, but p values and posterior probabilities of the null can differ by several orders of magnitude, the posterior probability always being larger (see Berger 1985; Berger and Delampady 1987; Berger and Sellke 1987). Most books, even at the elementary level, are aware of this misinterpretation and warn about it. It is rare, however, for these books to emphasize the practical consequences of falsely equating p values with posterior probabilities, namely, the conspicuous exaggeration of evidence against the null.

As we have shown throughout this article, researchers routinely confuse p values with error probabilities. This is not only wrong philosophically, but also has far-reaching practical implications. To see this we urge those teaching statistics to simulate the frequentist performance of p values in order to demonstrate the serious conflict between the student's intuition and reality. This can be done trivially on the Web, even at the undergraduate level, with an applet available at www.stat.duke.edu/~berger. The applet simulates repeated normal testing, retains the tests providing p values in a given range, and counts the proportion of those for which the null is true. The exercise is revealing. For example, if in a long series of tests on, say, no effect of new drugs (against AIDS, baldness, obesity, common cold, cavities, etc.) we assume that about half the drugs are effective (quite a generous assumption), then of all the tests resulting in a p value around .05 it is fairly typical to find that about 50% of them come, in fact, from the null (no effect) and 50% from the alternative. These percentages depend, of course, on the way the alternatives behave, but an absolute lower bound, for any way the alternatives could arise in the situation above, is about 22%. The upshot for applied work is clear. Most notably, about half (or at the very least over 1/5) of the times we see a p value of .05, it is actually coming from the null. That is, a p value of .05 provides, at most, very mild evidence against the null. When practitioners (and students) are not aware of this, they very likely interpret a .05 p value as much greater evidence against the null.

Finally, sophisticated statisticians (but very few students) might offer the argument that p values are just a measure of evidence in the sense that "either the null is false, or a rare event has occurred." The main flaw in this viewpoint is that the "rare event," whose probability (under the null) the p value computes, is *not* based on observed data, as the previous argument implies. Instead, the probability of the set of all data more extreme than the actual data is computed. It is obvious that in this set there can be data far more incompatible with the null than the data at hand,

and hence this set provides much more "evidence" against the null than does the actual data. This conditional fallacy, therefore, also results in an exaggeration of the evidence against the null provided by the observed data. Our informal argument is made in a rigorous way in Berger and Sellke (1987) and Berger and Delampady (1987).

5.2 Reconciling Fisher's and Neyman–Pearson's Methods of Statistical Testing

So, what should we do? One possible course of action is to use Bayesian measures of evidence (Bayes factors and posterior probabilities for hypothesis). Space constraints preclude debating this possibility here. Suffice it to say that there is a long-standing misconception that Bayesian methods are necessarily "subjective." In fact, objective Bayesian analyses can be carried out without incorporating any external information (see Berger 2000), and in recent years the objective Bayesian methodology for hypothesis testing and model selection has experienced rapid development (Berger and Pericchi 2001).

The interesting question, however, is not whether another methodology can be adopted, but rather can the ideas from the Neyman–Pearson and Fisher schools somehow be reconciled, thereby retaining the best of both worlds? This is what Lehmann (1993, p. 1248) had in mind, but he recognized that "A fundamental gap in the theory is the lack of clear principles for selecting the appropriate framework." There is, however, such a unifying theory which provides the "appropriate framework" Lehmann (1993) sought. This was clearly presented by Berger (2002). The intuitive notion behind it is that one should report *conditional* error probabilities. That is, reports that retain the unambiguous frequency interpretation, but that are allowed to vary with the observed data. The specific proposal is to condition on data that have the same "strength of evidence" as measured by p values. We see this as the ultimate reconciliation between the two opposing camps. Moreover, it has an added bonus: the conditional error probabilities can be interpreted as posterior probabilities of the hypotheses, thus guaranteeing easy computation as well as marked simplifications in sequential scenarios.

