The Fisher, Neyman–Pearson Theories of Testing Hypotheses: One Theory or Two?

E. L. LEHMANN*

The Fisher and Neyman–Pearson approaches to testing statistical hypotheses are compared with respect to their attitudes to the interpretation of the outcome, to power, to conditioning, and to the use of fixed significance levels. It is argued that despite basic philosophical differences, in their main practical aspects the two theories are complementary rather than contradictory and that a unified approach is possible that combines the best features of both. As applications, the controversies about the Behrens–Fisher problem and the comparison of two binomials (2 × 2 tables) are considered from the present point of view.

KEY WORDS: Behrens–Fisher problem; Conditioning; Power; p-value; Significance level.

1. INTRODUCTION

The formulation and philosophy of hypothesis testing as we know it today was largely created in the period 1915–1933 by three men: R. A. Fisher (1890–1962), J. Neyman (1894–1981), and E. S. Pearson (1895–1980). Since then it has expanded into one of the most widely used quantitative methodologies, and has found its way into nearly all areas of human endeavor. It is a fairly commonly held view that the theories due to Fisher on the one hand, and to Neyman and Pearson on the other, are quite distinct. This is reflected in the fact that separate terms are often used (although somewhat inconsistently) to designate the two approaches: significance testing for Fisher and hypothesis testing for Neyman and Pearson. (Since both are concerned with the testing of hypotheses, it is convenient here to ignore this terminological distinction and to use the term "hypothesis testing" regardless of whether the testing is carried out in a Fisherian or Neyman–Pearsonian mode.)

There clearly are important differences, both in philosophy and in the treatment of specific problems. These were fiercely debated by Fisher and Neyman in a way described by Zabell (1992) as "a battle which had a largely destructive effect on the statistical profession." I believe that the ferocity of the rhetoric has created an exaggerated impression of irreconcilability. The purpose of this article is to see whether there exists a common ground that permits a resolution of some of the principal differences and a basis for rational discussion of the remaining ones.

Some of the Fisher-Neyman debate is concerned with issues studied in depth by philosophers of science (see, for example, Braithwaite 1953; Hacking 1965; Kyburg 1974; and Seidenfeld 1979). I am not a philosopher, and this article is written from a statistical, not a philosophical, point of view.

Section 2 presents some historical background for the two points of view. Section 3 discusses the basic philosophical difference between Fisher and Neyman. (Although the main substantive papers [NP 1928 and 1933a] were joint by Neyman and Pearson, their collaboration stopped soon after Neyman left Pearson's Department to set up his own program in Berkeley. After that, the debate was carried on primarily by Fisher and Neyman.) Sections 4, 5, and 6 discuss three specific issues on which the two schools differ (fixed levels versus p values, power, and conditioning). Section 7 illustrates the effect of these differences on the treatment of two statistical problems, the 2 × 2 table and the Behrens–Fisher problem, that have become focal points of the controversy. Finally, Section 8 suggests a unified point of view that does not resolve all questions but provides a common basis for discussing the remaining issues.

For the sake of completeness, it should be said that in addition to the Fisher and Neyman–Pearson theories there exist other philosophies of testing, of which we shall mention only two. There is Bayesian hypothesis testing, which, on the basis of stronger assumptions, permits assigning probabilities to the various hypotheses being considered. All three authors were very hostile to this formulation and were in fact motivated in their work by a desire to rid hypothesis testing of the need to assume a prior distribution over the available hypotheses.

Finally, in certain important situations tests can be obtained by an approach also due to Fisher for which he used the term *fiducial.* Most comparisons of Fisher's work on hypothesis testing with that of Neyman and Pearson (see, for example, Barnett 1982; Carlson 1976; Morrison and Henkel 1970; Spielman 1974, 1978; Steger 1971) do not include a discussion of the fiducial argument, which most statisticians have found difficult to follow. Although Fisher himself viewed fiducial considerations to be a very important part of his statistical thinking, this topic can be split off from other aspects of his work, and here I shall consider neither the fiducial nor the Bayesian approach any further.

Critical discussion of the issues considered in this article with references to the extensive literature, in a wider context and from viewpoints differing from that presented here, can be found in, for example, Oakes (1986) and Gigerenzer et al. (1989).

2. TESTING STATISTICAL HYPOTHESES

The modern theory of testing hypotheses began with Student's discovery of the t test in 1908. This was followed by

* E. L. Lehmann is Professor Emeritus, Department of Statistics, University of California, Berkeley, CA 94720. This research was supported by National Science Foundation Grant DMS-8908670. The author thanks the referees for helpful suggestions, Sandy Zabell for suggesting improvement to an early version of the article, and his wife Juliet Shaffer for discussions and critical comments at all stages.

© 1993 American Statistical Association
Journal of the American Statistical Association
December 1993, Vol. 88, No. 424, Theory and Methods

Fisher with a series of papers culminating in his book *Statistical Methods for Research Workers* (1925), in which he created a new paradigm for hypothesis testing. He greatly extended the applicability of the t test (to the two-sample problem and the testing of regression coefficients) and generalized it to the testing of hypotheses in the analysis of variance. He advocated 5% as the standard level (with 1% as a more stringent alternative); through applying this new methodology to a variety of practical examples, he established it as a highly popular statistical approach for many fields of science.

