P Values are not Error Probabilities

Raymond Hubbard
College of Business and Public Administration
Drake University
Des Moines, IA 50311
Phone: (515) 271-2344
Fax: (515) 271-4518
E-mail: Raymond.Hubbard@drake.edu

and

M.J. Bayarri
Department of Statistics and Operations Research
University of Valencia
Burjassot
Valencia 46100
Spain
Phone: 34 96 354-4309
Fax: 34 96 354-4735
E-mail: Susie.bayarri@uv.es

November 2003

Author Footnotes

Raymond Hubbard is the Thomas F. Sheehan Distinguished Professor of Marketing, Drake University, Des Moines, IA 50311. M.J. Bayarri is Professor of Statistics, University of Valencia, Burjassot, Valencia 46100, Spain. The authors would like to thank Stuart Allen, Scott Armstrong, James Berger, Steven Goodman, Rahul Parsa, and Daniel Vetter for comments on earlier versions of this manuscript. 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.

P-values are not Error Probabilities

Abstract

Confusion surrounding the reporting and interpretation of results of classical statistical tests is widespread among applied researchers. The confusion stems from the fact that most of these researchers are unaware of the historical development of classical statistical testing methods, and the mathematical and philosophical principles underlying them. Moreover, researchers erroneously believe that the interpretation of such tests is 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 views of hypothesis testing and inductive behavior. Unfortunately, statistics textbooks tend to inadvertently cobble together elements from both of these schools of thought, thereby perpetuating the confusion. So complete is this misunderstanding over measures of evidence versus error that is not viewed as even being a problem among the vast majority of researchers. The upshot is that despite supplanting Fisher's significance testing paradigm some fifty years or so ago, recognizable applications of Neyman–Pearson theory are few and far between in empirical work. In contrast, Fisher's influence remains pervasive. Professional statisticians must adopt a leading role in lowering confusion levels by encouraging textbook authors to explicitly address the differences between Fisherian and Neyman–Pearson statistical testing frameworks.

KEY WORDS: Conditional Error Probabilities; Fisher Approach; Hypothesis Test; Inductive Behavior; Inductive Inference; Neyman–Pearson Approach; Significance Test; Teaching Statistics.

1. INTRODUCTION

Many users of statistical tests in the management, social, and medical sciences routinely invest them with properties they do not possess. (Use of the expression "statistical tests" rather than the more popular "significance tests" will become apparent shortly.) Thus, it has been pointed out, often by nonstatisticians (e.g., Carver 1978; Cohen 1994; Hubbard and Ryan 2000; Lindsay 1995; Nickerson 2000; Sawyer and Peter 1983), that the outcomes of these tests are mistakenly believed to yield the following information: the probability that the null hypothesis is true; the probability that the alternative hypothesis is true; the probability that an initial finding will replicate; whether a result is important; and whether a result will generalize to other contexts. These common misconceptions about the capabilities of statistical tests point to problems in classroom instruction.

Unfortunately, matters get worse: The extent of the confusion surrounding the reporting and interpretation of the results of statistical tests is far more pervasive than even the above misunderstandings suggest. It stems from the fact that most applied researchers are unfamiliar with the nature and historical origins of the classical *theories* of statistical testing. This, it should be added, is through no fault of their own. Rather, it reflects the way in which researchers are usually taught "statistics."

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) has 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. (Actually, it was Karl Pearson, and not Fisher, who introduced the p value in his chi-squared test—see Inman (1994)—but there is no doubt that Fisher was responsible for popularizing its use.) The distinction between *evidence* (p's) and *errors* (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 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 paper has three objectives. First, we outline the marked differences in the conceptual foundations of the Fisherian and Neyman–Pearson statistical testing approaches. Whenever possible, we let the protagonists speak for themselves. This is vitally important in view of the manner in which their own voices have been muted over the years, and their competing ideas unwittingly merged and distorted in many statistics textbooks. Because of the widespread practice of textbook authors' failing to credit Fisher and Neyman–Pearson for their respective methodologies, it is small wonder that present researchers remain unaware of them.