A very easy, approximate, calibration of p values was given by Sellke, Bayarri, and Berger (2001). It consists of computing, for an observed p value, the quantity $(1 + [-e\,p\log(p)]^{-1})^{-1}$ and interpreting this as a lower bound on the conditional Type I error probability. For example, a p value of .05 results in a *conditional* α of at least .289. (The calibration $-e\,p\log(p)$ can be interpreted as a lower bound on the Bayes factor.) That is, even though a p value of .05 might seem to give the impression that the evidence against H_0 is about 20 to 1, and hence quite strong, the conditional α of at least .289 (and corresponding Bayes factor of at least 0.4) tells us otherwise. In particular, as shown by Sellke, Bayarri, and Berger (2001), it can be interpreted (under some conditions) as saying that among all the experiments resulting in p values around .05, *at least* 28.9% come from those in which the null hypothesis is true. Also, the (Bayesian) odds against H_0 are *at most* 2.5 to 1. Both statements reveal that the practical evidence against H_0 provided by a p value of .05 is, at best, rather weak. This, of course, is in flagrant contradiction with usual interpretations. [For more discussion of this topic, and the corresponding implications, see Sellke, Bayarri, and Berger (2001) and Berger (2002).] The formulas introduced above are extremely simple and provide the correct order of magnitude for interpreting a p value as an error probability, as well as the evidence against the null given by a p value.

6. CONCLUSIONS

It is disturbing that the ubiquitous p value cannot be correctly interpreted by the majority of researchers. As a result, the p value is viewed simultaneously in Neyman–Pearson terms as a deductive assessment of error in long-run repeated sampling situations, and in a Fisherian sense as a measure of inductive evidence in a single study. In fact, a p value from a significance test has no place in the Neyman–Pearson hypothesis testing framework. Contrary to popular misconception, p's and α's are not the same thing; they measure different concepts.

Nevertheless, both concepts—evidence and error—can be important, and we briefly indicated a possible reconciliation by calibrating p values as conditional error probabilities. In the broader picture, we believe that it would be especially informative if those teaching statistics courses in the applied disciplines addressed the historical development of statistical testing in their classes and their textbooks. It is hoped that this article will stimulate discussions along these lines.

[Received October 2001. Revised May 2003.]

REFERENCES

Bayarri, M. J., and Berger, J. O. (2000), "P Values for Composite Null Models," *Journal of the American Statistical Association*, 95, 1127–1142.

Berger, J. O. (1985), *Statistical Decision Theory and Bayesian Analysis* (2nd ed.), New York: Springer-Verlag.

——— (2000), "Bayesian Analysis: A Look at Today and Thoughts of Tomorrow," *Journal of the American Statistical Association*, 95, 1269–1276.

——— (2002), "Could Fisher, Jeffreys, and Neyman Have Agreed on Testing?" ISDS (Duke University) discussion paper 02-01.

Berger, J. O., and Delampady, M. (1987), "Testing Precise Hypotheses" (with comments), *Statistical Science*, 2, 317–352.

Berger, J. O., and Pericchi, L. (2001), "Objective Bayesian Methods for Model Selection: Introduction and Comparison" (with comments), in *Model Selection*, ed. P. Lahiri, Institute of Mathematical Statistics Lecture Notes—Monograph Series, Volume 38, 135–207.

Berger, J. O., and Sellke, T. (1987), "Testing a Point Null Hypothesis: The Irreconcilability of P Values and Evidence" (with comments), *Journal of the American Statistical Association*, 82, 112–139.

Casella, G., and Berger, R. L. (1987), "Reconciling Bayesian and Frequentist Evidence in the One-Sided Testing Problem" (with comments), *Journal of the American Statistical Association*, 82, 106–139.

Donahue, R. M. J. (1999), "A Note on Information Seldom Reported Via the P Value," *The American Statistician*, 53, 303–306.

Fisher, R. A. (1925), *Statistical Methods for Research Workers*, Edinburgh: Oliver and Boyd.

——— (1926), "The Arrangement of Field Experiments," *Journal of the Ministry of Agriculture for Great Britain*, 33, 503–513.

——— (1929), "The Statistical Method in Psychical Research," in *Proceedings of the Society for Psychical Research*, London, 39, pp. 189–192.

——— (1935a), *The Design of Experiments*, Edinburgh: Oliver and Boyd.

——— (1935b), "The Logic of Inductive Inference," *Journal of the Royal Statistical Society*, 98, 39–54.

——— (1945), "The Logical Inversion of the Notion of the Random Variable," *Sankhyā*, 7, 129–132.

——— (1959), *Statistical Methods and Scientific Inference*, (2nd ed., revised), Edinburgh: Oliver and Boyd.

——— (1960), "Scientific Thought and the Refinement of Human Reasoning," *Journal of the Operations Research Society of Japan*, 3, 1–10.

——— (1966), *The Design of Experiments* (8th ed.), Edinburgh: Oliver and Boyd.