A question that Fisher did not raise was the origin of his test statistics: Why these rather than some others? This is the question that Neyman and Pearson considered and which (after some preliminary work in Neyman and Pearson 1928) they later answered (Neyman and Pearson 1933a). Their solution involved not only the hypothesis but also a class of possible alternatives and the probabilities of two kinds of error: false rejection (Error I) and false acceptance (Error II). The "best" test was one that minimized P_A (Error II) subject to a bound on P_H (Error I), the latter being the significance level of the test. They completely solved this problem for the case of testing a simple (i.e., single distribution) hypothesis against a simple alternative by means of the Neyman–Pearson lemma. For more complex situations, the theory required additional concepts, and working out the details of this program was an important concern of mathematical statistics in the following decades.

The Neyman–Pearson introduction to the two kinds of error contained a brief statement that was to become the focus of much later debate. "Without hoping to know whether each separate hypothesis is true or false", the authors wrote, "we may search for rules to govern our behavior with regard to them, in following which we insure that, in the long run of experience, we shall not be too often wrong." And in this and the following paragraph they refer to a test (i.e., a rule to reject or accept the hypothesis) as "a rule of behavior".

3. INDUCTIVE INFERENCE VERSUS INDUCTIVE BEHAVIOR

Fisher considered statistics, the science of uncertain inference, able to provide a key to the long-debated problem of induction. He started one paper (Fisher 1932, p. 257) with the statement "Logicians have long distinguished two modes of human reasoning, under the respective names of deductive and inductive reasoning. . . . In inductive reasoning we attempt to argue from the particular, which is typically a body of observational material, to the general, which is typically a theory applicable to future experience." He developed his ideas in more detail in a later paper (Fisher 1935a, p. 39)

> . . . everyone who does habitually attempt the difficult task of making sense of figures is, in fact, essaying a logical process of the kind we call inductive, in that he is attempting to draw inferences from the particular to the general. Such inferences we recognize to be uncertain inferences. . . .

He continued in the next paragraph:

> Although some uncertain inferences can be rigorously expressed in terms of mathematical probability, it does not follow that mathematical probability is an adequate concept for the rigorous expression of uncertain inferences of every kind. . . . The inferences of the classical theory of probability are all deductive in character. They are statements about the behaviour of individuals, or samples, or sequences of samples, drawn from populations which are fully known. . . . More generally, however, a mathematical quantity of a different kind, which I have termed mathematical likelihood, appears to take its place [i.e., the place of probability] as a measure of rational belief when we are reasoning from the sample to the population.

Neyman did not believe in the need for a special inductive logic but felt that the usual processes of deductive thinking should suffice. More specifically, he had no use for Fisher's idea of likelihood. In his discussion of Fisher's 1935 paper (Neyman, 1935, p. 74, 75) he expressed the thought that it should be possible "to construct a theory of mathematical statistics . . . based solely upon the theory of probability," and went on to suggest that the basis for such a theory can be provided by "the conception of frequency of errors in judgment." This was the approach that he and Pearson had earlier described as "inductive behavior"; in the case of hypothesis testing, the behavior consisted of either rejecting the hypothesis or (provisionally) accepting it.

Both Neyman and Fisher considered the distinction between "inductive behavior" and "inductive inference" to lie at the center of their disagreement. In fact, in writing retrospectively about the dispute, Neyman (1961, p. 142) said that "the subject of the dispute may be symbolized by the opposing terms "inductive reasoning" and "inductive behavior." How strongly Fisher felt about this distinction is indicated by his statement in Fisher (1973, p. 7) that "there is something horrifying in the ideological movement represented by the doctrine that reasoning, properly speaking, cannot be applied to empirical data to lead to inferences valid in the real world."

4. FIXED LEVELS VERSUS p VALUES

A distinction frequently made between the approaches of Fisher and Neyman–Pearson is that in the latter the test is carried out at a fixed level, whereas the principal outcome of the former is the statement of a p value that may or may not be followed by a pronouncement concerning significance of the result.

The history of this distinction is curious. Throughout the 19th century, testing was carried out rather informally. It was roughly equivalent to calculating an (approximate) p value and rejecting the hypothesis if this value appeared to be sufficiently small. These early approximate methods required only a table of the normal distribution. With the advent of exact small-sample tests, tables of χ^2, t, F, . . . were also required. Fisher, in his 1925 book and later, greatly reduced the needed tabulations by providing tables not of the distributions themselves but of selected quantiles. (For an explanation of this very influential decision by Fisher see Kendall [1963]. On the other hand Cowles and Davis [1982] argue that conventional levels of three probable errors or two standard deviations, both roughly equivalent [in the normal case] to 5% were already in place before Fisher.) These tables allow the calculation only of ranges for the p values; however, they are exactly suited for determining the critical values at which the statistic under consideration becomes significant at a given level. As Fisher wrote in explaining the use of his χ^2 table (1946, p. 80):

> In preparing this table we have borne in mind that in practice we do not want to know the exact value of P for any observed χ^2, but, in the first place, whether or not the observed value is open to suspicion. If P is between .1 and .9, there is certainly no reason to suspect the hypothesis tested. If it is below .02, it is strongly indicated that the hypothesis fails to account for the whole of the facts. We shall not often be astray if we draw a conventional line at .05 and consider that higher values of χ^2 indicate a real discrepancy.

Similarly, he also wrote (1935, p. 13) that "it is usual and convenient for experimenters to take 5 percent as a standard level of significance, in the sense that they are prepared to ignore all results which fail to reach this standard . . ."

Fisher's views and those of some of his contemporaries are discussed in more detail by Hall and Selinger (1986).