Second, we show how the rival ideas from the two schools of thought have been unintentionally mixed together. Curiously, this has taken place despite the fact that Neyman–Pearson, and not Fisherian, theory is regarded as classical statistical orthodoxy (Hogben 1957; Royall 1997; Spielman 1974). In particular, we illustrate how this mixing of statistical testing methodologies has resulted in widespread confusion over the interpretation of *p* values (evidential measures) and α levels (measures of error). We demonstrate that this confusion was a problem between the Fisherian and Neyman–Pearson camps, is not uncommon among statisticians, is prevalent in statistics textbooks, and is well nigh universal in the pages of leading (marketing) journals. This mass confusion, in turn, has rendered applications of classical statistical testing all but meaningless among applied researchers. And this points to the need for changes in the way in which such testing is approached in the classroom.

Third, we suggest how the confusion between *p*'s and α's may be resolved. This is achieved by reporting *conditional* (on p-values) error probabilities.

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. Despite beginning life as a Bayesian (Zabell 1992), Fisher soon grew disenchanted with the subjectivism involved, and 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 of the direct probability, or Pr(x | H). This transition was facilitated by his conviction that: "it is possible to argue from consequences to causes, from observations to hypotheses" (Fisher 1996, 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 | H_0). As intuitive as this might be, it is not useful for continuous variables. Thus, a significance test is defined as a procedure for establishing the probability of an outcome, as well as more extreme ones, on a null hypothesis of no effect or relationship. The distinction between the "probability" of the observed data given the null and the probability of the observed and *more extreme data* given the null is crucial: not only it has contributed to the confusion between *p*'s and α's, but also results in an exaggeration of the evidence against the null provided by the *observed* data.

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 ≤ .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 World War II. It is widely thought of as constituting the basis of classical statistical testing (Carlson 1976; Hogben 1957; LeCam and Lehmann 1974; Nester 1996; 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 (Fisher 1955). (We keep the traditional name "Neyman–Pearson" to denote this school of thought, although Lehmann [1993] mentions 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).

Or as Pearson (1990, p. 82) put it: "The rational human mind did not discard a hypothesis unless it could conceive at least one plausible *alternative* hypothesis." (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.

In contradistinction to Fisher's ideas about hypothetical infinite populations, Neyman–Pearson results are predicated on the assumption of repeated random sampling from a defined population. Consequently, Neyman–Pearson theory is best suited to situations in which repeated random sampling has meaning, as in the case of quality control experiments. In such restricted circumstances, the Neyman–Pearson frequentist interpretation of probability makes sense: α is the *long-run* frequency of Type I errors and β is the counterpart for Type II errors.

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. The power of a test to detect a particular effect size in the population can be calculated before conducting the research, and is therefore considered to be useful in the design of experiments. Because Fisher's statistical testing procedure admits of no alternative hypothesis (H_A), the concepts of Type II error and the power of the test are not relevant. Fisher made this clear when chastising Neyman and Pearson without naming them: “In fact … ‘errors of the second kind’ are committed only by those who misunderstand the nature and application of tests of significance” (Fisher 1935c, p. 474). And he subsequently added that “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), and in fact vigorously opposed its incorporation by Neyman–Pearson (Hacking 1965).

Fisher nevertheless hints at the idea of the power of a test when he refers to the “sensitiveness” of an experiment:

> “By increasing the size of the experiment we can render it more sensitive, meaning by this that it will allow of the detection of a lower degree of sensory discrimination, or, in other words, of a quantitatively smaller departure from the null hypothesis. Since in every case the experiment is capable of disproving, but never of proving this hypothesis, we may say that the value of the experiment is increased whenever it permits the null hypothesis to be more readily disproved” (Fisher 1966, pp. 21-22).

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.