Gibbons, J. D. (1986), "P-Values," in *Encyclopedia of Statistical Sciences*, eds. S. Kotz and N. L. Johnson, New York: Wiley, pp. 366–368.

Gibbons, J. D., and Pratt, J. W. (1975), "P-values: Interpretation and Methodology," *The American Statistician*, 29, 20–25.

Gigerenzer, G. (1993), "The Superego, the Ego, and the Id in Statistical Reasoning," in *A Handbook for Data Analysis in the Behavioral Sciences—Methodological Issues*, eds. G. Keren and C. A. Lewis, Hillsdale, NJ: Erlbaum, pp. 311–339.

Goodman, S. N. (1992), "A Comment on Replication, P-Values and Evidence," *Statistics in Medicine*, 11, 875–879.

——— (1993), "p Values, Hypothesis Tests, and Likelihood: Implications for Epidemiology of a Neglected Historical Debate," *American Journal of Epidemiology*, 137, 485–496.

——— (1999), "Toward Evidence-Based Medical Statistics. 1: The P Value Fallacy," *Annals of Internal Medicine*, 130, 995–1004.

Hinkley, D. V. (1987), "Comment," *Journal of the American Statistical Association*, 82, 128–129.

Hung, H. M. J., O'Neill, R. T., Bauer, P., and Köhne, K. (1997), "The Behavior of the P-Value When the Alternative Hypothesis is True," *Biometrics*, 53, 11–22.

Johnstone, D. J. (1986), "Tests of Significance in Theory and Practice" (with comments), *The Statistician*, 35, 491–504.

——— (1987a), "On the Interpretation of Hypothesis Tests Following Neyman and Pearson," in *Probability and Bayesian Statistics*, ed. R. Viertl, New York: Plenum Press, pp. 267–277.

——— (1987b), "Tests of Significance Following R.A. Fisher," *British Journal for the Philosophy of Science*, 38, 481–499.

Kyburg, H. E. (1974), *The Logical Foundations of Statistical Inference*, Dordrecht: Reidel.

Lehmann, E. L. (1993), "The Fisher, Neyman–Pearson Theories of Testing Hypotheses: One Theory or Two?" *Journal of the American Statistical Association*, 88, 1242–1249.

Neyman, J. (1950), *First Course in Probability and Statistics*, New York: Holt.

——— (1952), *Lectures and Conferences on Mathematical Statistics and Probability* (2nd ed., revised and enlarged), Washington, DC: Graduate School, U.S. Department of Agriculture.

——— (1955), "The Problem of Inductive Inference," *Communications on Pure and Applied Mathematics*, 8, 13–45.

——— (1967), "R.A. Fisher (1890–1962), An Appreciation," *Science*, 156, 1456–1460.

——— (1976), "The Emergence of Mathematical Statistics: A Historical Sketch with Particular Reference to the United States," in *On the History of Statistics and Probability*, ed. D.B. Owen, New York: Marcel Dekker, pp. 149–193.

——— (1977), "Frequentist Probability and Frequentist Statistics," *Synthese*, 36, 97–131.

Neyman, J., and Pearson, E. S. (1928a), "On the Use and Interpretation of Certain Test Criteria for Purposes of Statistical Inference. Part I," *Biometrika*, 20A, 175–240.

——— (1928b), "On the Use and Interpretation of Certain Test Criteria for Purposes of Statistical Inference. Part II," *Biometrika*, 20A, 263–294.

——— (1933), "On the Problem of the Most Efficient Tests of Statistical Hypotheses," *Philosophical Transactions of the Royal Society of London*, Ser. A, 231, 289–337.

Pearson, E. S. (1962), "Some Thoughts on Statistical Inference," *Annals of Mathematical Statistics*, 33, 394–403.

Rao, C. R. (1992), "R.A. Fisher: The Founder of Modern Statistics," *Statistical Science*, 7, 34–48.

Royall, R. M. (1997), *Statistical Evidence: A Likelihood Paradigm*, New York: Chapman and Hall.

Schervish, M. J. (1996), "P Values: What They Are and What They Are Not," *The American Statistician*, 50, 203–206.

Seidenfeld, T. (1979), *Philosophical Problems of Statistical Inference: Learning from R. A. Fisher*, Dordrecht: Reidel.

Sellke, T., Bayarri, M. J., and Berger, J. O. (2001), "Calibration of p Values for Testing Precise Null Hypotheses," *The American Statistician*, 55, 62–71.