Neyman and Pearson followed Fisher's adoption of a fixed level. In fact, Pearson (1962, p. 395) acknowledged that they were influenced by "[Fisher's] tables of 5 and 1% significance levels which lent themselves to the idea of choice, in advance of experiment, of the risk of the 'first kind of error' which the experimenter was prepared to take." He was even more outspoken in a letter to Neyman of April 28, 1978 (unpublished; in the Neyman collection of the Bancroft Library, University of California, Berkeley): "If there had not been these % tables available when you and I started work on testing statistical hypotheses in 1926, or when you were starting to talk on confidence intervals, say in 1928, how much more difficult it would have been for us! The concept of the control of 1st kind of error would not have come so readily nor your idea of following a rule of behaviour. . . . Anyway, you and I must be grateful for those two tables in the 1925 Statistical Methods for Research Workers." (For an idea of what the Neyman–Pearson theory might have looked like had it been based on p values instead of fixed levels, see Schweder 1988.)

It is interesting to note that unlike Fisher, Neyman and Pearson (1933a, p. 296) did not recommend a standard level but suggested that "how the balance [between the two kinds of error] should be struck must be left to the investigator," and (1933b, p. 497) "we attempt to adjust the balance between the risks P_{I} and P_{II} to meet the type of problem before us."

It is thus surprising that in SMSI Fisher (1973, p. 44–45) criticized the NP use of a fixed conventional level. He objected that

> 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. A man who 'rejects' a hypothesis provisionally, as a matter of habitual practice, when the significance is 1% or higher, will certainly be mistaken in not more than 1% of such decisions. . . . However, the calculation is absurdly academic, for in fact 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.

The difference between the reporting of a p value or that of a statement of acceptance or rejection of the hypothesis was linked by Fisher in Fisher (1973, pp. 79–80), to the distinction between drawing conclusions or making decisions.

> The conclusions drawn from such tests constitute the steps by which the research worker gains a better understanding of his experimental material, and of the problems which it presents. . . . More recently, indeed, a considerable body of doctrine has attempted to explain, or rather to reinterpret, these tests on the basis of quite a different model, namely as means to making decisions in an acceptance procedure.

Responding to earlier versions of these and related objections by Fisher to the Neyman–Pearson formulation, Pearson (1955, p. 206) admitted that the terms "acceptance" and "rejection" were perhaps unfortunately chosen, but of his joint work with Neyman he said that "from the start we shared Professor Fisher's view that in scientific inquiry, a statistical test is 'a means of learning' " and "I would agree that some of our wording may have been chosen inadequately, but I do not think that our position in some respects was or is so very different from that which Professor Fisher himself has now reached."

The distinctions under discussion are of course related to the argument about "inductive inference" vs. "inductive behavior," but in this debate Pearson refused to participate. He concludes his response to Fisher's 1955 attack with: "Professor Fisher's final criticism concerns the use of the term 'inductive behavior'; this is Professor Neyman's field rather than mine."

5. POWER

As was mentioned in Section 2, a central consideration of the Neyman–Pearson theory is that one must specify not only the hypothesis H but also the alternatives against which it is to be tested. In terms of the alternatives, one can then define the type II error (false acceptance) and the power of the test (the rejection probability as a function of the alternative). This idea is now fairly generally accepted for its importance in assessing the chance of detecting an effect (i.e., a departure from H) when it exists, determining the sample size required to raise this chance to an acceptable level, and providing a criterion on which to base the choice of an appropriate test.

Fisher never wavered in his strong opposition to these ideas. Following are some of his objections:

1. A type II error consists in falsely accepting H, and Fisher (1935b, p.) emphasized that there is no reason for "believing that a hypothesis has been proved to be true merely because it is not contradicted by the available facts." This is of course correct, but it does not diminish the usefulness of power calculations.

2. A second point Fisher raised is, in modern terminology, that the power cannot be calculated because it depends on the unknown alternative. For example (Fisher 1955, p. 73), he wrote:

> The frequency of the 1st class [type I error] . . . is calculable and therefore controllable simply from the specification of the null hypothesis. The frequency of the 2nd kind must depend . . . greatly on how closely they [rival hypotheses] resemble the null

> hypothesis. Such errors are therefore incalculable . . . merely from the specification of the null hypothesis, and would never have came into consideration in the theory only of tests of significance, had the logic of such tests not been confused with that of acceptance procedures. (He discussed the same point in Fisher 1947, p. 16–17.)

Fisher was of course aware of the importance of power, as is clear from the following remarks (1947, p. 24): "With respect to the refinements of technique, we have seen above that these contribute nothing to the validity of the experiment and of the test of significance by which we determine its result. They may, however, be important, and even essential, in permitting the phenomenon under test to manifest itself." The section in which this statement appears is tellingly entitled "Qualitative Methods of Increasing Sensitiveness." Fisher accepted the importance of the concept but denied the possibility of assessing it quantitatively.

Later in the same book Fisher made a very similar distinction regarding the choice of test. Under the heading "Multiplicity of Tests of the Same Hypothesis," he devoted a section (sec. 61) to this topic. Here again, without using the term, he referred to alternatives when he wrote (Fisher 1947, p. 182) that "we may now observe that the same data may contradict the hypothesis in any of a number of different ways." After illustrating how different tests would be appropriate for different alternatives, he continued (p. 185):

> The notion that different tests of significance are appropriate to test different features of the same null hypothesis presents no difficulty to workers engaged in practical experimentation but has been the occasion of much theoretical discussion among statisticians. The reason for this diversity of view-point is perhaps that the experimenter is thinking in terms of observational values, and is aware of what observational discrepancy it is which interests him, and which he thinks may be statistically significant, before he inquires what test of significance, if any, is available appropriate to his needs. He is, therefore, not usually concerned with the question: To what observational feature should a test of significance be applied?

