Statistical Methods and Scientific Induction

By Sir Ronald Fisher

Department of Genetics, University of Cambridge

SUMMARY

The attempt to reinterpret the common tests of significance used in scientific research as though they constituted some kind of acceptance procedure and led to "decisions" in Wald's sense, originated in several misapprehensions and has led, apparently, to several more.

The three phrases examined here, with a view to elucidating the fallacies they embody, are:

(i) "Repeated sampling from the same population",
(ii) Errors of the "second kind",
(iii) "Inductive behaviour".

Mathematicians without personal contact with the Natural Sciences have often been misled by such phrases. The errors to which they lead are not always only numerical.

1. *Introduction*

During the present century a good deal of progress seems to have been made in the business of interpreting observational data, so as to obtain a better understanding of the real world. The three aspects of principle importance for this progress have been, first, the use of better mathematics and more comprehensive ideas in mathematical statistics; leading to more correct or exact methods of calculation, applied to the given body of data (a unique sample in the language of W. S. Gosset, writing under the name of "Student") which comprehends all the numerical information available on the topic under discussion. Secondly, as methods of summarizing and drawing correct conclusions approached adequacy, the wide subject of experimental design was opened up, aimed at obtaining data more complete and precise, and at avoiding waste of effort in the accumulation of ill-planned, indecisive, or irrelevent observations. Thirdly, as a natural or even inevitable concomitant of the first two, a more complete understanding has been reached of the structure and peculiarities of inductive logic—that is of reasoning from the sample to the population from which the sample was drawn, from consequences to causes, or in more logical terms, from the particular to the general.

Much that I have to say 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; different in particular from that expressed in numerous papers by Neyman, Pearson Wald and Bartlett. There is no difference to matter in the field of mathematical analysis, though different numerical results are arrived at, but there is a clear difference in logical point of view, and I owe to Professor Barnard of The Imperial College 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.

Now, acceptance procedures are of great importance in the modern world. When a large concern like the Royal Navy receives material from an engineering firm it is, I suppose, subjected to sufficiently careful inspection and testing to reduce the frequency of the acceptance of faulty or defective consignments. The instructions to the Officers carrying out the tests must also, I conceive, be intended to keep low both the cost of testing and the frequency of the rejection of satisfactory lots. Much ingenuity and skill must be exercised in making the acceptance procedure a really effectual and economical one. I am casting no contempt on acceptance procedures, and

I am thankful, whenever I travel by air, that the high level of precision and reliability required can really be achieved by such means. But the logical differences between such an operation and the work of scientific discovery by physical or biological experimentation seem to me so wide that the analogy between them is not helpful, and the identification of the two sorts of operation is decidedly misleading.

I shall hope to bring out some of the logical differences more distinctly, but there is also, I fancy, in the background an ideological difference. Russians are made familiar with the ideal that research in pure science can and should be geared to technological performance, in the comprehensive organized effort of a five-year plan for the nation. How far, within such a system, personal and individual inferences from observed facts are permissible we do not know, but it may be safer, and even, in such a political atmosphere, more agreeable, to regard one's scientific work simply as a contributary element in a great machine, and to conceal rather than to advertise the selfish and perhaps heretical aim of understanding for oneself the scientific situation. In the U.S. also the great importance of organized technology has I think made it easy to confuse the process appropriate for drawing correct conclusions, with those aimed rather at, let us say, speeding production, or saving money. There is therefore something to be gained by at least being able to think of our scientific problems in a language distinct from that of technological efficiency.

I believe I can best illustrate the contrast I want to make clear by taking a few current phrases which are foreign to my own point of view, and after examining these, by setting out in a more constructive spirit, some of the special characteristics of inductive reasoning. The phrases I should choose for the fallacies they embody are:

(i) Repeated sampling from the same population.
(ii) Errors of the "second kind".
(iii) "Inductive behaviour".

But first I must exemplify the extent to which divergence in language has been carried by quoting some rather simple phrases from Wald's book on Decision Functions.