Spielman, S. (1974), "The Logic of Tests of Significance," *Philosophy of Science*, 41, 211–226.

Discussion

Kenneth N. Berk

One could cynically describe Sections 1 to 4 of this article as an attempt to tear down traditional statistical testing in an effort to sell the books and articles by Berger and his coauthors listed in Section 5. It seems likely that statistical testing will remain unchanged, although these references offer interesting and creative alternative interpretations of p values.

The theme of Sections 2 and 3 seems to be that Fisher did not like what Neyman and Pearson did, and Neyman did not like what Fisher did, and therefore we should not do anything that combines the two approaches. There is no escaping the premise here, but the reasoning escapes me. I can see the image of two gods standing on twin mountains and hurling thunderbolts at each other, but I still do not see why we cannot talk about p values and Type I errors in the same sentence. Section 4 begins by saying how terrible it is that "statistics textbooks combine (sometimes incongruous) ideas from both schools of thought" without mentioning the controversy.

Many books indeed do introduce the p value and then discuss how it can be used to decide whether to accept or reject the null hypothesis at the .05 level. And then they are supposed to say how the originators of these ideas would have hated to see them combined? This is asking a bit much in a situation where the student is trying to learn new terms and difficult concepts, so we cannot expect to see a discussion of this historical controversy in an introductory text. I doubt that my coauthor and I will include it in our new introductory mathematical statistics text.

Section 4.2 points an accusing finger at some who supposedly got it wrong. Gibbons and Pratt are among the chosen ones, because they say that reporting a p value allows each individual to choose "a level of significance as the maximum tolerable probability of a Type I error." If they are guilty, they have a lot of company. Textbooks typically describe a similar procedure, and that is what a lot of people are doing in statistical practice. The computer output comes with a p value, and this allows the user to make a judgment about whether it is less than .05 or .01 or whatever it takes to convince the user. Often the p value is communicated in the published paper, rather than a statement about rejecting the null hypothesis at the .05 level. This is in accord with what Moore and McCabe (1997, p. 476) said: "Different persons may feel that different levels of significance are appropriate. It is better to report the p value, which allows each of us to decide individually if the evidence is sufficiently strong."

It is especially difficult to see the objections to Gibbons

Kenneth N. Berk is Professor Emeritus, Department of Mathematics, Box 4520, Illinois State University, Normal, IL 61790 (E-mail: kberk@ilstu.edu).

and Pratt (1975) (and presumably Moore and McCabe and pretty much the whole statistical world) when you read further in Section 4.2 and see that, "for a prespecified α, the Neyman–Pearson decision rule . . . could be defined in terms of the p value." There must be a linguistic subtlety here that I and many others are missing. Maybe the issue is in the "prespecified" part, which perhaps is not made sufficiently explicit by some. Indeed, Miller and Miller (1999, p. 415) warned against the hazards of allowing the choice of α afterward:

> . . . consider the temptation one might be exposed to when choosing α after having seen the P value with which it is compared. Suppose, for instance, that an experiment yields a P value of 0.036. If we are anxious to reject the null hypothesis and prove our point, it would be tempting to choose $\alpha = 0.05$; if we are anxious to accept the null hypothesis and prove our point it would be tempting to choose $\alpha = 0.01$.

Similarly, Anderson and Finn (1996, p. 421) gave this caution:

> One caution must be observed in using a P value to decide if H_0 is accepted or rejected. It is always tempting to look at a P value, especially when it is printed by a computer program, and use it after the fact to decide what the α level should be. For example, suppose we did not set α in advance of a study and obtained the P value of .0359. If we have a large personal investment in the outcome of the study, we might be tempted to set α at .05 in order to reject H_0, whereas we might have chosen .01 if α was set in advance.

From the point of view of this article, the worst thing in the world is the confusion of the Type I error probability α and the p value. One can only wonder about the reaction to this passage from Boniface (1995, p. 21):

> The *level of significance* is the probability that a difference in means has been erroneously declared to be significant. Typical values for significance levels are 0.05 and 0.01 (corresponding to 5% and 1% chance of error). Another name for significance level is *p-value*.

As is pointed out in Section 5, students frequently misinterpret the p value as the probability that the null hypothesis is true. You might think therefore that texts would hammer away at this fallacy with the hope of defeating it. However, I found only one text that makes any effort in this regard. Rice (1995) has true/false questions like this: "The p-value of a test is the probability that the null hypothesis is correct."