The idea that there is no need for a theory of test choice, because an experienced experimenter knows what is the appropriate test, is expressed more strongly in a letter to W. E. Hick of October 1951 (Bennett 1990, p. 144), who, in asking about "one-tail" vs. "two-tail" in χ^2, had referred to his lack of knowledge concerning "the theory of critical regions, power, etc.":

> I am a little sorry that you have been worrying yourself at all with that unnecessarily portentous approach to tests of significance represented by the Neyman and Pearson critical regions, etc. In fact, I and my pupils throughout the world would never think of using them. If I am asked to give an explicit reason for this I should say that they approach the problem entirely from the wrong end, i.e., not from the point of view of a research worker, with a basis of well grounded knowledge on which a very fluctuating population of conjectures and incoherent observations is continually under examination. In these circumstances the experimenter does know what observation it is that attracts his attention. What he needs is a confident answer to the question "ought I to take any notice of that?" This question can, of course, and for refinement of thought should, be framed as "Is this particular hypothesis overthrown, and if so at what level of significance, by this particular body of observations?" It can be put in this form unequivocally only because the genuine experimenter already has the answers to all the questions that the followers of Neyman and Pearson attempt, I think vainly, to answer by merely mathematical consideration.

6. CONDITIONAL INFERENCE

While Fisher's approach to testing included no detailed consideration of power, the Neyman–Pearson approach failed to pay attention to an important concern raised by Fisher. To discuss this issue, we must begin by considering briefly the different meanings that Fisher and Neyman attach to probability.

For Neyman, the idea of probability is fairly straightforward: It represents an idealization of long-run frequency in a long sequence of repetitions under constant conditions (see, for example, Neyman 1952, p. 27; 1957, p. 9). Later (Neyman 1977), he pointed out that by the law of large numbers, this idea permits an extension: If a sequence of independent events is observed, each with probability p of success, then the long-run success frequency will be approximately p even if the events are not identical. This property adds greatly to the appeal and applicability of a frequentist probability. In particular, it is the way in which Neyman came to interpret the value of a significance level.

On the other hand, the meaning of probability is a problem with which Fisher grappled throughout his life. Not surprisingly, his views too underwent some changes. The concept at which he eventually arrived is much broader than Neyman's: "In a statement of probability, the predicand, which may be conceived as an object, as an event, or as a proposition, is asserted to be one of a set of a number, however large, of like entities of which a known proportion, P, have some relevant characteristic, not possessed by the remainder. It is further asserted that no subset of the entire set, having a different proportion, can be recognized" (Fisher 1973, p. 113). It is this last requirement, Fisher's version of the "requirement of total evidence" (Carnap 1962, sec. 45), which is particularly important to the present discussion.

Example 1 (Cox 1958). Suppose that we are concerned with the probability $P(X \leq x)$, where X is normally distributed as $N(\mu, 1)$ or $N(\mu, 4)$, depending on whether the spin of a fair coin results in heads (H) or tails (T). Here the set of cases in which the coin falls heads is a recognizable subset; therefore, Fisher would not admit the statement

$$P(X \leq x) = \frac{1}{2}\,\Phi(x - \mu) + \frac{1}{2}\,\Phi\left(\frac{x-\mu}{2}\right) \qquad (1)$$

as legitimate. Instead, he would have required $P(X \leq x)$ to be evaluated conditionally as

$$P(X \leq x \mid \mathrm{H}) = \Phi(x - \mu) \quad \text{or}$$

$$P(X \leq x \mid \mathrm{T}) = \Phi\left(\frac{x-\mu}{2}\right), \qquad (2)$$

depending on the outcome of the spin.

On the other hand, Neyman would have taken (1) to provide the natural assessment of $P(X \leq x)$. Despite this preference, there is nothing in the Neyman–Pearson (frequentist) approach to prevent consideration of the conditional probabilities (2). The critical issue from a frequentist viewpoint is what to consider as the relevant replications of the experiment: a sequence of observations from the same normal distribution or a sequence of coin tosses, each followed by an observation from the appropriate normal distribution.

Consider now the problem of testing H: $\mu = 0$ against the simple alternative $\mu = 1$ on the basis of a sample $X_1, \ldots, X_n$ from the distribution (1). The Neyman–Pearson lemma would tell us to reject H when

$$\frac{1}{2}\cdot\frac{1}{\sqrt{2\pi}}e^{-\Sigma(x_i-1)^2/2}+\frac{1}{2}\frac{1}{2\sqrt{2\pi}}e^{-\Sigma(x_i-1)^2/8} \geq K\left[\frac{1}{2}\frac{1}{\sqrt{2\pi}}e^{-\Sigma x_i^2/2}+\frac{1}{2}\frac{1}{2\sqrt{2\pi}}e^{-\Sigma x_i^2/8}\right], \quad (3)$$

where K is determined so that the probability of (3) when $\mu = 0$ is equal to the specified level α.

