Statistical Concepts in their Relation to Reality

By E. S. Pearson

University College, London

SUMMARY

This paper contains a reply to some criticisms made by Sir Ronald Fisher in his recent article on "Statistical Methods and Scientific Induction".

Controversies in the field of mathematical statistics seem largely to have arisen because statisticians have been unable to agree on how theory is to provide, in terms of probability statements, the numerical measures most helpful to those who have to draw conclusions from observational data. We are concerned here with the ways in which mathematical theory may be put, as it were, into gear with the common processes of rational thought, and there seems no reason to suppose that there is one best way in which this can be done. If, therefore, Sir Ronald Fisher recapitulates and enlarges on his views upon statistical methods and scientific induction we can all only be grateful, but when he takes this opportunity to criticize the work of others through misapprehension of their views as he has done in his recent contribution to this *Journal* (Fisher 1955), it is impossible to leave him altogether unanswered.

In the first place it seems unfortunate that much of Fisher's criticism of Neyman and Pearson's approach to the testing of statistical hypotheses should be built upon a "penetrating observation" ascribed to Professor G. A. Barnard, the assumption involved in which happens to be historically incorrect. There was no question of a difference in point of view having "originated" when Neyman "re-interpreted" Fisher's early work on tests of significance "in terms of that technological and commercial apparatus which is known as an acceptance procedure". There was no sudden descent upon British soil of Russian ideas regarding the function of science in relation to technology and to five-year plans. It was really much simpler—or worse. The original heresy, as we shall see, was a Pearson one!

As has often been pointed out, the break with the traditional approach to the handling of tests for the significance of differences came with Student's paper of 1908, although the implications of the step which he had taken were not realized for some time. In puzzling over the relation to this step of Fisher's early theoretical papers and the first edition of his *Statistical Methods for Research Workers*, during the years 1925–27, I could not satisfy myself that the reasons which had been given for the choice of a particular test function in terms of the theory of estimation were altogether adequate. It was a question which I discussed from time to time with Student, and I have already quoted (Pearson, 1938, p. 243) a letter of his written in 1926 which contained the germ of that fruitful idea about the hypothesis tested and its alternatives. Apart from Student, I had no contact with industry at that time and it was some years before the publications of W. A. Shewhart appeared, showing the scope for statistical method in problems of acceptance sampling. Indeed, to dispel the picture of the Russian technological bogey, I might recall how certain early ideas came into my head as I sat on a gate overlooking an experimental blackcurrant plot at the East Malling Research Station!

To the best of my ability I was searching for a way of expressing in mathematical terms what appeared to me to be the requirements of the scientist in applying statistical tests to his data. After contact was made with Neyman in 1926, the development of a joint mathematical theory proceeded much more surely; it was not till after the main lines of this theory had taken shape with its necessary formalization* in terms of critical regions, the class of admissible hypotheses, the two sources of error, the power function, etc., that the fact that there was a remarkable parallelism of ideas in the field of acceptance sampling became apparent. Abraham Wald's contri-

* Necessary just as was the introduction of such terms as "sufficiency" and "amount of information" in the formal development of Fisher's theory.

butions to decision theory of ten to fifteen years later were perhaps strongly influenced by acceptance sampling problems, but that is another story.

So much for historical clarification. Turn now to some of Professor Fisher's strictures. It seems to me that he is often tilting at views which those whom he attacks have never held. Where, for example, do we really stand in regard to the phrase "repeated sampling from the same population"? As Fisher points out (first paragraph of his p. 71), if we have before us a single sample of observations resulting perhaps from some experimental procedure, the population of possible samples, which may be termed the reference set, will often have no objective reality, being only a product of the statistician's imagination. Further, he remarks:

> "In respect of tests of significance, therefore, there is need for further guidance as to how this imagination is to be exercised. In fact a careful choice has to be made, based on the understanding of the question or questions to be answered".

I agree entirely with this need for careful choice, but this was just what Neyman and I were pointing out long ago. The difference often appears to lie in the particular population of samples considered most appropriate. The two examples which Fisher first discusses, those of linear regression and the 2×2 table, do in fact as a pair throw much light on the question of what is involved in this exercise of the imagination.

Professor Fisher's choice of reference set is based, I think, on his theory of information. Thus in writing of the 2×2 table he speaks (p. 73) of "the reasonable principle that in testing the significance with a unique sample, we should compare it only with other possibilities in all relevant respects like that observed", and again, in the regression problem, he refers to "a population of samples in all relevant respects like that observed". The meaning of terms such as "relevant' is not of course self-evident without definition, but such phrases form part of Fisher's general approach to estimation theory and the reference sets adopted in these two examples are made perfectly clear.