Whereas 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. According to Neyman:

> “The description of the theory of statistics involving a reference to behavior, for example, *behavioristic statistics,* has been introduced to contrast with what has been termed *inductive reasoning.* Rather than speak of inductive reasoning I prefer to speak of *inductive behavior”* (Neyman 1971, p. 1; original emphasis).

And “The term ‘inductive behavior’ means simply the habit of humans and other animals (Pavlov’s dog, etc.) to adjust their actions to noticed frequencies of events, so as to avoid undesirable consequences” (Neyman 1961, p. 48). 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. As they indicated:

> “… in some cases it will be more important to avoid the first [type of error], in others the second [type of error]… From the point of view of mathematical theory all we can do is to show how the risk of errors may be controlled or minimised. The use of these statistical tools in any given case, in determining just how the balance should be struck, must be left to the investigator” (Neyman and Pearson 1933, p. 296).

After taking such advice into account, 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). In determining what this bound on α should be, Neyman later stated that the control of Type I errors was more important than that of Type II errors:

> “The problem of testing statistical hypotheses is the problem of selecting critical regions. When attempting to solve this problem, one must remember that the purpose of testing hypotheses is to avoid errors insofar as possible. Because an error of the first kind is more important to avoid than an error of the second kind, our first requirement is that the test should reject the hypothesis tested when it is true very infrequently… To put it differently, when selecting tests, we begin by making an effort to control the frequency of the errors of the first kind (the more important errors to avoid), and then think of errors of the second kind. The ordinary procedure is to fix arbitrarily a small number α … and to require that the probability of committing an error of the first kind does not exceed α (Neyman 1950 p. 265).

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 (Lehmann 1993), or fixed size (Seidenfeld 1979) approach. 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 H_0 is rejected in favor of H_A, otherwise 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 specific 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). But since scientists are in the business of gleaning evidence from individual studies, this limitation of Neyman–Pearson theory is severe.

Neyman–Pearson theory is *non-evidential*. Fisher recognized this deficiency, commenting that their "procedure is devised for a whole class of cases. No particular thought is given to each case as it arises, nor is the tester's capacity for learning exercised" (Fisher 1959, p. 100). Instead, the researcher can only make a decision about the likely outcome of a hypothesis as if it had been subjected to numerous and identical repetitions, a condition that Fisher (1956, p. 99) charged "will never take place" in normal scientific research. In most applied work, repeated random sampling is a myth because empirical results tend to be based on a single sample.

Fisher did agree that what he called the Neyman–Pearson "acceptance procedures" approach could play a part in quality control decisions: "I am casting no contempt on acceptance procedures, and I am thankful, whenever I travel by air, that the high level of precision and reliability required can really be achieved by such means" (Fisher 1955, pp. 69-70). 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:

> "The 'Theory of Testing Hypotheses' was a later attempt, by authors who had taken no part in the development of [significance] tests, or in their scientific application, to reinterpret them in terms of an imagined process of acceptance sampling, such as was beginning to be used in commerce; although such processes have a logical basis very different from those of a scientist engaged in gaining from his observations an improved understanding of reality" (Fisher 1959, pp. 4–5).

And in drawing further distinctions between the Fisherian and Neyman–Pearson paradigms, Fisher reminds 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–Pearson 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

The rank and files of users of statistical tests in the management, social, and medical sciences are unaware of the above distinctions between the Fisherian and Neyman–Pearson camps (Gigerenzer 1993; Goodman 1993; Royall 1997). As previously acknowledged, this is not their fault; after all, they have been taught from numerous well-regarded textbooks on statistical analysis. Unfortunately, many of these same textbooks combine (sometimes incongruous) ideas from both schools of thought, usually without acknowledging, or worse yet, recognizing, this. That is, although the Neyman–Pearson approach has long since attained the status of orthodoxy in classical statistics, Fisher's methods continue to permeate the literature (Hogben 1957; Spielman 1974).