On the other hand, a Fisherian approach would adjust the test to whether the coin falls H or T and would use the rejection region

$$\frac{1}{\sqrt{2\pi}}e^{-\Sigma(x_i-1)^2/2} \geq K_1\frac{1}{\sqrt{2\pi}}e^{-\Sigma x_i^2/2} \quad \text{when the coin falls H} \quad (4)$$

and

$$\frac{1}{2\sqrt{2\pi}}e^{-\Sigma(x_i-1)^2/8} \geq K_2\frac{1}{2\sqrt{2\pi}}e^{-\Sigma x_i^2/8} \quad \text{when the coin falls T,} \quad (5)$$

where K_1 and K_2 are determined so that the null probability of both (4) and (5) is equal to α. It is easily seen that these two tests are not equivalent. Which one should we prefer?

Test (3) has the advantage of being more powerful in the sense that when the full experiment of spinning a coin and then taking n observations on X is repeated many times, and when $\mu = 1$, this test will reject the hypothesis more frequently.

The second test has the advantage that its conditional level given the outcome of the spin is α both when the outcome is H and when it is T. [The conditional level of the first test will be $<\alpha$ for one of the two outcomes and $>\alpha$ for the other.]

Which of these considerations is more important depends on the circumstances. Echoing Fisher, we might say that we prefer (1) in an acceptance sampling situation where interest focuses not on the individual cases but on the long-run frequency of errors, but that we would prefer the second test in a scientific situation where long-run considerations are irrelevant and only the circumstances at hand (i.e., H or T) matter. As Fisher put it (1973, p. 101–102), referring to a different but similar situation: "It is then obvious at the time that the judgment of significance has been decided not by the evidence of the sample, but by the throw of a coin. It is not obvious how the research worker is to be made to forget this circumstance, and it is certain that he ought not to forget it, if he is concerned to assess the weight only of objective observational facts against the hypothesis in question."

The present example is of course artificial, but the same issue arises whenever there exists an ancillary statistic (see, for example, Cox and Hinkley 1974; Lehmann 1986), and it seems to lie at the heart of the cases in which the two theories disagree on specific tests. The two most prominent of these cases are discussed in the next section.

7. TWO EXAMPLES

For many problems, a pure Fisherian or Neymann–Pearsonian approach will lead to the same test. Suppose in particular that the observations X follow a distribution from an exponential family with density

$$p_{\theta,\partial}(x) = C(\theta, \partial)e^{\theta U(x)+\Sigma_{i=1}^{k}\partial_i T_i(x)} \quad (6)$$

and consider testing the hypothesis

$$H\colon\ \theta = \theta_0 \quad (7)$$

against the one-sided alternatives $\theta > \theta_0$. Then Fisher would condition on $T = (T_1, \ldots, T_k)$ and would in the conditional model consider it natural to calculate the p value as the conditional probability of $U \geq u$, where u is the observed value of U. At a given level α, the result would be declared significant if $U \geq C(t)$, where $C(t)$ is determined by

$$P[U > C(t) \mid T = t] = \alpha. \quad (8)$$

A Neyman–Pearson viewpoint would lead to the same test as being uniformly most powerful among all similar tests.

But as we have seen in Example 1, the two approaches do not always lead to the same result. We next consider the two examples that have engendered the most controversy.

Example 2: The 2 × 2 table with one fixed margin. Let X, Y be two independent binomial variables with success probabilities p_1 and p_2 and corresponding to m and n trials. The problem of testing H: $p_2 = p_1$ against the alternatives $p_2 > p_1$ is of the form given by (6) and (7) with $\theta = \log[(p_2/q_2)/(p_1/q_1)]$, $T = X + Y$ and $U = Y$. Basically, there is therefore no conflict between the two approaches. However, because of the discreteness of the conditional distribution of U given t, condition (8) typically cannot be satisfied. Fisher's exact test then chooses $C(t)$ to be the largest constant for which

$$P[U > C(t) \mid T = t] \leq \alpha. \quad (9)$$

For small values of t, this may lead to conditional levels substantially less than α; for small m and n, the same may be true for the unconditional level. For this reason, Fisher's exact test has been criticized as being too conservative. Many alternatives have been proposed for which the unconditional level (which is a function of $p_1 = p_2$) is much closer to α. Upton (1982) lists 22; for other surveys, see Yates (1984) and Agresti (1992).

The issues are similar to those encountered in Example 1. If conditioning is considered appropriate (and in the present case it typically is), and if control of type I error at level α is considered essential, then the only sensible test available is of the form $U > C(t)$, where $C(t)$ is the largest value satisfying (9). If, on the other hand, only the unconditional performance is considered relevant, then we may allow the conditional level of the region $U > C(t)$ to exceed α for some values of t in such a way that the unconditional level (which is the expected value of the conditional level) gets closer to

α while remaining $\leq\alpha$ for all values of $p_2 = p_1$. This is essentially what the alternatives to Fisher's exact tests try to achieve. (The same issues arise also when analyzing 2×2 tables in which none of the margins are fixed.)

Example 3: Behrens–Fisher problem. Here we are dealing with independent samples $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ from normal distributions $N(\xi, \sigma^2)$ and $N(\eta, \tau^2)$ and we wish to test the hypothesis H: $\eta = \xi$. Against the two-sided alternatives $\eta \neq \xi$, there is general agreement that the rejection region should be of the form

$$\sqrt{\frac{|\bar{Y} - \bar{X}|}{\dfrac{S_X^2}{m} + \dfrac{S_Y^2}{n}}} \geq g\left(\frac{S_Y^2}{S_X^2}\right), \tag{10}$$

where S_Y^2 and S_X^2 are the usual estimators of σ^2 and τ^2.