On the outside of the cover we read, "Particularly noteworthy is the treatment of experiment design as a part of the general decision problem".

On the inside, "The design of experimentation is made a part of the general decision problems—a major advance beyond previous results", and in the first paragraph of the author's preface "A major advance beyond previous results is the treatment of the design of experimentation as a part of the general decision problem".

These claims seem very much like an afterthought, of a kind which is sometimes suggested by a publisher; for, apart from these three quotations, the design of experiments is scarcely mentioned in the rest of the book. For example, the index does not contain the word "replication", or "control", or "randomization"; there is no discussion of the functions and purposes of these three elements of design. Of authorities, the bibliography does not contain the names of Yates, of Finney, or of Davies; or, on the other side of the Atlantic, of Goulden, who was the first of transatlantic writers on the design of experiments, or of Cochran and Cox. My own book is indeed mentioned, but no use seems to have been made of it. The obvious inference is that Wald was quite unaware of the nature and scope of the subject of experimental design, but had simply assumed that it *must* be included in that of acceptance procedures, to which his book is devoted. Rather similar, equally innocent and unfounded presumptions, have been not uncommon in the last twenty years. They would scarcely have been possible without that insulation from all living contact with the natural sciences, which is a disconcerting feature of many mathematical departments.

The first questionable phrase and the one responsible for the greatest amount of *numerical* error is:

2. "*Repeated Sampling from the Same Population*"

The operative properties of an acceptance procedure, single or sequential, are ascertained practically or conceptually by applying it to a series of successive similar samples from the same source of supply, and determining the frequencies of the various possible results. It is doubtless in consequence of this that it has been thought, and frequently asserted, that the validity of a test of significance is to be judged in the same way. However, a rather large number of examples are now known in which this rule is seen to be misleading. The root of the difficulty of carrying over the idea from the field of acceptance procedures to that of tests of significance is that, where acceptance procedures are appropriate, the source of supply has an objective reality, and the population of lots, or one or more, which could be successively chosen for examination is uniquely defined; whereas if we possess a unique sample in student's sense on which significance tests are to be performed, there is always, as Venn (1876) in particular has shown, a multiplicity of populations to each of which we can legitimately regard our sample as belonging; so that the phrase "repeated sampling from the same population" does not enable us to determine which population is to be used to define the probability level, for no one of them has objective reality, all being products of the statistician's imagination. 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 an understanding of the question or questions to be answered. By ignoring this necessity a "theory of testing hypotheses" has been produced in which a primary requirement of any competent test has been overlooked.

Consider the case of simple linear regression. Let us suppose that the numerical data consist of N pairs of values (x, y), while the qualitative data tell us that for each value of x the variate y is normally distributed with variance σ about a mean given by

$$Y \equiv E_x(y) = \alpha + x\beta, \qquad E_x(y - Y)^2 = \sigma^2,$$

being a linear function of the variate x. The qualitative data may also tell us how x is distributed, with or without specific parameters; this information is irrelevant.

In such cases the unknown parameter, β, may be estimated and the precision of estimation determined by a standard and well known procedure; let

$$A = S(x - \bar{x})^2,\ \ B = S(x - \bar{x})(y - \bar{y}), \quad C = S(y - \bar{y})^2.$$

Then we may take as our estimate of β the statistic

$$b = B/A,$$

and of σ^2 the statistic

$$s^2 = (C - B^2/A) \div (N - 2).$$

For samples having the same value A it is easy to show that the estimate b is normally distributed about β with variance σ^2/A, so that we have a typical analysis of variance:

d.f.	*Sum of Squares*	*Mean Square*
1	$A(b - \beta)^2$	$A(b - \beta)^2$
$N - 2$	$C - B^2/A$	s^2

and the significance of the deviation of b from zero, or any other proposed value of β, is a simple t-test with $N - 2$ degrees of freedom, with

$$t = (b - \beta_0)\frac{\sqrt{A}}{s},$$

where β_0 is the theoretical value proposed for comparison.