Johnstone (1986) remarks 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. In addition, textbooks descriptions of Neyman–Pearson theory often refer to the Type I error probability as the "significance level" (Goodman 1999; Kempthorne 1976; Royall 1997).

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 (significance probability) 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.

The confusion surrounding researchers 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 and is also prevalent in textbooks. 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, i.e., 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) agree 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

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. On the contrary, Kalbfleisch and Sprott (1976) allege that statisticians commonly make the mistake of equating *p* values with Type I error rates. And their allegations find ready support in the literature. For example, Gibbons and Pratt (1975, p. 21), in an article titled “*P* Values: Interpretation and Methodology,” erroneously state: “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.” Barnard (1985, p. 7) is similarly at fault when he remarks, “For those who need to interpret probabilities as [long run] frequencies, a *P*-value ‘measures’ the possibility of an ‘error of the first kind,’ arising from rejection of H_0 when it is in fact true.” Again, Hung, O’Neill, Bauer, and Köhne (1997, p. 12) note 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 paper:

> “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 (1986) and Berger and Delampady (1987, p. 329) correctly insist 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 100α% of the true H_0, but the data-dependent-*P*-values have no such interpretation.” (Original emphasis). Lindsey (1999) agrees that the *p* value has no clear long-run meaning in classical frequentist inference. In sum, although *p*’s and α’s have very different meanings, Bayarri and Berger (2000) nevertheless contend that among statisticians there is a near ubiquitous misinterpretation of *p* values as frequentist error probabilities. And inevitably, this fallacy shows up in statistics textbooks, as when Canavos and Miller (1999, p. 255) stipulate: “If the null hypothesis is true, then a type I error occurs if (due to sampling error) the *P*-value is less than or equal to α.”

Indeed, in his effort to partially resolve differences between the Fisherian and Neyman–Pearson viewpoints, Lehmann (1993) also fails to distinguish between measures of evidence versus error. He calls the Type I error rate the significance level of the test, when for Fisher this was determined by *p* values and not α’s. And we have seen that misconstruing the evidential *p* value as a Neyman–Pearson Type I error rate was anathema to Fisher.

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 a paper published in the *Encyclopedia of Statistical Sciences* intended to clarify the meaning of *p* values, for example, Gibbons (1986, p. 367) falsely concludes 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) states: "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). Sackrowitz and Samuel-Cahn (1999) also subscribe to this approach, as do Lehmann (1978), and Bhattachayra and Habtzhi (2002).

Given the above, it is easy to see how similar misinterpretations are perpetuated in statistics textbooks. Canavos and Miller (1999, p. 254), for example, who earlier confused *both p* values and α levels with the Type I error rate, do likewise with regard to the significance level: "When a specific cutoff level for the p value is agreed upon in advance as the basis for a formal conclusion, it is called the **level of significance** and is denoted by α." (Original emphasis). Berenson and Levine's (1996, p. 394) textbook does the same:

"- If the p value is greater than or equal to α, the null hypothesis is not rejected.
- If the p value is smaller than α, the null hypothesis is rejected."

A few remarks from Keller and Warrack (1997) further demonstrate the widespread nature of the anonymous mixing of Fisherian with Neyman–Pearson ideas in some statistics textbooks, and the conceptual headaches this is likely to create for students and researchers. In a section titled "The p Value of a Hypothesis Test," they state:

> "What is really needed [in a study] is a measure of how much statistical evidence exists…. In this section we present such a measure: the *p*-value of a test…. The *p*-value of a test of hypothesis is the smallest value of α that would lead to rejection of the null hypothesis…. It is important to understand that the calculation of the p value depends on, among other things, the alternative hypothesis…. The *p*-value is an important number because it measures the amount of statistical evidence that supports the alternative hypothesis (Keller and Warrack 1997, p. 346, 347, 349).