We may, however, ask whether there are not other "reasonable principles" which might be used to guide the statistician's imagination. Here, for example, is one. If probability is to be justly applied to the analysis of data, it follows that a random process must have been introduced or been naturally present at some stage in the collection of these data. Is there not then an appeal to the imagination in taking as the hypothetical population of samples that which would have been generated by repetition of this random process?

If we follow this principle in the regression problem, we see that the reference set will depend on the character of the experiment or investigation. If the values of x were chosen in advance, then the population of samples consists of those having these fixed x's, but with varying y values. If the data were obtained by sampling N pairs of observations (x, y) freely from a bivariate population,* the population of samples may be imagined as enlarged accordingly. In both cases, however, $t = (b - \beta)\sqrt{A}/s$ will follow Student's distribution, although in the second case $A = S(x - \bar{x})^2$ will vary from sample to sample, as do b and s. From the Neyman and Pearson point of view, t would be regarded in both instances as the appropriate function of the sample to use in testing the hypothesis that $b = \beta_0$ and the awkwardness of the distribution of b itself in the second situation would be irrelevant. The population of samples having A fixed, which is the reference set of Professor Fisher's approach, can clearly be imagined but does not seem to have any experimental counterpart which, of course, from his point of view it need not.

The case of the 2×2 table provides an interesting companion example. Here, as Barnard first pointed out, data presented in the same form of table may have been obtained from a sampling or an experiment conducted in several different ways, For example, they may arise: (i) after the random partition of a number, N, of individuals into two groups which receive different "treatments"; (ii) by drawing randomly and independently a sample from each of two populations; (iii) by drawing a single random sample from a population of individuals possessing two qualitative characters. Following Fisher and Yates, the statistician should in each case relate his test of significance to the same reference set, that of the population of samples giving to the table the same marginal totals as those observed. The other principle to which I have referred would define three different reference sets, of which only that for case (i) corresponds with the Fisher and Yates set.

* In which the array distributions of y for fixed x are, of course, normal and homoscedastic.

Because we are dealing with discontinuous hypergeometric distributions and not with the normal curve, we do not obtain from this second approach, as in the case of linear regression, a test function whose distribution is the same for all three reference sets. All that we can say* is that if the table is denoted by

a	c	m
b	d	n
r	s	N

then under the null hypothesis

$$u = \frac{a - mr/N}{\left\{\frac{mnrs}{N^2(N-1)}\right\}^{\frac{1}{2}}}$$

will for all reference sets have zero expectation and unit variance.

But does the existence of this limitation establish that one principle is right, the other wrong? I think not, because there is still a further matter to be considered which is often overlooked. Having decided on the reference set that he regards as appropriate, Professor Fisher has still to set out the logical justification for measuring the level of significance in terms of the integral or the sum of the separate probabilities in the tail of the relevant probability distribution. This is a matter which has been raised by Harold Jeffreys (1948, p. 357) and again by G. A. Barnard (1949, p. 137). Starting from the reference set which they considered appropriate, Neyman and I arrived at the critical or rejection region for the sample point through a formulation of the alternatives to the null hypothesis, and as soon as these are considered in the problem of the 2×2 table it appears necessary to differentiate the cases (i), (ii) and (iii). Given the critical region, there is clearly more than one numerical measure which could be associated with it. We deliberately chose the integral or sum of the probabilities (under the null hypothesis) of the sample point falling within the region rather than, say, the value of the ratio of likelihoods on its boundary because it seemed to us the more relevant and meaningful measure to use.

It seems to me that there is still a good deal here that is worth thinking over and that we shall get no nearer to a solution of the logical problems involved by throwing up the question "repeated samples from the same population?" and answering, in effect, "what nonsense!" We have only to turn to D. V. Lindley's recent paper (1953) and the discussion which followed to realize the continued value of an unrestricted play of thought round these problems.

