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

By

E.L. Lehmann

Technical Report No. 333
January 1992

Research supported by NSF Grant No. DMS-8908670.

Department of Statistics
University of California
Berkeley, California 94720

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

By E.L. Lehmann
Univ. of Calif., Berkeley

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.

* Research supported by NSF Grant No. DMS-8908670.

1. **Introduction.**

The formulation and philosophy of hypothesis testing as we know it today was largely created by three men: R.A. Fisher (1890-1962), J. Neyman (1894-1981), and E.S. Pearson (1895-1980) in the period 1915-1933. 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's and Hypothesis testing for that of Neyman and Pearson.* But are they really that different?

It is interesting to see what Fisher, Neyman, and Pearson themselves have to say about this question. Fisher frequently attacked the Neyman-Pearson (NP) approach as completely inappropriate to the testing of scientific hypotheses (although perhaps suitable in the context of acceptance sampling). In his last book ''Statistical Methods and Scientific Inference'' (3rd ed., published posthumously in 1973, to which we shall refer as SMSI), he writes (p.103):

''The examples elaborated in the foregoing sections of numerical discrepancies... constitute only one aspect of the deep-seated difference in point of view ...''

On the other hand, Neyman (1976) stated that he ''is not aware of a conceptual difference between a 'test of a statistical hypothesis' and a 'test of significance' and [that he] uses these terms interchangeably''.

Pearson (1974) took an intermediate position by acknowledging the existence of differences but claiming that they were of little importance in practice. After referring to inference as ''the manner in which we bring the theory of probability into gear with the way our mind works in reaching decisions and practical conclusions'', he continues: ''If, as undoubtedly seems the case, the *same* mechanism of this 'putting into gear operation' does not work for everyone in identical ways, this does not seem to matter''.

In the present paper, written just ten years after the death of the last protagonist, I examine yet another possibility: that important differences do exist but that it may be possible to formulate a unified theory that combines the best features of both approaches.

* 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.

For the sake of completeness it should be said that in addition to the Fisher and Neyman-Pearson theories there exist still 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 Morrison and Henkel (1970), Steger (1971), Spielman (1974, 1978), Carlson (1976), Barnett (1982)) 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 easily be split off from other aspects of his work, and we shall here not consider either the fiducial or the Bayesian approach any further.

It seems appropriate to conclude this introduction with two personal statements.

(i) I was a student of Neyman's and later for many years his colleague. As a result I am fairly familiar with his thinking. On the other hand, I have seriously studied Fisher's work only in recent years and, perhaps partly for this reason, have found his ideas much harder to understand. I shall therefore try to follow Fisher's advice to a correspondent (Bennett, 1990, p.221):

''If you must write about someone else's work it is, I feel sure, worth taking even more than a little trouble to avoid misrepresenting him. One safeguard is to use actual quotations from his writing;''

(ii) 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 the present paper is written from a statistical, not a philosophical, point of view.

* 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.

2. Testing Statistical Hypotheses

The modern theory of testing hypotheses began with Student's discovery of the t-distribution in 1908. This was followed by 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 teting of hypotheses in the analysis of variance. He advocated 5% as the standard level (with 1% as a more stringent alternative); and 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 NP (1928)) they answered in NP (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 NP-program was an important concern of mathematical statistics in the following decades.

The NP 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 vs. inductive behavior

Fisher (1932) started a paper entitled "Inverse probability and the use of likelihood" 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 1935 paper, "The logic of inductive inference" where he explains:

"... 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 continues 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".

The paper was presented at a meeting of the Royal Statistical Society and was not well received. The last discussant was Neyman who began in a very complimentary vein. He then suggested that some readers might react by thinking: "What an interesting problem is raised! How could I develop it further?, but, he continues "I personally seem to have another kind of psychology and can't help thinking: What an interesting way of asking and answering questions, but can't I do it differently?"

More specifically Neyman asks "granted that the conception of likelihood is independent of the classical theory of probability, isn't it possible to construct a theory of mathematical statistics which would be based soley upon the theory of probability (thus independent of the conception of likelihood) and be adequate from the point of view of practical statistical work?"

And later, still more directly: "Now what could be considered as a sufficiently simple and unquestionable principle in statistical work? I think the basic conception here is the conception of frequency of errors in judgement." He points out that this idea applies to both hypothesis testing and estimation and completes the paragraph with the statement that "the complex of results in this direction may be considered as a system of mathematical statistics alternative to that of Professor Fisher, and entirely based on the classical theory of probability."