These points are incorrect. It has already been shown that interpreting p values in single (or ongoing) experiments is not permissible in a Neyman–Pearson hypothesis testing context. Their model is behavioral, not evidential. Next, Keller and Warrack (1997), like Berenson and Levine (1996), falsely equate p's with α's when recommending the $p < \alpha$ statistical significance result strategy. They then compound their misconceptions about statistical testing when claiming that both the calculation and interpretation of a p value depend on the alternative hypothesis. This is not so. The calculation of a p value depends only on the truth of the null hypothesis. Fisher, as we have seen, had no time for the alternative hypothesis introduced by Neyman–Pearson. What is more, the p value does not measure the amount of evidence supporting H_A; it is a measure of inductive evidence against H_0. Moreover, Neyman and Pearson would not endorse this evidential interpretation of p values espoused by Keller and Warrack

(1997). In the first place, *the p value plays no role in their theory*. Secondly, and to reiterate, Neyman–Pearson theory is non-evidential.

Instead, the Neyman–Pearson 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. Interestingly, this last assertion contradicts Fisher's (1966, p. 16) adage that "the null hypothesis is never proved or established, but is possibly disproved, in the course of experimentation." Otherwise expressed, the familiar claim that "one can never accept the null hypothesis, only fail to reject it" is a characteristic of Fisher's significance test, and not the Neyman–Pearson hypothesis test. In the latter's paradigm one can indeed "accept" the null hypothesis.

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 100α% 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 problematical. From a pragmatic point of view we do not recommend it either, since 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 Sections 5 and 6.) 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) note that because of its superficial resemblance to the Neyman–Pearson Type I error rate, α, Fisher's *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 points 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).

Devore and Peck's (1993, p. 451) statistics textbook illustrates Goodman's point: "The smallest α for which H_0 could be rejected is determined by the tail area captured by the computed value of the test statistic. This smallest α is the *P*-value." Or consider in this context another erroneous passage from a statistics textbook:

> "We sometimes take one final step to assess the evidence against H_0. We can compare the *P*-value with a fixed value that we regard as decisive. This amounts to announcing in advance how much evidence against H_0 we will insist on. The decisive value of *P* is called the **significance level**. We write it as α, the Greek letter alpha" (Moore 2000, p. 326; original emphasis).

1.3 *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).

That Fisher popularized such nominal levels of statistical significance is itself an interesting, not to say extremely influential, historical quirk. While working on *Statistical Methods for Research Workers* Fisher was denied permission by Karl Pearson to reproduce W.P. Elderton's table of χ^2 from the first volume of *Biometrika*, and therefore prepared his own version. In doing so, Egon Pearson (1990, p. 52) informs us: "[Fisher] gave the values of [Karl Pearson's] χ^2 [and Student's *t*] for selected values of *P* … instead of *P* for arbitrary χ^2, *and thus introduced the concept of nominal levels of significance.*" (Our emphasis). As noted, Fisher's use of 5% and 1% levels was similarly adopted, and ultimately institutionalized, by Neyman–Pearson. And Fisher (1959, p. 42) rebuked them for doing so, explaining: "…no scientific worker has a fixed level of significance at which from year to year, and in all circumstances, he rejects hypotheses; he rather gives his mind to each particular case in the light of his evidence and his ideas." Despite this rebuke, 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 the 1% levels.

Many researchers will not 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 widely 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. Because some statisticians and textbooks on the subject are unclear about the differences between *p*’s and α’s, it is anticipated that confusion levels in journal articles will be high.

1.4 *Confusion Over p’s and α’s in Marketing Journals*

The manner in which the results of statistical tests are reported in marketing journals is used as an empirical barometer for practices in other applied disciplines. We doubt whether the findings reported here would differ substantially from those in other fields.

More specifically, *two* randomly selected issues of each of three leading marketing journals—the *Journal of Consumer Research, Journal of Marketing,* and *Journal of Marketing Research*—were analyzed for the eleven-year period 1990 through 2000 in order to assess the number of empirical articles and notes published therein. This procedure yielded a sample of 478 empirical papers. These papers were then examined to see whether classical statistical tests had been used in the data analysis. Some 435, or 91.0%, employed such testing.