I do not believe that anyone doubts the validity of this simple test. It does, however, violate the rule of determining levels of significance by frequencies of occurrence of the proposed events in repeated samples from the same population. For if a succession of sets of N pairs of observations (x, y) were taken from the same population, the value of A would not be the same for each set. Consequently, the frequency distribution of $b - \beta$ in the aggregate of all such sets would not be the same as that which I have calculated taking A constant, and would indeed be unknown until the sampling variation of A were investigated. In reality, therefore, no one uses the rule of determining the level of significance by successive sampling from the population of *all* random samples of N pairs of values, but, ever since the right approach was indicated (Fisher 1922), the *selection* of all random samples having a constant value A, equal to that actually observed in the sample under test, is what has in fact been used. The normal distribution of b about β with variance σ^2/A does not correspond with any realistic process of sampling for acceptance, but to a population of samples in all relevant respects like that observed, neither more precise nor less precise, and which therefore we think it appropriate to select in specifying the precision of the estimate, b. In relation to the estimation of β the value A is known as an *ancillary statistic.* Had it been necessary we should not have hesitated to specify all values of x ($x_1, \ . \ . \ . \ , x_N$) individually, but this would have made no difference once the comprehensive value A had been specified.

The confusion introduced, even in the case of the most fundamental and logically simple of tests of significance, by the introduction of the notion of basing the test on repeated sampling from the same population, is well illustrated by some episodes, which ought not to be forgotten, in the curious history of testing proportionality in a two-by-two table.

In the solution of the problem of the 2 × 2 table, put forward concurrently by Dr. F. Yates and myself in 1934, the essential point was the recognition that the probabilities of occurrence of different possible tables, *having the same marginal totals*

$$\begin{array}{cc|c} a & b & a+b \\ c & d & c+d \\ \hline a+c & b+d & n \end{array}$$

were proportional simply to

$$1/a\,!\,b\,!\,c\,!\,d\,!$$

where a, b, c and d are the four frequencies observed in the double dichotomy, whatever might be the probabilities governing the marginal distributions. Within sets of tables having the same margins, therefore, each may be assigned an absolute probability:

$$\frac{(a+b)\,!\,(a+c)\,!\,(b+d)\,!\,(c+d)\,!}{n\,!}\cdot\frac{1}{a\,!\,b\,!\,c\,!\,d\,!}$$

where the new factor depends only on the margins and not on the contents.

In this case the margins of the table, which by themselves supply no information as to the proportionality of the contents, do, like the value A in the regression example, determine how much information the contents will contain. 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, will lead us to set aside the various possible tables having different margins, the relative frequencies of which must depend on unknown factors of the population sampled.

On two occasions in the intervening twenty years distinguished statisticians have attempted to bring into the account populations of fourfold tables not having fixed margins. In both cases, such is the reasonableness of human nature in favourable cases, the authors of these innovations withdrew them after some discussion, and expressed themselves as completely satisfied that the apparent advance they had made was illusory. The first was Professor E. B. Wilson of the Harvard School of Public Health, writing in *Science* in 1941, and later taking occasion to expound the method of Fisher and Yates in two papers in the *Proceedings of the National Academy of Sciences* in the following year. The second case was that of Professor Barnard, who started on the assumption that the method expounded by Neyman and Pearson could be relied on, and in the first flush of success reported a test using the language of that theory "much more powerful than Fisher's", but who also, after some discussion, had the generosity to go out of his way to explain that further meditation had led him to the conclusion that Fisher was right after all.

Professor Barnard has a keen and highly trained mathematical mind, and the fact that he was misled into much wasted effort and disappointment should be a warning that the theory of testing hypotheses set out by Neyman and Pearson has missed at least some of the essentials of the problem, and will mislead others who accept it uncritically. Indeed, in the matter of Behren's test for the significance of the difference between the means of two small samples, objection was taken on exactly the ground that the significance level is not the same as the frequency found on repeated sampling.