Of Fisher, L.J. Savage (1976) in his insightful overview of Fisher's great accomplishments "On rereading R.A. Fisher" wrote: "Fisher burned even more than the rest of us, it seems to me, to be original, right, important, famous, and respected." One can then imagine Fisher's reaction to this attack on his cherished and ambitious attempt to put scientific thinking on an entirely new basis.

Neyman's message was: We have no need for your inductive inference and its new concept of likelihood. The problem can be solved in a very satisfactory manner using only the classical theory of probability and deductive arguments, by minimizing the probability of errors [i.e. of wrong conclusions].

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) said that ''the subject of the dispute may be symbolized by the opposing terms ''inductive reasoning'' and ''inductive behavior''. That Fisher also assigned a central role to this distinction is indicated by his statement in SMSI (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.''

Actually, the interpretation of acceptance or rejection of a hypothesis as behavior or inference, as a decision or conclusion, is largely a matter of terminology which diverts attention from the more central issue: whether only deductive arguments are needed to reach the desired end, or whether there is a need also for induction, a concept which inspired Fisher while for Neyman it was imbued with an aura of suspect mysticism. This issue had in fact a long history and a resolution (albeit in a deterministic rather than a stochastic setting) in the description of the scientific method as ''hypothetico-deductive''. According to this view of science, induction is required in deciding on the experiment to be performed, in the formulation of the model and the hypothesis, while the testing of the model and the hypothesis can be carried out deductively.

Surprisingly, Fisher himself seemed to view the situation somewhat similarly, when in his 1939 obituary of Student he wrote: ''Many mathematicians must possess the penetration necessary to perceive, when confronted with concrete experimental results, that it must be possible to use them, by rigorously objective calculations, to throw light on the plausibility or otherwise of the interpretations that suggest themselves. A few must also possess the pertinacity needed to convert this intuition into such a completed procedure as we know as a test of significance. It is, I believe, nothing but an illusion to think that this process can ever be reduced to a self-contained mathematical theory of tests of significance. Constructive imagination, together with much knowledge based on experience of data of the same kind, must be exercised before deciding on what hypotheses are worth testing, and in what respects. Only when this fundamental thinking has been accomplished can the problem be given a mathematical form.''

4. Errors of the Second Kind

Fisher did not respond immediately to the attack Neyman had mounted in his discussion of Fisher's paper. However, in a note in Nature (1935b) which was ostensibly a reply to an only trangentially related statement by Karl Pearson, be lashed out at Neyman and E.S. Pearson without however mentioning their names. Karl Pearson, in a letter to Nature had complained that his χ^2-test of goodness of fit was not a rule for deciding whether or not to reject a possibly correct hypothesis, but rather an attempt to see whether a distribution, although not expected to be exactly correct, would provide an adequate fit. After a relatively mild rejection of this position, Fisher adds in a last paragraph: ''For the logical fallacy of believing that a hypothesis has been proved to be true, merely because it is not contradicted by the available facts, has no more right to insinuate itself in statistical than in other kinds of scientific reasoning. Yet it does so only too frequently. Indeed, the ''error of accepting an hypothesis when it is false'' has been specially named by some writers ''errors of the second kind''. It would, therefore, add greatly to the clarity with which the tests of significance are regarded if it were generally understood that tests of significance, when used accurately, are capable of rejecting or invalidating hypotheses, in so far as these are contradicted by the data; but that they are never capable of establishing them as certainly true. In fact that ''errors of the second kind'' are committed only by those who misunderstand the nature and application of tests of significance.''

After this outburst, the dispute appeared to die down. Undoubtedly it helped that in 1938 Neyman, the more combative of Fisher's opponents, left London for Berkeley, thereby removing the irritation of offices in the same building and frequent encounters at meetings of the Royal Statistical Society. Then, twenty years after the Nature article, Fisher (1955) published a paper devoted entirely to an attack on the point of view ''expressed in numerous papers by Neyman, Pearson, Wald and Bartlett''.

The first introductory sections suggest two reasons for Fisher's writing such a paper at that time. He begins by describing the progress that had been made during the present century ''in the business of interpreting observational data, so as to obtain a better understanding of the real world.'' He mentions in particular ''the use of better mathematics and more comprehensive ideas in mathematical statistics''; ''the new theory of experimental design''; and ''a more complete understanding. .. of the structure and peculiarities of inductive logic''.