Paradoxically, Rice's text is used mainly for very well-prepared students who presumably have less need for drill of this kind. A case could be made that all introductory texts, especially those at lower levels, should have such exercises.

It should be acknowledged that some statisticians might answer "none of the above" to this article. That is, there are those who prefer not to do hypothesis testing in any form, and instead prefer confidence intervals. This includes statisticians in psychology (Harlow, Mulaik, and Steiger 1997), medicine (Rothman 1978), and engineering (Deming 1975; Hahn 1995). Deming pointed out what has occurred to many of us in dealing with large datasets, that with enough data any hypothesis is rejected.

[Received October 2001. Revised May 2003.]

REFERENCES

Anderson, T. W., and Finn, J.D. (1996), *The New Statistical Analysis of Data*, New York: Springer-Verlag.

Boniface, D. R. (1995), *Experiment Design and Statistical Methods for Behavioural and Social Research*, Boca Raton, FL: CRC Press.

Deming, W. E. (1975), "On Probability as a Basis for Action," *The American Statistician*, 29, 146–152.

Gibbons, J. D., and Pratt, J. W. (1975), "P-values: Interpretation and Methodology," *The American Statistician*, 29, 20–25.

Hahn, G. J. (1995), "Deming's Impact on Industrial Statistics: Some Reflections," *The American Statistician*, 49, 336–341.

Harlow, L. L., Mulaik, S. A., and Steiger, J. H. (1997), *What If There Were No Significance Tests?*, Mahwah, NJ: Lawrence Erlbaum.

Miller, I., and Miller, M. (1999), *John E. Freund's Mathematical Statistics*, Saddle River, NJ: Prentice-Hall.

Moore, D. S., and McCabe, G. P. (1999), *Introduction to the Practice of Statistics* (3rd ed.), New York: Freeman.

Rice, J. A. (1995), *Mathematical Statistics and Data Analysis* (2nd ed.), Belmont, CA: Wadsworth.

Rothman, K. J. (1978), "A Show of Confidence," *New England Journal of Medicine*, 299, 1362–1363.

Discussion

Matthew A. CARLTON

The authors of this article rightly point out that many statistics students—and, we assume by extension, many statistics practitioners—regularly misinterpret the p value of a hypothesis test. My students most commonly fall into the trap that a p value indicates the probability that the null hypothesis is true. However, the authors adamantly insist that most confusion stems from the marriage of Fisherian and Neyman–Pearsonian ideas, that such a marriage is a catastrophic error on the part of modern statisticians, and that a radically different, Bayesian approach is the solution. In this response, I will try to highlight my agreements and disagreements with the article and to discuss briefly some key considerations for statistics teachers when they tackle hypothesis testing.

Matthew A. Carlton is Assistant Professor, Department of Statistics, California Polytechnic State University, San Luis Obispo, CA 93407 (E-mail: mcarlton@calpoly.edu).

One point made by the authors is well worth repeating: a p value is *not* the same thing as a Type I error rate. In fact, it is questionable whether we can interpret the p value as an error rate at all. Statisticians generally agree on the following definition:

Given a hypothesis H and a random sample of data, we define the P value to be the probability of observing data at least as contradictory to H as our data, under the assumption that H is true.

That definition does not sound anything like an error rate, so why would someone make that misinterpretation? As the authors note, the numerical comparison of the p value to the *significance level* α, combined with the identification of α as the probability of a Type I error, make the two seem analogous. Our own statistical terminology increases the confusion: some texts refer to the

p value as the *observed significance level*, further cementing in the students' minds that p values and α's are somehow siblings.

In the authors' words: "P values are not a repetitive error rate." This conclusion, which seems quite reasonable in light of the above definition, distinguishes the p value from α, which we understand to represent the long-run error rate across all tests of true null hypotheses. That is,

$$\alpha = \Pr_{\mathrm{H}}(\text{Reject H}),$$

where H is the null hypothesis being tested and "Reject H" abbreviates the rejection rule for our particular test. In a frequentist sense, a 5% significance level ensures that, in the long run, (at most) 5% of all *true* hypotheses will be mistakenly rejected by our standard hypothesis testing procedure.