The examples I have given from simpler problems show clearly that it should never have been put forward in the field of significance tests, though perhaps perfectly appropriate to acceptance sampling.

3. *Errors of the "Second Kind"*

The phrase "Errors of the second kind", although apparently only a harmless piece of technical jargon, is useful as indicating the type of mental confusion in which it was coined.

In an acceptance procedure lots will sometimes be accepted which would have been rejected had they been examined fully, and other lots will have been rejected when, in this sense, they ought to have been accepted. A well-designed acceptance procedure is one which attempts to minimize the losses entailed by such events. To do this one must take account of the costliness of each type of error, if errors they should be called, and in similar terms of the costliness of the testing process; it must take account also of the frequencies of each type of event. For this reason probability *a priori*, or rather knowledge based on past experience, of the frequencies with which lots of different quality are offered, is of great importance; whereas, in scientific research, or in the process of "learning by experience", such knowledge *a priori* is almost always absent or negligible.

Simply from the point of view of an acceptance procedure, though we may by analogy think of these two kinds of events as "errors" and recognize that they are errors in opposite directions, I doubt if anyone would have thought of distinguishing them as of two kinds, for in this *milieu* they are essentially of one kind only and of equal theoretical importance. It was only when the relation between a test of significance and its corresponding null hypothesis was confused with an acceptance procedure that it seemed suitable to distinguish errors in which the hypothesis is rejected wrongly, from errors in which it is "accepted wrongly" as the phrase does. The frequency of the first class, relative to the frequency with which the hypothesis is true, is calculable, and therefore controllable simply from the specification of the null hypothesis. The frequency of the second kind must depend not only on the frequency with which rival hypotheses are in fact true, but also greatly on how closely they resemble the null hypothesis. Such errors are therefore incalculable both in frequency and in magnitude merely from the specification of the null hypothesis, and would never have come into consideration in the theory only of tests of significance, had the logic of such tests not been confused with that of acceptance procedures.

It may be added that in the theory of estimation we consider a continuum of hypotheses each eligible as null hypothesis, and it is the aggregate of frequencies calculated from each possibility in turn as true—including frequencies of error, therefore only of the "first kind", without any assumptions of knowledge *a priori*—which supply the likelihood function, fiducial limits, and other indications of the amount of information available. The introduction of allusions to errors of the second kind in such arguments is entirely formal and ineffectual.

The fashion of speaking of a null hypothesis as "accepted when false", whenever a test of significance gives us no strong reason for rejecting it, and when in fact it *is* in some way imperfect, shows real ignorance of the research workers' attitude, by suggesting that in such a case he has come to an irreversible decision.

The worker's real attitude in such a case might be, according to the circumstances:

(*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 it might be:

(*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."

These examples show how badly the word "error" is used in describing such a situation. Moreover, it is a fallacy, so well known as to be a *standard* example, to conclude from a test of significance that the null hypothesis is thereby established; at most it may be said to be confirmed or strengthened.

In an acceptance procedure, on the other hand, acceptance is irreversible, whether the evidence for it was strong or weak. It is the result of applying mechanically rules laid down in advance;

no *thought* is given to the particular case, and the tester's state of mind, or his capacity for *learning*, is inoperative.

By contrast, the conclusions drawn by a scientific worker from a test of significance are *provisional*, and involve an intelligent attempt to *understand* the experimental situation.

4. "*Inductive Behaviour*"

The erroneous insistence on the formula of "repeated sampling from the same population" and the misplaced emphasis on "errors of the second kind" seem both clearly enough to flow from the notion that the process by which experimenters learn from their experiments might be equated to some equivalent acceptance procedure. The same confusion evidently takes part in the curious preference expressed by J. Neyman for the phrase "inductive behaviour" to replace what he regards as the mistaken phrase "inductive reasoning".