''Much that I have to say''; Fisher continues, ''will not command universal assent. I know this for it is just because I find myself in disagreement with some of the modes of exposition of this new subject which have from time to time been adopted, that I have taken this opportunity of expressing a different point of view''.

What Fisher was referring to are developments that had occurred since the publication of his early papers and the two books, "Statistical Methods for Research Washers" (1925) and "The Design of Experiments" (1935c). His methods had been enormously successful; his tests, the analysis of variance, the experimental designs had become the staple of working statisticians. His books had reached a wide public. (By 1946, Statistical Methods had reach the 10th Edition) but — and this must have been tremendously galling to him — his philosophical approach had not found acceptance. On the one hand, his central concept of fiducial inference had found few adherents; on the other, perhaps even more annoying, developments growing out of Neyman's philosophy had been grafted onto his framework and were highly successful. There had been considerable elaboration of the NP theory of optimal tests; more importantly, the idea of power (1 − P(Error II)) was generally accepted as a concept of interest in itself and as the basis for sample size determination; and finally Neyman's philosophy of inductive behavior had been formalized by Wald into a comprehensive theory of Statistical Decision Functions.

An additional stimulus for Fisher's paper appears to have been a suggestion by George Barnard which Fisher acknowledges in a letter of Feb. 9, 1954: (Bennett (1990, p.9) "I find, looking up the old papers, that I can now understand, much better than before, the early work of Neyman, or Neyman and Pearson, in the light of what you said the other afternoon, for it now seems clear to me, as it did not before, that Neyman, thinking all the time of acceptance procedures, was under the misapprehension that my own work on estimation had only the same end in view".

Fisher accepts in the introduction to his 1955 paper that "there is no difference to matter in the field of mathematical analysis [i.e. typically the different approaches lead to essentially the same methods]... but, he says, "there is a clear difference in logical point of view". He then acknowledges his debt to Barnard and strikes a theme which will be dominant in his discussion of these issues from now on: "I owe to Professor Barnard .. the penetrating observation that this difference in point of view originated when Neyman, thinking that he was correcting and improving my own early work on tests of significance, as a means to 'the improvement of natural knowledge', in fact reinterpreted them in terms of that technological and commercial apparatus which is known as an acceptance procedure".

With this remark, Fisher cedes to Neyman's idea of inductive behavior the lower spheres of technology and commerce, while reserving his own deeper, more difficult, and hence less understood and accepted, idea of inductive inference for scientific work. One must admit that the NP terms behavior, error, acceptance, and rejection, and their extension by Wald to decision and loss function, encourage such an interpretation.

More specifically, Fisher's attack in the paper under discussion concentrated on three targets: repeated sampling from the same population; errors of the second kind' and 'inductive behavior'. Neyman replied in the following year with a "Note on an article by Sir Ronald Fisher". The year 1956 also saw the publication of Fisher's last book (SMSI), which sets out once more in full his own position and his criticism of the opposing view, and the next year Neyman followed with a paper, " 'Inductive behavior' as a basic concept of philosophy of science". The exchange ended with a last flurry: A paper by Fisher (1960) entitled "Scientific thought and the refinement of human reason" and Neyman's reply the following year: "Silver Jubilee of my dispute with Fisher".

It is tempting to quote some of the interesting and colorful statements that can be found in these publications, but in fact not much new ground was covered. At the end of his life Fisher continued to feel strongly that the ideas conveyed by the terms rules of behavior and its long-run consequences, particularly errors of the second kind, had no place in scientific inference.

5. Conditional inference

While Fisher's approach to testing included no consideration of power, the NP approach failed to pay attention to an important concern raised by Fisher. In order to discuss this issue we must begin by considering briefly the different meanings 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) and Neyman (1957, p.9). Later (Neyman (1977)), he points out that by the law of large numbers this idea permits an extension: that 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 greatly adds 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 and, 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", he says on p.113 of SMSI, "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.