Suppose that we consider it appropriate, as Fisher does, to carry out the analysis conditionally on the value of S_Y^2/S_X^2. If the conditional distribution of the left side of (10) given $S_Y^2/S_X^2 = c$ were independent of the parameters and hence known, there would be no problem. Everyone would agree to calculate g so that the conditional level is α for each c, which would then also result in an unconditional level identically equal to α. Unfortunately, the conditional distribution depends on the unknown variances. Two principal ways out of this difficulty have been proposed.

1. From a Neyman–Pearson point of view, the attempt has been to construct a function g for which the probability of (10) is $\equiv\alpha$ under H for all σ and τ (it actually depends only on the ratio $\theta = \tau^2/\sigma^2$). After much effort in this direction, it became clear that an acceptable function g satisfying this condition does not exist. But Welch and Aspin have produced tests whose level differs from α so little over the entire range of θ that, for all practical purposes, they can be viewed as solutions to the problem. (For a discussion and references see, for example, Stuart and Ord 1991, sec. 20.33.)

2. These tests are unacceptable to Fisher, however, because they admit recognizable subsets. In particular, Fisher (1956) produced an example for which the conditional level given $S_Y^2/S_X^2 = 1$ is always $>\alpha + \varepsilon$ for some positive ε. Fisher's own solution to the problem is the so-called Behrens–Fisher test, which he derived by means of a fiducial argument. Although it does not follow from this derivation, numerical evidence (Robinson 1976) strongly suggests that this test is conservative; that is, its unconditional level is always $<\alpha$. But a proof of this fact for all m, n, and θ is not yet available.

Let us call a set C in the sample space for which there exists $\varepsilon > 0$ such that

$$P_H[\text{rejecting} \mid X \in C] > \alpha + \varepsilon \quad \text{for all distributions in } H,$$

a liberally biased relevant subset. (The corresponding concept for confidence intervals is called negatively biased.) Robinson (1976) showed that no such subsets exist for the Behrens–Fisher test. (Because of this test's conservative nature, this is perhaps not too surprising.) He proposed calling a test conservative if its unconditional level is always $\leq\alpha$ and if it does not admit a liberally biased relevant subset, and expressed the hope that "perhaps the Behrens–Fisher test is optimal in some sense among the class of procedures which are conservative" (Robinson 1976, p. 970). This conjecture seems to have been disproved by Linssen (1991).

8. ONE THEORY OR TWO?

From the preceding sections it is clear that considerable differences exist between the viewpoints of Fisher and Neyman–Pearson. Are these sufficiently contradictory to preclude a unified theory that would combine the best features of both?

A first difference, discussed in Section 4, concerns the reporting of the conclusions of the analysis. Should this consist merely of a statement of significance or nonsignificance at a given level, or should a p value be reported? The original reason for fixed, standardized levels—unavailability of more detailed tables—no longer applies, and in any case reporting the p value provides more information. On the other hand, definite decisions or conclusions are often required. Additionally, in view of the enormously widespread use of testing at many different levels of sophistication, some statisticians (and journal editors) see an advantage in standardization; fortunately, this is a case where you can have your cake and eat it too. One should routinely report the p value and, where desired, combine this with a statement on significance at any stated level. (This was in fact common practice throughout the 19th Century and is the procedure frequently used by Fisher.) Two other principal differences, considered in Sections 5 and 6, are the omissions of power (by Fisher) and of conditioning (by Neyman–Pearson). It seems clear that a unified approach needs to incorporate both of these ideas.

For some problems this will cause no difficulty, because both approaches will lead to the same test, as illustrated at the beginning of Section 7. But the principles of conditioning on the one hand and of maximizing the unconditional power on the other may be in conflict, as is seen from Examples 1–3. This conflict disappears when it is realized that in such cases priority must be given to deciding on the appropriate frame of reference; that is, the real or hypothetical sequence of events that determine the meaning of any probability statement. Only after this has been settled do probabilistic concepts such as level and power acquire meaning, and it is only then that the problem of maximizing power comes into play.

This leaves the combined theory with its most difficult issue: What is the relevant frame of reference? It seems clear that even in the simplest situations (such as Ex. 1), no universal answer is possible. In any specific case, the solution will depend on contextual considerations that cannot easily be captured by a general theory.

That conflicting considerations argue for different solutions in specific cases is not an indictment of a theory, provided that the theory furnishes a basis for discussing the issues. Although Neyman and Pearson never seem to have raised the problem of just what constitutes a replication of an experiment, this question is as important for a frequentist as it is for an adherent of Fisherian probability. This was recognized, for example, by Bartlett (1984, p. 453), who stated "I regard the 'frequence requirement of repeated sampling' as including conditional inferences." A common basis for the discussion of various conditioning concepts, such as ancillaries and relevant subsets, thus exists. The proper choice of framework is a problem needing further study.

We conclude by considering some more detailed issues and by reviewing Examples 2 and 3 from the present point of view.

1. Both Neyman–Pearson and Fisher would give at most lukewarm support to standard significance levels such as 5% or 1%. Fisher, although originally recommending the use of such levels, later strongly attacked any standard choice. Neyman–Pearson, in their original formulation of 1933, recommended a balance between the two kinds of error (i.e., between level and power). For a disucssion of how to achieve such a balance, see, for example, Sanathanan (1974). Both level and power should of course be considered conditionally whenever conditioning is deemed appropriate. Unfortunately, this is not possible at the planning stage.