Although the evidential *p* value from a significance test violates the orthodox Neyman–Pearson behavioral hypothesis testing schema, Table 1 shows that *p* values are commonplace in marketing’s empirical literature. Conversely, α levels are in short supply.

Of the 435 papers using statistical tests, fully 312, or 71.7%, employed what Goodman (1993) calls “roving alphas,” i.e., a discrete, graduated number of *p* values masquerading variously as Type I error rates and/or measures of evidence against H_0, usually at the $p < .05$, $p < .01$, $p < .001$ values, etc. In other words, these *p* values may sometimes constitute an “observed” Type I error rate in the sense that they are not even pre-assigned, or fixed, *p*’s/α’s; rather, they are variable, *de facto,* “error rates” determined solely by the data. In addition, these same *p* values will be interpreted simultaneously in a quasi-evidential manner as a basis for rejecting H_0 if $p < \alpha$. This includes, in many cases, erroneously using the *p* value as a proxy measure for effect sizes (e.g., $p < .05$ is “significant,” $p < .01$ is “very significant,” $p < .001$ is “extremely significant,” and so on). In sum, these “roving alphas” are habitually misinterpreted by applied researchers.

A further 19 (4.4%) chose to report “exact” p values, while an additional 61 (14.0%) opted to present various combinations of exact p’s with either “roving alphas” or fixed p values. Conservatively, therefore, 392, or 90.1%, of empirical articles in a sample of marketing journals report the results of statistical tests in a manner that is incompatible with Neyman–Pearson orthodoxy. Another 4 (0.9%) studies were not sufficiently clear about the disposition of a finding (beyond statements such as “this result was statistically significant at conventional levels”) in their accounts.

This leaves 39 (9.0%) studies as eligible for the reporting of “fixed” level α values in the fashion intended by Neyman–Pearson. Unfortunately, 21 of these 39 studies reported “fixed p” rather than fixed α levels. After subtracting this group, only 18 (4.1%) studies remain eligible. Of these 18, some 13 simply refer to their published results as being “significant” at the .05, .01 levels, etc. No information about p values or α levels is provided. Finally, only 5 of 435 empirical papers using statistical tests, or 1.1%, explicitly used fixed α levels.

Insert Table 1 about here

2. DISCUSSION

Confusion over the interpretation of classical statistical tests is so complete as to render their application almost meaningless. As we have seen, this chaos extends throughout the scholarly hierarchy from the originators of the test themselves—Fisher and Neyman–Pearson—to some fellow professional statisticians to textbook authors to applied researchers.

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—and the synopsis provided in Table 2—the two concepts of statistical significance could scarcely be further apart in meaning.

Insert Table 2 about here

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. Usually, in such texts, an anonymous account of standard Neyman–Pearson doctrine is put forward initially, and is often followed by an equally anonymous discussion of "the p value approach." This transition from (and mixing of) α levels to p values is typically seamless, as if it constitutes a natural progression through different parts of the same coherent statistical whole. It is revealed in the following passage from one such textbook: "In the next subsection we illustrate testing a hypothesis by using various values of α, and we see that this leads to defining the ***p* value**…." (Bowerman et al., 2001, p. 300; original emphasis).

Unfortunately, this nameless amalgamation of the Fisherian and Neyman–Pearson paradigms, with the p value serving as the conduct, has indeed 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. It is now ensconced in college curricula, textbooks, and journals.

3. 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.

3.1 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 paper, 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 around .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 (like 1 in 20).

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).

3.2 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 longstanding 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 is clearly presented in 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 is given in 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. This is an extremely simple formula, and it provides the correct order of magnitude for interpreting a *p* value. (The calibration $- e\ p\ \log(p)$ can be interpreted as a lower bound on the Bayes factor.)