It is this last requirement which is particularly important to the present discussion. As an example, suppose that we are concerned with the probability $P(X \leq 1)$ 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 and therefore Fisher would not admit the statement

$$P(X \leq x) = \frac{1}{2}\Phi(x-\mu) + \frac{1}{2}\Phi(\frac{x-\mu}{2}) \tag{5.1}$$

as legitimate. On the other hand, Neyman would have no trouble with (5.1), although he might also be interested in the conditional statements

$$P(X \leq x | H) = \Phi(x-\mu) \text{ and } P(X \leq x | T) = \Phi(\frac{x-\mu}{2}). \tag{5.2}$$

From a frequentist point of view, the critical issue is whether we 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 (5.1). The Neyman-Pearson lemma would tell us to reject H when

$$\begin{aligned} &\frac{1}{2} \cdot \frac{1}{\sqrt{2\pi}} e^{-\Sigma(x_i - 1)^2/2} + \frac{1}{2}\frac{1}{2\sqrt{2\pi}} e^{-\Sigma(x_i - 1)^2/8} \\ &\geq K[\frac{1}{2}\frac{1}{\sqrt{2\pi}} e^{-\Sigma x_i^2/2} + \frac{1}{2}\frac{1}{2\sqrt{2\pi}} e^{-\Sigma x_i^2/8}] \end{aligned} \tag{5.3}$$

where K is determined so that the probability of (5.3) 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

$$\frac{1}{\sqrt{2\pi}} e^{-\Sigma(x_i-1)^2/2} \geq K_1 \frac{1}{\sqrt{2\pi}} e^{-\Sigma x_i^2/2} \text{ when the coin falls H} \tag{5.4}$$

and

$$\frac{1}{2\sqrt{2\pi}} e^{-\Sigma(x_i-1)^2/8} \geq K_2 \frac{1}{2\sqrt{2\pi}} e^{-\Sigma x_i^2/8} \text{ when the coin falls T}. \tag{5.5}$$

It is easily seen that these two tests are not equivalent. Which one should we prefer?

The test (5.3) 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 (5.1) in an acceptance sampling situation where interest focuses not on the individual cases but on the long-run frequency of errors; however, 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 (SMSI, p.102) referring to a different but similar situation: "... it is then obvious at the time that the judgement 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) and Lehmann (1986)), and it seems to lie at the heart of the cases in which the two theories disagree on specific tests. The most widely discussed example of this kind is the Behrens-Fisher problem, in which one is dealing with independent samples from normal distributions $N(\xi, \sigma^2)$ and $N(\eta, \tau^2)$ respectively and wishes to test the hypothesis $H: \eta = \xi$.

It turns out that no sensible test exists for which the rejection probability under the hypothesis is exactly α for all values of $\eta = \xi$, σ and τ. However, except for very small sample sizes the Welch and Welch-Aspin tests satisfy this condition to a very close approximation and constitute practically satisfactory solutions from a Neyman-Pearson point of view. Fisher's objection to the Welch-Aspin Test is based on his belief that the relevant level is the conditional probability of rejection given the ratio of the sample variances S_Y^2 / S_X^2. In Fisher (1956a) he shows for a particular example with sample sizes $n_1 = n_2 = 7$ that the conditional level given $S_Y^2 / S_X^2 = 1$ exceeds the nominal level for all values of the ratio τ^2 / σ^2 of the true variances. The set $S_Y^2 / S_X^2 = 1$ is thus what Fisher calls "a recognizable subset".

Why does Fisher feel so strongly that the conditional rather than the unconditional level is appropriate? In a letter written in 1955 to Yates (Bennett, 1990, p.243) he states categorically: "From my point of view this "fixing" [of S_Y^2 / S_X^2] is not a voluntary act to be done or abstained from, but a fact to be recognized...", and in SMSI (p.96) he comments regarding the table for the Aspin-Welch test published by Pearson and Hartley (1954) that it is "indeed misleading in just this way. The editors' justification is: "It is the result of an attempt to produce a test which *does* satisfy the condition that the probability of the rejection of the hypothesis tested will be equal to the specified figure. The logical basis of a test of significance as a means of learning from experimental data is here completely overlooked; for the potential user has no warning of the concealed laxity of the test.''

Fisher's own solution to the two-means problem is the test now known as the Behrens-Fisher test, which he derives by means of a fiducial argument. From a Neyman-Pearson point of view this test suffer from the defect that its rejection probability depends on τ^2/σ^2. Since the test appears to be conservative i.e. its level always to be less than the nominal level, its power is thus less than it could be. (For a more detailed comparison of these tests see for example Robinson (1976 and 1982), Pedersen (1978), and Wallace (1980).)

6. Fixed levels vs p-values

A distinction that is frequently made between the approaches of Fisher and Neyman-Pearson is that in the latter the test is carried out at a fixed level while 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 to 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.* These allow the calculation only of ranges for the p-values. However, they are exactly suited to determining the critical values at which the statistic under consideration becomes significant at a given level. As Fisher writes** in explaining the use of his χ^2-table (10th Ed., 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

* 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.