2. A second point on which there appears to be no conflict between the two approaches is "truth in advertising." Even if a particular nominal level α, say 5%, is the target, when it cannot be achieved because of discreteness the test should not just be described as conservative or liberal relative to the nominal level; instead, the actual (conditional or unconditional) level should be stated. If this level is not known because it depends on unknown parameters, at least its range should be given and, if possible, also an estimated value.

3. In both the 2 × 2 example and the Behrens–Fisher problems, the conflict between the solutions proposed by the two schools is often discussed as that of a desire for a similar test (i.e., one for which the unconditional level is $\equiv\alpha$) versus a suitable conditional test. The issue becomes clearer if one asks for the reason that Neyman–Pearson proposed the condition of similarity. The explanation begins with the case of a simple hypothesis where these authors take it for granted that in order to maximize the power, one would want the attained level to be equal to rather than less than α. For a composite hypothesis H, they therefore stated that the level should equal α for each of the simple hypotheses making up H. The requirement for similarity thus has its origin in the desire to maximize power, the issue discussed in Section 5.

In the light of (1) and (2), a unified theory less concerned with standard nominal levels might jettison not only the demand for similarity but also that of conservatism relative to a nominal level.

When similarity cannot be achieved and conservation is not required, various compromise solutions may be available. Thus in the 2 × 2 case of Example 2, one could, for example, select for each t the conditional level closest to α. If this seems too permissive, then the rule could be modified by adding a cap on the conditional level beyond which one would not go. Tests with a variable conditional level that will sometimes be $<\alpha$ and sometimes $>\alpha$ have been discussed by Barnard (1989) under the name "flexible Fisher." Alternatively, one might give up on a nominal level altogether and instead for each t adjust the level to the attainable (conditional) power.

The situation is much more complicated for the Behrens–Fisher problem. On the one hand, the arguments for conditioning an S_Y^2/S_X^2 seems less compelling; on the other hand, even if this conditioning requirement is accepted, the conditional distribution depends on unknown parameters, and thus it is less clear how to control the conditional level. Robinson's formulation, mentioned in Section 7, provides an interesting possibility but requires much further investigation. But such work can be carried out from the present point of view by combining considerations of both conditioning and power.

To summarize, p values, fixed-level significance statements, conditioning, and power considerations can be combined into a unified approach. When long-term power and conditioning are in conflict, specification of the appropriate frame of reference takes priority, because it determines the meaning of the probability statements. A fundamental gap in the theory is the lack of clear principles for selecting the appropriate framework. Additional work in this area will have to come to terms with the fact that the decision in any particular situation must be based not only on abstract principles but also on contextual aspects.

[Received January 1992. Revised February 1993.]

REFERENCES

Agresti, A. (1992), "A Survey of Exact Inference for Contingency Tables" (with discussion), *Statistical Science*, 7, 131–177.

Barnard, G. A. (1989), "On Alleged Gains in Power from Lower p Values," *Statistics in Medicine*, 8, 1469–1477.

Barnett, V. (1982), *Comparative Statistical Inference* (2nd ed.), New York: John Wiley.

Bartlett, M. S. (1984), "Discussion of 'Tests of Significance for 2 × 2 Contingency Tables,' by F. Yates." *Journal of the Royal Statistical Society*, Ser. A, 147, 453.

Bennett, J. H. (1990), *Statistical Inference and Analysis (Selected Correspondence of R. A. Fisher)*, Oxford, U.K.: Clarendon Press.

Braithwaite, R. B. (1953), *Scientific Explanation*, Cambridge, U.K.: Cambridge University Press.

Brown, L. (1967), "The Conditional Level of Student's t Test," *Annals of Mathematical Statistics*, 38, 1068–1071.

Carlson, R. (1976), "The Logic of Tests of Significance" (with discussion), *Philosophy of Science*, 43, 116–128.

Carnap, R. (1962), *Logical Foundations of Probability* (2nd ed.), Chicago: the University of Chicago Press.

Cowles, M., and Davis, C. (1982), "On the Origins of the .05 Level of Statistical Significance," *American Psychologist*, 37, 553–558.

Cox, D. R. (1958), "Some Problems Connected With Statistical Inference," *Annals of Mathematical Statistics*, 29, 357–372.

Cox, D. R., and Hinkley, D. V. (1974), *Theoretical Statistics*, London: Chapman and Hall.

Fisher, R. A. (1925) (10th ed., 1946), *Statistical Methods for Research Workers*, Edinburgh: Oliver & Boyd.

——— (1932), "Inverse Probability and the Use of Likelihood," *Proceedings of the Cambridge Philosophical Society*, 28, 257–261.

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

——— (1935b), "Statistical Tests," *Nature*, 136, 474.

——— (1935c) (4th ed., 1947), *The Design of Experiments*, Edinburgh: Oliver & Boyd.

——— (1939), "Student," *Annals of Engenics*, 9, 1–9.

——— (1947), *The Design of Experiments* (4th ed.), New York: Hafner Press.

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

——— (1956), "On a Test of Significance in Pearson's Biometrika Tables (No. 11), *Journal of the Royal Statistical Society*, Ser. B, 18, 56–60.

——— (1958), "The Nature of Probability," *Centennial Review*, 2, 261–274.

——— (1959), "Mathematical Probability in the Natural Sciences," *Technometrics,* 1, 21–29.

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

——— (1973), *Statistical Methods and Scientific Inference,* (3rd ed.) London: Collins Macmillan.

Gigerenzer, G., et al. (1989), *The Empire of Chance,* New York: Cambridge University Press.