The confusion brought about by the interplay of these two quantities (the p value and α) and the vocabulary we use to describe them indeed warrants attention, and I am grateful to the authors for cautioning the statistical community on their misuse. But the authors claim this confusion runs rampant among researchers, without any real evidence to back up that concern. More troubling, they seem intent on establishing that P values and Type I errors cannot coexist in the same universe. It is unclear whether the authors have given any substantive reason why we cannot utter "p value" and "Type I error" in the same sentence. Indeed, a good portion of their article is devoted to attacking all those who would dare combine the work of Fisher, Neyman, and Pearson without any justification for their hostility, other than the confusion between p values and α's per se. They write: "[C]onsider the widely unappreciated fact that [Fisher's] p value is *incompatible* with the Neyman–Pearson hypothesis test in which it has become embedded" (their emphasis). The "fact" of their incompatibility comes as surprising news to me, as I'm sure it does to the thousands of qualified statisticians reading the article. The authors even seem to suggest that among the reasons statisticians should now divorce these two ideas is that Fisher and Neyman were not terribly fond of each other (or each other's philosophies on testing). I have always viewed our current practice, which integrates Fisher's and Neyman's philosophies and permits discussion of both P values and Type I errors—though certainly not in parallel—as one of our discipline's greater triumphs.

The authors then argue two seemingly contradictory lines of thought on the respective worth of p values and α's. They correctly note that a rejection decision in invariant under a one-to-one transformation, so that, for example, "$z > z_\alpha$" might equate to "p value $< \alpha$" in the appropriate setting. But they state: "only the fact that $p < \alpha$ is of any interest, not the specific value of p itself." That statement follows the strict Neyman–Pearson philosophy, but most statisticians recognize that the p value also gives us a sense of the magnitude of disparity between our data and the null hypothesis (Fisher's "measure of evidence"). The authors follow up this statement shortly thereafter with: "Indeed, if a researcher is interested in the 'measure of evidence' provided by the p value, we see no use in also reporting the error probabilities, since they do not refer to any property that the p value has." Arguably, the researcher following a strictly Fisherian philosophy of measuring significance may not need α, but those who will make real-world decisions based upon the research (the FDA, the actuary's boss) do.

Finally, the authors offer their solution to what they allege is "near-universal confusion among researchers over the meaning of p values and α levels" (though, again, they fail to present significant data to support this allegation). They propose a Bayesian approach: in essence, they have imposed a prior distribution on their hypotheses (e.g., half of all new treatments are effective) and thereby computed that using a "p value $< 5\%$" rule results in 22% of rejected hypotheses being true. They imply that the disparity between 5% and 22% is an indictment of the p value method while failing to mention that these two percents do not measure the same thing. A 5% cutoff means that 5% of true hypotheses are rejected, not that 5% of rejected hypotheses are true. Conceding for the moment the inherent validity of their Bayesian framework, they have compared $\Pr(x|\mathrm{H}) = 5\%$ to $\Pr(\mathrm{H}|x) = 22\%$ and become distraught at their dissimilarity.

Back to the main issue: they profess that, by having statistical practitioners compute the "*conditional* α" (their emphasis), $(1 + [-e\,p \log p]^{-1})^{-1}$, the Fisher–Neyman–Pearson feud shall be resolved, and all this unnecessary confusion about evidence versus error shall go away. If the whole point is ostensibly to *avoid* confusion among modern researchers, how does the Bayesian methodology in general, and this nonintuitive formula in particular, help? Surely the authors are not suggesting that we teach Bayesian hypothesis testing, let alone this lower bound for a conditional error rate, at the high school or undergraduate service course level.

If researchers really do misinterpret and misapply statistical testing methodology, then a main initiative as statisticians should be to *access* these individuals so we can re-educate them, not to discard the procedural paradigm itself.

So, what should a statistics educator do with p values and α's—throw away one (or both) of these notions in favor of conditional errors and Bayesian methodology? Not in my view. Although the authors' warnings do make me reconsider whether I should focus more on the "classical" rejection region method to avoid confusion, I still believe students (and, by extension, professional researchers) can handle both p values and α's if taught carefully and thought about carefully.

Drawing on my own experiences teaching statistics, I would advocate the following game plan for teaching hypothesis testing in a manner that incorporates both Fisherian and Neyman–Pearsonian ideas.

- Begin with an informal example, avoiding the terms "p value" and "level of significance (α)." Lead them through the logical process: assuming this particular claim is true, what does the sampling distribution of our statistic look like? What would be the chances of observing a sample statistic value of ___ or more extreme? Now, suppose we take a random sample and we do observe a statistic value of ___. What can we say about the credibility of the original claim?