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

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 writes in his next book (1935, p.13)

''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,...''

Neyman and Pearson follow Fisher's adoption of a fixed level. In fact, Pearson (1962) acknowledges 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 NP 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) do not recommend a standard level but suggest that ''how the balance [between the two kinds of error] should be struck must be left to the investigator'' and again in (1933b):'' 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 (p.44/45) Fisher criticizes the NP use of a fixed conventional level. He objects 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, be 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 is linked by Fisher to the distinction between drawing conclusions or making decisions.* On p.79/80 of SMSI he writes:

"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 NP formulation, E.S. Pearson (1955) admits that the terms "acceptance" and "rejection" were perhaps unfortunately chosen but of his joint work with Neyman he writes: "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 refuses to participate. The last sentence in his response to Fisher's 1955 attack is; "Professor Fisher's final criticism concerns the use of the term "inductive behaviour; this is Professor Neyman's field rather than mine."

7. One theory or two?

It is clear from the preceding sections that there are considerable differenes between the two approaches. Are they sufficiently contradictory to preclude a unified methodology which would combine the best features of both?

Two of the most important differences, discussed in Sections 3 and 4 are omissions. They pose no obstacles to a unified theory which would include consideration of power and sample size on the one hand, and the appropriateness of conditioning on the other. (Note however that the theory of conditioning is still incomplete and requires further work. For an account of the present status see Lehmann (1986, Chapter 10).

* For further discussion of this distinction see Tukey (1960).

A third difference, which has been much emphasized, is whether to report a p-value or to decide on acceptance or rejection at a fixed level. The original reason for fixed, standardized levels - unavailability of suitable tables - no longer applies. On the other hand, with the enormously widespread use of testing at many different levels of sophistication, some statisticians (and journal editors) see an advantage in standardization. In any case, reporting the p-value (possibly combined with a statement of significant at a conventional level) provides additional information and should be done routinely (but see the next section for an alternative approach).

By utilizing from each theory what is missing in the other, a unified approach to testing thus appears surprisingly easy to achieve. While central differences of interpretation remain (particularly with respect to the meaning of probability), there turn out to be no serious practical inconsistencies that would present major obstacles. The impression of irreconcilable differences was greatly exaggerated on the one hand by the basic difference in philosophy: between Neyman's pragmatic efforts of reducing the process of science to deductive reasoning as a guide to appropriate behavior, opposed by Fisher's more ambitious aspirations toward a deeper understanding of the process of human reasoning involved in scientific progress; and on the other hand by the fierce and insistent rhetoric that fueled the controversy.

In hindsight we can see that Fisher on the one side and Neyman-Pearson on the other shared certain basic assumptions that were perhaps more fundamental than the differences that divided them. To begin with, they both formulated their theories in terms of parametric models (initiated by Fisher (1922)) and they were clear about the fact that these were only mathematical representations of a much more complex reality. A strong motivation for both Fisher and NP was the desire to create a theory that was independent of any assumed prior distribution over the possible hypotheses or parameter values. (Since Laplace, the assumption of a uniform prior had been the much debated dominant approach, although calculations of p-values without such an assumption were also performed). Finally it seems safe to say that all the contestants would have deplored a purely data-analytic approach without any probability calculations.

This common basis of clearly defined models, with considerations of power as well as significance, and with the inference restricted by appropriate conditioning where suitable, seems to provide an attractive framework for hypothesis testing that combines some of the most important ideas from both schools.

8. Beyond hypothesis testing

Hypothesis testing, along the lines discussed here, is frequently derided (see for example Savage (1976, p.473)) but nevertheless often encountered in practice. It refers to situations in which a single experiment is to be used to test a hypothesis, with both model and hypothesis clearly specified before the observations are taken, at a level which is also prespecified.

On the other hand, there are many situations that do not fit this description. One set of difficulties arises when the hypothesis is suggested by the data. Such "data snooping", and the related problem of multiplicity stemming from the deliberate consideration of several hypotheses simultaneously, have given rise to a body of techniques of multiple comparisons and simultaneous inference (see for example the books by Miller (1981) and Hochberg and Tamhane (1987). Furthermore, not only the hypothesis but also the model may have been chosen in light of the data. This problem has not yet been adequately explored although model selection techniques provide a start. (See for example Linhart and Zucchini (1986).)