4. 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.

We have, nevertheless, indicated how the confusion over the meaning of *p*'s and α's may be resolved 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 the present paper will help to stimulate discussions along these lines.

REFERENCES

Barnard, G.A. (1985), *A Coherent View of Statistical Inference.* Technical Report Series, Department of Statistics & Actuarial Science, University of Waterloo, Ontario, Canada.

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

Berenson, M.L., and Levine, D.M. (1996), *Basic Business Statistics: Concepts and Applications* (6th ed.), Englewood Cliffs, NJ: Prentice-Hall.

Berger, J.O. (1986), "Are *P*-Values Reasonable Measures of Accuracy?" in *Pacific Statistical Congress,* eds. I.S. Francis, B.F.J. Manly and F.C. Lam, Amsterdam: Elsevier, 21–27.

, (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.

(2003), "Could Fisher, Jeffreys, and Neyman Have Agreed on Testing?" (with comments), Statistical Science, 18, 1–32.

and Delampady, M. (1987), "Testing Precise Hypotheses" (with comments), Statistical Science, 2, 317-352.

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.

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.

Bhattacharya, B., and Habtzghi, D. (2002), "Median of the p Value Under the Alternative Hypothesis," The American Statistician, 56, 202–206.

Bowerman, B.L. O'Connell, R.T., and Hand, M.L. (2001), Business Statistics in Practice (2nd ed.), New York: McGraw Hill.

Canavos, G.C., and Miller, D.M. (1999), An Introduction to Modern Business Statistics, New York: Duxbury Press.

Carlson, R. (1976), “The Logic of Tests of Significance,” Philosophical of Science, 43, 116–128.

Carver, R.P. (1978), “The Case Against Statistical Significance Testing,” Harvard Educational Review, 48, 378–399.

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.

Cohen, J. (1994), “The Earth is Round (p < .05),” American Psychologist, 49, 997–1003.

Devore, J., and Peck, R. (1993), Statistics: The Exploration and Analysis of Data, New York: Duxbury Press.

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,” Proceedings of the Society for Psychical Research, London, 39, 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.

(1935c), “Statistical Tests,” Nature, 136, 474.

(1945), “The Logical Inversion of the Notion of the Random Variable,” SankhyÔ, 7, 129-132.

(1955), “Statistical Methods and Scientific Induction,” Journal of the Royal Statistical Society, Ser. B., 17, 69–78.

(1956), Statistical Methods and Scientific Inference, Edinburgh: Oliver and Boyd.

(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, 366–368.

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, 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.

Hacking, I. (1965), Logic of Statistical Inference, New York: Cambridge University Press.

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

Hogben, L. (1957), Statistical Theory, New York: Norton.

Hubbard, R., and Ryan, P.A. (2000), "The Historical Growth of Statistical Significance Testing in Psychology—and Its Future Prospects," Educational and Psychological Measurement, 60, 661–684.

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.

Inman, H.F. (1994), “Karl Pearson and R. A. Fisher on Statistical Tests: A 1935 Exchange from Nature,” The American Statistician, 48, 2–11.

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, 267-277.

(1987b), “Tests of Significance Following R.A. Fisher,” British Journal for the Philosophy of Science, 38, 481-499.

Kalbfleisch, J.G., and Sprott, D.A. (1976), “On Tests of Significance,” in Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science, eds. W.L. Harper and C.A. Hooker, Dordrecht: Reidel, 259–270.

Keller, G., and Warrack, B. (1997), Statistics for Management and Economics (4th ed.), Belmont, CA: Duxbury.

Kempthorne, O. (1976), “Of What Use are Tests of Significance and Tests of Hypothesis,” Communications in Statistics, Part A—Theory and Methods, 8, 763–777.

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

LeCam, L., and Lehmann, E.L. (1974), “J. Neyman: On the Occasion of His 80th Birthday,” Annals of Statistics, 2, vii–xiii.