This leads into Fisher's dichotomy that either a rare event has occurred, or the predicate claim is false. Incidentally, I see a trichotomy when presented with a small p value: a rare event has occurred, or the predicate claim is false, *or* the original sample was not random. When I ask students, "why would we get a

probability so low?" they generally offer up a nonrandom sample as the first explanation.

- Next, formalize the procedure. Introduce the terminology: null hypothesis, alternative hypothesis, probability value (aka p value). Identify these elements of the original informal example. Emphasize that we questioned the credibility of the claim because this probability was small.

- Almost invariably, students will ask what qualifies as a "small p value." *Now* is the time to introduce the *significance level*, α. I usually describe α as a predetermined cutoff for what we consider "small" and "not small" probabilities in the context of our hypothesis test. At this point, students should accept that the significance level for a test is an agreed upon (e.g., 5%) but, in some sense, arbitrary choice. I teach my students to (1) report their p value, not just "Reject H_0" or "Fail to Reject H_0" and (2) frame their conclusions relative to α; for example, at the 5% significance level, we cannot reject the claim that. . . .

- Emphasize the difference between statistical significance (the p value was less than 5%) and practical significance (the evidence suggests a process change has occurred and that we should react to that change). Many introductory textbooks give nice examples of this seemingly minor idea, but at its heart is a key philosophical difference between Neyman–Pearson and Fisher.

Notice that, thus far, I have made no mention of Type I error, though I claimed it can and should be maintained in our current hypothesis testing framework. I generally wait until I am well into hypothesis testing before I introduce the topics of Type I and Type II error. This distance between when the students learn about p values and when they learn about error rates seems to greatly reduce their tendency to then misinterpret a p value itself as an error rate.

- Finally, introduce Type I and Type II errors. Explain what the two potential errors are, both generically and in the context of a particular example. Then, inform them that our heretofore "arbitrary" α is, in fact, the probability of committing a Type I error (assuming a simple null, of course). I generally do not prove this fact to my introductory level students; they seem to take my word for it, and a "proof" with any real rigor would only confuse them. They should understand that Type I and II errors balance each other, in the sense that to reduce the chance of one error a priori we must raise the chance of the other. This brings up an important point: emphasize to the students that α must be selected in the context of the problem, but *in advance of observing the data*. The choice of α (5%, lower, or higher) reflects the researcher's judgment on the relative severity of the *practical* consequences of Type I and Type II errors in the context of his or her situation.

One last note: How can we make students understand the definition of a p value? The classic analogy of hypothesis testing to a jury trial serves well here: suppose you are seated on a jury in a criminal trial. You are required by law to operate under a "presumption of innocence"; that is, to assume the defendant's claim to be true until proven otherwise. In the jury room, your deliberations boil down to the following question: What's the chance the prosecution could amass this much evidence if the defendant were innocent? That is what the p value, in a loose sense, calculates: the chance of acquiring this much evidence against a true null hypothesis. (This also allows us to interpret α as the cut-off between "reasonable doubt" and "beyond reasonable doubt.")

Rejoinder

We thank the discussants for their valuable insights and interesting comments on this difficult and fascinating area. In fact, Matthew Carlton has already answered many of the issues raised by Kenneth Berk, so we will not repeat his clarifying comments. We organize our brief rejoinder around three main themes that interest both discussants.

1. Emphasis of the article. We want to clearly state that the main goal of this article is to warn against the extensive misconception that a p value is a frequentist measure just because it is computed as a tail area. Frequentism aims at reporting measures of performance that behave nicely in the long run, in the sense that the average *reported* performance is no better than the long run *actual* performance of the statistical procedure. P values simply do not have this property. We wanted to use Fisher's and Neyman's words because they perfectly understood their respective philosophies. It was later that these philosophies were cobbled together, the end result being that the p value is now indelibly linked in researchers' minds with the Type I error rate, α. And this is precisely what Fisher (1955, p. 74) had complained about when he accused Neyman–Pearson of attempting "to assimilate a test of significance to an acceptance procedure." This assimilation is now all but complete in the statistics curriculum.

As we stated in our article, Fisher saw the p value as a measure of evidence, not as a frequentist evaluation. Unfortunately, as a measure of evidence it is very misleading; but this is the topic of other papers, not this one. The main problem is that the p value computes not the probability (or density) of the *observed* data given the null hypothesis, but the probability of this and *more extreme data*. Most difficulties with p values stem from the natural (but wrong) tendency to interpret the latter as the former. If the p value had been introduced as the probability (or density) of the observed data (and only that data) under the null model, the confusion between p's and α's and the exaggeration of the evidence against H_0 provided by a p value would probably have never happened. (Notice that in the computation of the p value, very extreme data, strongly unfavorable to the null hypothesis is included.)