Another assumption mentioned at the beginning of the section is that the hypothesis is to be tested on the basis of a single experiment. However, most scientific studies are carried out over a period of time and involve replication of the experiment by the same or other investigators, possibly under a range of different conditions. It may then be important to combine information from the separate experiments into an overall assessment of the combined results. A survey of methods developed for this purpose can be found for example in Hedges and Olkin (1985).

The statistical procedures required for the extensions mentioned so far in this section can be discussed in terms of the general approach described in Section 7. We shall now conclude by a somewhat different approach that is often preferred. As was discussed earlier, reporting not only the conclusion (reject or accept) but also the p-value provides an important strengthening of the fixed level Neyman-Pearson testing procedure. An alternative strengthening often is available when a clearly defined class of alternatives has been formulated. To be specific, suppose the hypothesis specifies the value of a parameter to be θ. Then it was pointed out by Neyman (1937) that consideration of the class of hypotheses

$$H(\theta): \text{the parameter has the value } \theta,$$

tested for varying θ at a common level α on the basis of observations X by means of acceptance regions $A(\theta)$, leads to confidence sets $S(x)$ through the relation

$$\theta \in S(x) \text{ if and only if } x \in A(\theta). \tag{7.1}$$

Here x denotes the observed value of X and the sets $S(X)$ have the property that they cover the unknown true θ with probability $\gamma = 1 - \alpha$, the confidence coefficient. Reporting of the confidence set $S(x)$ provides solutions both to the problem of estimating θ and that of testing the original hypothesis $H: \theta = \theta_0$ by means of the acceptance region $A(\theta_0)$ obtained from (7.1) (i.e. H is accepted if and only if $S(x)$ covers the point θ_0.)

This approach is not restricted to parametric situations. Suppose for example that $X_1, \ldots, X_n$ are iid according to an unknown continuous cdf F, and consider the hypothesis H that F is equal to a specified distribution F_0. Then confidence bands for F can be based, for example on the Kolmogorov statistic and provide solutions both of the problem of estimating F and of testing H.

As was the case for hypothesis testing there is a rival theory of estimation due to Fisher, his fiducial theory which was central to Fisher's statistical thinking. However, a discussion of this approach and its comparison with Neyman's go beyond the purpose of the present paper.

References

Barnett, V. (1982). Comparative Statistical Inference (2nd Ed). Wiley. New York.

Bennett, J.H. (199). Statistical Inference and Analysis (Selected Correspondence of R.A. Fisher.) Clarendon Press. Oxford.

Braithwaite, R.B. (1953). Scientific Explanation. Cambridge University Press.

Brown, L. (1967). The conditional level of Student's t-test. *Ann. Math. Statist.* **38,** 1068-1071.

Carlson, R. (1976). The logic of tests of significance (discussion). *Phil. of Sci.* **43,** 116-128.

Cowles, M. and Davis, C. (1982). On the origins of the .05 level of statistical significance. *Amer. Psychologist* **37,** 553-558.

Cox, D.R. and Hinkley, D.V. (1974). Theoretical statistics.

Fisher, R.A. (1925); (10th Ed., 1946). Statistical Methods for Research Workers. Oliver and Boyd. Edinburgh

Fisher, R.A. (1932). Inverse probability and the use of likelihood. *Proc. Cambridge Phil. Soc.* **28,** 257-261.

Fisher, R.A. (1935a). The logic of inductive inference. *J. Roy. Statist. Soc.* **98,** 39-54.

Fisher, R.A. (1935b). Statistical tests. *Nature* **136,** 474.

Fisher, R.A. (1935c). (4th Ed. 1947). The Design of Experiments. Olivers and Boyd. Edinburh.

Fisher, R.A. (1939). 'Student'. *Am. Engenics* **9,** 1-9.

Fisher, R.A. (1955). Statistical methods and scientific induction. *J. Roy. Statist. Soc. (B)* **17,** 69-78.

Fisher, R.A. (1956). On a test of significance in Pearson's Biometrika Tables (No.11). *J. Roy. Statist. Soc (B)* **18,** 56-60.

Fisher, R.A. (1956). (3rd Ed, 1973) Statistical Methods and Scientific Inference. Collins Macmillan London.

Fisher, R.A. (1958). The nature of probability, *Centennial Rev..* **2,** 261-274.