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

Hall, P., and Selinger, B. (1986), "Statistical Significance: Balancing Evidence Against Doubt," *Australian Journal of Statistics,* 28, 354–370.

Hedges, L., and Olkin, I. (1985), *Statistical Methods for Meta-Analysis,* Orlando, FL: Academic Press.

Hockberg, Y., and Tamhane, A. C. (1987), *Multiple Comparison Procedures,* New York: John Wiley.

Kendall, M. G. (1963), "Ronald Aylmer Fisher, 1890–1962," *Biometrika,* 50, 1–15.

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

Linhart, H., and Zucchini, W. (1986), *Model Selection,* New York: John Wiley.

Linssen, H. N. (1991), "A Table for Solving the Behrens–Fisher Problem," *Statistics and Probability Letters,* 11, 359–363.

Miller, R. G. (1981), *Simultaneous Statistical Inference,* (2nd ed.), New York: Springer-Verlag.

Morrison, D. E., and Henkel, R. E. (1970), *The Significance Test Controversy,* Chicago: Aldine.

Neyman, J. (1935), "Discussion of Fisher (1935a)." *Journal of the Royal Statistical Society,* 98, 74–75.

——— (1938), "L'Estimation Statistique Traitée Comme un Problème Classique de Probabilité," *Actualités Scientifiques et Industrielles,* 739, 25–57.

——— (1952), *Lectures and Conferences on Mathematical Statistics and Probability* (2nd ed.), Graduate School, Washington, D.C.: U.S. Dept. of Agriculture.

——— (1955), "The Problem of Inductive Inference," *Communications in Pure and Applied Mathematics,* 8, 13–46.

——— (1956), "Note on an Article by Sir Ronald Fisher," *Journal of the Royal Statistical Society,* Ser. B, 18, 288–294.

——— (1957), " 'Inductive behavior' as a Basic Concept of Philosophy of Science," *Review of the International Statistical Institute,* 25, 7–22.

——— (1961), "Silver Jubilee of My Dispute With Fisher," *Journal of the Operations Research Society of Japan,* 3, 145–154.

——— (1966), "Behavioristic Points of View on Mathematical Statistics," in *On Political Economy and Econometrics: Essays in Honour of Oscar Lange,* Warsaw: Polish Scientific Publishers, pp. 445–462.

——— (1976), "Tests of Statistical Hypotheses and Their Use in Studies of Natural Phenomena," *Communications in Statistics, Part A—Theory and Methods,* 5, 737–751.

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

Neyman, J., and Pearson, E. S. (1928), "On the Use and Interpretation of Certain Test Criteria for Purposes of Statistical Inference," *Biometrika,* 20A, 175–240, 263–294.

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

——— (1933b), "The Testing of Statistical Hypotheses in Relation to Probabilities A Priori," *Proceedings of the Cambridge Philosophical Society,* 29, 492–510.

Oakes, M. (1986), *Statistical Inference: A Comment for the Social and Behavioral Sciences,* New York: John Wiley.

Pearson, E. S. (1955), "Statistical Concepts in Their Relation to Reality," *Journal of the Royal Statistical Society,* Ser. B, 17, 204–207.

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

——— (1974), "Memories of the Impact of Fisher's Work in the 1920's," *International Statistical Review,* 42, 5–8.

Pearson, E. S., and Hartley, H. O. (1954), *Biometrika Tables for Statisticians (Table No. 11),* New York: Cambridge University Press.

Pedersen, J. G. (1978), "Fiducial Inference," *International Statistical Review,* 46, 147–170.

Robinson, G. K. (1976), "Properties of Student's *t* and of the Behrens–Fisher Solution to the Two-Means Problem," *The Annals of Statistics,* 4, 963–971.

——— (1982), "Behrens–Fisher Problem," in Encyclopedia of Statistical Sciences (Vol. 1, eds. S. Kotz and N. L. Johnson), New York: John Wiley, pp. 205–209.

Savage, L. J. (1976), "On Rereading R. A. Fisher" (with discussion), *Annals of Statistics,* 4, 441–500.

Schweder, T. (1988), "A Significance Version of the Basic Neyman–Pearson Theory for Scientific Hypothesis Testing," *Scandanavian Journal of Statistics,* 15, 225–242.

Seidenfeld, T. (1979), *Philosophical Problems of Statistical Inference,* Boston: D. Reidel.

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

——— (1978), "Statistical Dogma and the Logic of Significance Testing," *Philosophy of Science,* 45, 120–135.

Steger, J. A. (ed.) (1971), *Readings in Statistics for the Behavioral Scientist,* New York: Holt, Rinehart and Winston.

Stuart, A., and Ord, J. K. (1991), *Kendall's Advanced Theory of Statistics, Vol. II* (5th ed.), New York: Oxford University Press.

Tukey, J. W. (1960), "Conclusions vs. Decisions," *Technometrics,* 2, 424–432.

Upton, G. J. G. (1982), "A Comparison of Alternative Tests for the 2 × 2 Comparative Trial," *Journal of the Royal Statistical Society,* Ser. A, 145, 86–105.

Wallace, D. L. (1980), "The Behrens–Fisher and Fieller–Creasy Problems," in *R. A. Fisher: An Application,* eds. S. E. Fienberg and D. V. Hinkley, New York: Springer-Verlag, pp. 119–147.

Yates, F. (1984), "Tests of Significance for 2 × 2 Contingency Tables" (with discussion), *Journal of the Royal Statistical Society,* Ser. A, 147, 426–463.

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