2. The proposed solution. Our original intent was not to propose a solution to the p's versus α's dilemma. We did this on purpose so as not to distract from the main goal of our article,

stated previously, which is a crucial one in our view. Our aim was to reveal how the anonymous merger of two incompatible testing paradigms now dominates statistical teaching and practice. We only touched on our favorite solution to this incompatibility at the request of a referee, and did so just to introduce the reader to relevant papers addressing this topic. Section 5.2 of our article has only this modest role; it is not intended to fully explain or defend the adequacy of the proposed novel procedure, since other papers we cited already do this.

The proposed solution is a novel methodology, and hence difficult to grasp without careful reading of the mentioned papers, and definitely impossible from the very little mention we make of it in Section 5.2. We do, however, want to emphasize that the conditional frequentist test to which we briefly refer is *not* a Bayesian solution. Rather, this solution is derived under strict frequentist methodology, computes frequentist error probabilities, and conditions on p values as statistics reflecting the "evidence" in the data (see Sellke, Bayarri, and Berger 2001, for details).

3. The correct teaching of hypothesis testing. This extremely important issue itself deserves several articles (with discussion). Again, however, it was not the focus of this article. Let us, nevertheless, make some brief remarks on Berk's and Carlton's comments:

a. Hypothesis testing problems. Both discussants refer to the convenience of treating some hypothesis testing problems as estimation problems. We agree that many problems posed and solved as hypothesis testing problems are in reality estimation problems. Only "real" hypotheses (those having some clear, distinctive, meaning and plausibility) should ever enter a hypothesis testing problem. We should never use hypothesis testing tools for a problem that is intrinsically an estimation problem (nor vice versa, although this misuse is less common). It is especially important when the null hypothesis is a point null or a precise null approximated by a point null. [For conditions under which this approximation is valid, see Berger and Sellke (1987); in fact, when n is very large, this is usually not possible. Realization of this fact would alleviate many of the problems encountered when testing point nulls with very large n.]

b. Role of p values in testing. As a measure of evidence, the p value is so misleading that there is real danger in teaching it (unless it is used in a much weaker sense as a conditioning statistic). In a purely frequentist framework, there seems to be no place for p values at all. A very modest role for them might be as handy one-to-one transformations (when possible) of the test statistics, one that allows easy checking of how far on the tail of the null distribution the observed value of the test statistic is. This makes them convenient in reporting (in papers, or in outputs from computer software). However, even this modest role poses more real dangers than advantages. Indeed, the use of the $p < \alpha$ criterion has been the source of much of the confusion that we have attempted to convey in our article. We would be much happier if its usage were dropped altogether.

Measures of evidence (as p values or likelihood ratios) have radically different interpretations to frequentist measures of performance (as Type I errors and power functions), and mixing both of them in the same analysis is always delicate. It certainly requires a fair amount of statistical training and sophistication, and we believe that it is way too dangerous to encourage its use by the casual, statistically untrained, user.

c. What should we do? This is an extremely important issue, and one which cannot be adequately addressed in this brief rejoinder. We wish we had the magic solution that would please everyone, but we do not. We have our own preferred solutions, but we do not want to enter into this argument here. This article warns against one of the things we should *not* do: *We should not take a p value as a Type I error adjustable with data*. Smart applied statisticians are aware of the problem, but users at large are not. Other articles have already warned against interpreting a small p value as important evidence against the null hypothesis. Last, the confusion between p values and posterior probabilities is well understood and taught in most elementary courses (however, this is still probably the mistake most frequently committed by students and users).

[Received October 2001. Revised May 2003.]

REFERENCES

Berger, J. O., and Sellke, T. (1987), "Testing a Point Null Hypothesis: The Irreconcilability of P Values and Evidence" (with comments), *Journal of the American Statistical Association*, 82, 112–139.

Fisher, R. A. (1955), "Statistical Methods and Scientific Induction," *Journal of the Royal Statistical Society*, Ser. B, 17, 69–78.

Sellke, T., Bayarri, M. J., and Berger, J. O. (2001), "Calibration of p Values for Testing Precise Null Hypotheses," *The American Statistician*, 55, 62–71.