Logicians, in introducing the terms "inductive reasoning" and "inductive inference" evidently imply that they are speaking of processes of the mind falling to some extent outside those of which a full account can be given in terms of the traditional deductive reasoning of formal logic. Deductive reasoning in particular supplies no essentially new knowledge, but merely reveals or unfolds the implications of the axiomatic basis adopted. Ideally, perhaps, it should be carried out mechanically. It is the function of inductive reasoning to be used, in conjunction with observational data, to add new elements to our theoretical knowledge. That such a process existed, and was possible to normal minds, has been understood for centuries; it is only with the recent development of statistical science that an analytic account can now be given, about as satisfying and complete, at least, as that given traditionally of the deductive processes.

When, therefore, Neyman denies the existence of inductive reasoning he is merely expressing a verbal preference. For him "reasoning" means what "deductive reasoning" means to others. He does not tell us what in his vocabulary stands for inductive reasoning, for he does not clearly understand what that is. What he tells us to call "inductive behaviour" is merely the practice of making some assertion of the form

$$T < \theta$$

in some circumstances, and refraining from this assertion in others. This is evidently an effort to assimilate a test of significance to an acceptance procedure. From a test of significance, however, we learn more than that the body of data at our disposal would have passed an acceptance test at some particular level; we may learn, if we wish to, and it is to this that we usually pay attention, at what level it would have been doubtful; doing this we have a genuine measure of the confidence with which any particular opinion may be held, in view of our particular data. From a strictly realistic viewpoint we have no expectation of an unending sequence of similar bodies of data, to each of which a mechanical "yes or no" response is to be given. What we look forward to in science is further data, probably of a somewhat different kind, which may confirm or elaborate the conclusions we have drawn; but perhaps of the same kind, which may then be added to what we have already, to form an enlarged basis for induction.

Neyman reinforces his choice of language by arguments much less defensible. He seems to claim that the statement (*a*) "θ has a probability of 5 per cent. of exceeding T" is a different statement from (*b*) "T has a probability of 5 per cent. of falling short of θ". Since language is meant to be used I believe it is essential that such statements, whether expressed in words or symbols, should be recognized as equivalent, even when θ is a parameter, defined as an objective character of the real world, entering into the specification of our hypothetical population, whilst T is directly calculable from the observations. To prevent the kind of confusion that Neyman has introduced we may point out that both statements are statements of the relationship in which T, or θ, stands to the other. Also, since *probability* is specified, the statements have meaning only in relation to a sufficiently well-defined population of pairs of these values. The statements do not imply that in this population of pairs of values either T or θ is constant, but also they do not exclude the possibility that one should be constant, and that variability should be confined to the other. Reference to the mode of calculating our limits in an ordinary test of significance will generally establish that in these calculations the parameter θ has been treated provisionally as constant, and variations calculated of T for given θ. The possible variation of θ is left arbitrary, and is irrelevant to the calculations, much as is the distribution of the independent variate in the regression problem.

A complementary doctrine of Neyman violating equally the principles of deductive logic is to accept a general symbolical statement such as

$$Pr\{(\bar{x} - ts) < \mu < (\bar{x} + ts)\} = \alpha\,,$$

as rigorously demonstrated, and yet, when numerical values are available for the statistics $\bar{x}$ and s, so that on substitution of these and use of the 5 per cent. value of t, the statement would read

$$Pr\{92{\cdot}99 < \mu < 93{\cdot}01\} = 95 \text{ per cent.},$$

to *deny* to this *numerical* statement any validity. This evidently is to deny the syllogistic process of making a substitution in the major premise of terms which the minor premise establishes as equivalent. By this, which is surely a desperate measure, Neyman supports the assertion that if μ stand for some objective constant of nature, or property of the real world, such as the distance of the sun, its probability of lying between any named numerical limits is necessarily either 0 or 1, and we cannot know which, unless the true distance is known to us. The paradox is rather childish, for it requires that we should wilfully misinterpret the probability statement so as to pretend that the population to which it refers is not defined by our observations and their precision, but is absolutely independent of them. As this is certainly not what any astronomer means, and is not in accordance with the origin of the statement he makes, it seems rather like an acknowledgement of bankruptcy to pretend that it is.