Lehmann, E.L. (1978), “Hypothesis Testing,” in International Encyclopedia of Statistics, Volume 1, eds. W.H. Kruskal and J.M. Tanur, New York: The Free Press, pp. 441-449.

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.

Lindsay, R.M. (1995), “Reconsidering the Status of Tests of Significance: An Alternative Criterion of Adequacy,” Accounting, Organizations and Society, 20, 35–53.

Lindsey, J.K. (1999), “Some Statistical Heresies” (with comments), The Statistician, 48, 1–40.

Moore, D.S. (2000), The Basic Practice of Statistics (2nd ed.), New York: Freeman.

Nester, M.R. (1996), “An Applied Statistician’s Creed,” The Statistician, 45, 401–410.

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.

(1957), “‘Inductive Behavior’ as a Basic Concept of Philosophy of Science,” International Statistical Review, 25, 7–22.

(1961), “Silver Jubilee of My Dispute with Fisher,” Journal of the Operations Research Society of Japan, 3, 145–154.

(1967), “R.A. Fisher (1890–1962), An Appreciation,” Science, 156, 1456-1460.

(1971), “Foundations o Behavioristic Statistics” (with comments), in Foundations of Statistical Inference, eds. V.P. Godambe and D.A. Sprott, Toronto: Holt, Rinehart and Winston of Canada, Limited, 1–19.

(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, 149-193.

(1977), “Frequentist Probability and Frequentist Statistics,” Synthese, 36, 97-131.

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.

and (1928b), “On the Use and Interpretation of Certain Test Criteria for Purposes of Statistical Inference. Part II,” Biometrika, 20A, 263-294.

and (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.

Nickerson, R.S. (2000), “Null Hypothesis Significance Testing: A Review of an Old and Continuing Controversy,” Psychological Methods, 5, 241–301.

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

(1990), 'Student' A Statistical Biography of William Sealy Gosset. Edited and augmented by R.L. Plackett and G.A. Barnard. Oxford: Clarendon Press.

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.

Sackrowitz, H., and Samuel-Cahn, E. (1999), "P Values as Random Variables—Expected P Values," The American Statistician, 53, 326–331.

Sawyer, A.G., and Peter, J.P. (1983), "The Significance of Statistical Significance Tests in Marketing Research," Journal of Marketing Research, 20, 122–133.

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.

Zabell, S.L. (1992), "R.A. Fisher and the Fiducial Argument," Statistical Science, 7, 369–387.

Table 1. The Reporting of Results of Statistical Tests in Three Leading Marketing Journals

“Roving alphas” (R)	*Exact P values (E_p)*	*Combination of E_n’s with fixed P values and “roving alphas”*	*“Fixed” level values*				*Unspecified*	*Total*
			Level	*P’s*	*“Significant”*	*α ’s*		
	19		*.10*	1	—	—	4	
		E_p + .05 2	*.05*	9	13	4		
			.01	7	—	1		
			.001	2	—	—		
		E_p + R 59	*Other*	2	—	—		
Total								
312	19	61		21	13	5	4	435
Percentage								
71.7	4.4	14.0		4.8	3.0	1.1	0.9	99.9

Table 2. Contrasting P's and α's

P–Value	α*–Level*
Fisherian Significance Level	Neyman–Pearson Significance Level
Significance Test	Hypothesis Test
Evidence Against H_0	Type I Error— Erroneous Rejection of H_0
Inductive Philosophy— From Particular to General	Deductive Philosophy— From General to Particular
Inductive Inference— Guidelines for Interpreting Strength of Evidence in Data	Inductive Behavior— Guidelines for Making Decisions Based on Data
Data-based Random Variable	Pre-Assigned Fixed Value
Property of Data	Property of Test
Short-Run— Applies to any Single Experiment/Study	Long-Run— Applies only to Ongoing Repetitions of Original Experiment/Study— Not to any Given Study
Hypothetical Infinite Population	Clearly Defined Population