Professor Fisher's next objection is to the use of such terms as the "acceptance" or "rejection" of a statistical hypothesis, and "errors of the first and second kinds". It may be readily agreed that in the first Neyman and Pearson paper of 1928, more space might have been given to discussing how the scientific worker's attitude of mind could be related to the formal structure of the mathematical probability theory that was introduced. Nevertheless it should be clear from the first paragraph of this paper that we were not speaking of the *final* acceptance or rejection of a *scientific* hypothesis on the basis of statistical analysis. We speak of accepting or rejecting a hypothesis with a "greater or less degree of confidence". Further, we were very far from suggesting that statistical methods should force an irreversible acceptance procedure upon the experimenter. Indeed, from the start we shared Professor Fisher's view that in scientific enquiry, a statistical test is "a means of learning", for we remark: "the tests themselves give no final verdict, but as tools help the worker who is using them to form his final decision". No doubt we could more aptly have said "his final or provisional decision"; even scientists, if they are employed in research departments by industry or government, may sometimes have to give a final decision.

As already mentioned, a certain simplification of real situations and a formalization in the verbal expression of ideas seems unavoidable when one attempts to put mathematical theory into gear with the way the mind works. 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

* Unless we follow K. D. Tocher's (1950) suggestion of adding to a a random variable uniformly distributed in the interval (0, 1). Then, we can use a test function whose probability distribution is the same for all the three reference sets of cases (i), (ii) and (iii).

different from that which Professor Fisher himself has now reached. On p. 73 of his last (1955) paper, he sets out as alternatives (*a*) and (*b*) what he thinks may be, according to the circumstances, the worker's attitude in a case where the test of significance applied gives no strong reason for rejecting the null hypothesis. The phrases used, cautious though they are, are yet so relevant to the understanding of the Neyman and Pearson approach, that I shall quote them here. The worker is stated to express himself as follows:

> (*a*) "The possible deviation from truth of my working hypothesis, to examine which the test is appropriate, seems not to be of sufficient magnitude to warrant any immediate modification".

or

> (*b*) "The deviation is in the direction expected for certain influences which seemed to me not improbable, and to this extent my suspicion has been confirmed; but the body of data available so far is not by itself sufficient to demonstrate their reality".

What ideas seem to underlie these statements? There is a recognition that the probable deviations from the working hypothesis lie in a particular direction. This seems to imply that the appropriate test is one which should be sensitive, indeed as sensitive as possible, to deviations in that direction. If he is going to have to discard his working hypothesis, the scientist would presumably like to be able to reach the conclusion that this is necessary with the greatest economy of effort in experimentation. Under these conditions, as part of the mathematical structure which would help to determine the appropriate test (or to compare alternative tests), Neyman and I introduced the notions of the class of admissible hypotheses and the power function of a test. The class of admissible alternatives is formally related to the direction of probable deviations —changes in mean, changes in variability, departure from linear regression, existence of interactions, or what you will. The power function will help to indicate what amount of data may be required to demonstrate the reality of specific departures from the working hypothesis.

It seems to me that continuing on the lines of statements (*a*) and (*b*), we may imagine our worker to go further and to enlarge on the term "appropriate" as follows:

> (*c*) "The appropriate test is one which, while involving (through the choice of its significance level) only a very small risk of discarding my working hypothesis prematurely will enable me to to demonstrate with assurance (but without an unnecessary amount of experimentation) the reality of the influences which I suspect may be present".

If we accept (*c*) as a reasonable expression of attitude, it seems to follow that our worker has among other things two balancing considerations in his mind; he wants to avoid:

> (1) discarding his working hypothesis prematurely,
>
> (2) waiting an unnecessarily long time before reaching the conclusion that suspected factors are influencing the situation.

The formal description of this situation as involving the Scylla and Charybdis of two possible "sources of error", may be abhorrent to him. But perhaps, cautious as this ideal scientist is, he would admit to a desire to avoid being wrong in a tentative opinion expressed, let us say, in an informal discussion following another scientific colleague's paper read before a learned society!

Professor Fisher's final criticism concerns the use of the term "inductive behaviour"; this is Professor Neyman's field rather than mine.

References

BARNARD, G. A. (1949), *J. R. Statist. Soc.* B, **11**, 115–139.
FISHER, R. A. (1955), *J. R. Statist. Soc.* B, **17**, 69–78.
JEFFREYS, H. (1948), *Theory of Probability*. Oxford University Press.
LINDLEY, D. V. (1953), *J. R. Statist. Soc.* B, **15**, 30–76.
NEYMAN, J., & PEARSON, E. S. (1928), *Biometrika*, **20A**, 175–240.
PEARSON, E. S. (1938), *Biometrika*, **30**, 210–50.
TOCHER, K. D. (1950), *Biometrika*, **37**, 130–44.