Fisher, R.A. (1959). Mathematical probability in the natural sciences. *Technometrics* **1,** 21-29.

Fisher, R.A. (1960). Scientific thought and the refinement of human reason. *J. Oper. Res. Soc. of Japan* **3,** 1-10.

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

Hall, P. and Selinger, B. (1986). Statistical significance: Balancing evidence against doubt. *Austral. J. Statist.* **28**, 354-370.

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

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

Kendall, M.G. (1963). Rowald Aylmer Fisher, 1890-1962. *Biometrika* **50**, 1-15.

Kyburg, H.E., Jr. (1974). The Logical Foundations of Statistical Inference. D. Reidel Publ. Co.

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

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

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

Neyman, J. (1938). L'Estimation Statistique traitée comme un problème classique de probabilité. *Actual. Sci. et Industr.* **739**, 25-57.

Neyman, J. (1952). Lectures and Conferences on Mathematical Statistics and Probability (2nd Ed). Graduate School, U.S. Dept. of Agriculture, Washington.

Neyman, J. (1955). The problem of inductive inference. *Commun. Pure and Applied Math.* **8**, 13-46.

Neyman, J. (1956). Note on an article by Sir Ronald Fisher. *J. Roy. Statist. Soc. (B)* **18**, 288-294.

Neyman, J. (1957). "Inductive behavior" as a basic concept of philosophy of science. *Rev. Int. Statist. Inst.* **25**, 7-22.

Neyman, J. (1961). Silver Jubilee of my dispute with Fisher. *J. Oper. Res. Soc. Japan* **3**, 145-154.

Neyman, J. (1966). Behavioristic points of view on mathematical statistic. In: On Political Economy and Econometrics; Essays in honour of Oscar Lange. Polish Scient. Publ. Warsaw.

Neyman, J. (1976). Tests of statistical hypotheses and their use in studies of natural phenomena. Commun. Statist. - Theor. Meth. A5(8) 737-751.

Neyman, J. (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.

Neyman, J. and Pearson, E.S. (1933a). On the problem of the most efficient tests of statistical hypotheses. *Phil. Trans. Roy. Soc. A* **231**, 289-337.

Neyman, J. and Pearson, E.S. (1933b). The testing of statistical hypotheses in relation to probabilities a priori. *Proc. Comb. Phil. Soc.* **24**, 492-510.

Pearson, E.S. (1955). Statistical concepts in their relation to reality. *J. Roy. Statist. Soc. (B)* **17**, 204-207.

Pearson, E.S. (1962). Some thoughts on statistical inference. *Am. Math. Statist.* **33**, 394-403.

Pearson, E.S. (1974). Memories of the impact of Fisher's work in the 1920's. *Int. Statist. Rev.* **42**, 5-8.

Pearson, E.S. and Hartley, H.O. (1954). Biometrika Tables for Statisticians (Table No.11). Cambridge Univ. Press.

Pedersen, J.G. (1978). Fiducial inference. *Intern. Statist. Rev* **46**, 147-170.

Robinson, G.K. (1976). Properties of Student's t and of the Behrens-Fisher solution to the two means problem. *Ann. Statist.* **4**, 963-971.

Robinson, G.K. (1982). Behrens-Fisher problem in Enc. Statist. Sci. (Kotz and Johnson, Eds) vol. 1, 205-209. Wiley. New York.

Savage, L.J. (1976). On rereading R.A. Fisher (with discussion). *Am. Statist.* **4**, 441-500.

Schweder, T. (1988). A significance version of the basic Neyman-Pearson theory for scientific hypothesis testing. *Scand. J. Statist.* **15**, 225-242.

Seidenfeld, T. (1979). Philosophical Problems of Statistical Inference. D. Reidel Publ. Co.

Spielman, S. (1974). The logic of tests of significance. *Phil. of Sci.* **41**, 211-226.

Spielman, S. (1978). Statistical dogma and the logic of significance testing. *Phil. of Sci* **45**, 120-135.

Steger, J.A. (Ed) (1971). Readings in Statistics for the behavioral scientist. Holt, Rinehart and Winston. New York.

Tukey, J.W. (1960). Conclusions vs Decisions. *Technometrics* **2**, 424-432. (Reprinted in Steger (1971)).

Wallace, D.L. (1980). The Behrens-Fisher and Fieller-Creasy problems. In "R.A. Fisher: An Application (S.E. Fienberg and D.V. Hinkley, Eds). Springer-Verlag.

New York.