Finally let me add some notes on what appear to me to be distinctive requirements of valid inductive inference.

5. *Requirements of Inductive Inferences*

(*a*) Since some inductive inferences are expressed in terms of *probability* (fiducial probability) the first requirement is a clear understanding that probability statements always have reference to some sufficiently defined population, and never to individuals, save as typical members of such a population. This understanding is needed for deductive inferences also, when statements of probability are made.

(*b*) A very important feature of inductive inference, unknown in the field of deductive inference, is the framing of the hypothesis in terms of which the data are to be interpreted. This hypothesis must fulfill several requirements: (i) it must be in accordance with the facts of nature as so far known; (ii) it must specify the frequency distribution of all observational facts included in the data, so that the data as a whole may be taken as a typical sample; (iii) it must incorporate as parameters all constants of nature which it is intended to estimate, in addition possibly to special, or *ad hoc*, parameters; (iv) it must not be contradicted, *in any way judged relevant*, by the data in hand. If it satisfies these conditions it is therefore a scientific construct of a fairly elaborate type. It is by no means obvious that different persons should not put forward different successful hypotheses, among which the data can supply little or no discrimination. The hypothesis is sometimes called a model, but I should suggest that the word model should only be used for aspects of the hypothesis between which the data cannot discriminate. As an act of construction the hypothesis is not altogether impersonal, for the scientist's personal capacity for theorizing comes into it; moreover, the criteria by which it is approved require a certain honesty, or integrity, in their application.

(*c*) In one respect inductive reasoning is more strict than is deductive reasoning, since in the latter any item of the data may be ignored, and valid inferences may be drawn from the rest; i.e. from any selected sub-set of the set of axioms used, whereas in inductive inference the whole of the data must be taken into account. This seems to be very difficult to be understood by workers trained in deductive methods only, though more easily understood by statisticians. The political principle that anything can be proved by statistics arises from the practice of presenting only a selected sub-set of the data available.

In some early results of my own I rely on the datum "There is no knowledge of probabilities *a priori*". They would not certainly have been legitimate without this datum, but they have been mistakenly described as a kind of greatest common factor of the inferences which could be drawn for different possible data giving probabilities *a priori*.

It is revealing that this logical distinction was overlooked by Neyman and Pearson, in 1933, in one of their earliest papers after they had learnt of the possibility of inferring fiducial limits, the argument for which I had set out in a paper on *inverse probability* in the *Proceedings of the Cambridge Philosophical Society*, 1930. It is particularly instructive that although in that paper I speak of "learning by experience", of "inductive processes", and of "the probability of causes", much as others had done since the eighteenth century, these authors read into my work "rules of behaviour", which indeed I had not mentioned at all. Both misapprehensions become intelligible if we realise that the authors had no idea of a test of significance as a means of learning, but conceived it only under the form of an acceptance procedure. The passage is as follows:

"In a recent paper [(Neyman & Pearson, 1933*b*)] we have discussed certain general principles underlying the determination of the most efficient tests of statistical hypotheses, but the method of approach did not involve any detailed consideration of the question of *a priori* probability. We propose now to consider more fully the bearing of the earlier results on this question and in particular to discuss what statements of value to the statistician in reaching his final judgement can be made from an analysis of observed data, which would not be modified by any change in the probabilities *a priori*. In dealing with the problem of statistical estimation, R. A. Fisher has shown how, under certain conditions, what may be described as *rules of behaviour* can be employed which will lead to results independent of these probabilities; in this connection he has discussed the important conception of what he terms fiducial limits.[8, 9]. But the testing of statistical hypotheses cannot be treated as a problem in estimation, and it is necessary to discuss afresh in what sense tests can be employed which are independent of *a priori* laws."

There seems here an entirely genuine inability to conceive that when new data are added in an inductive problem, previously correct conclusions are no longer correct. Or, in this case that the conclusions proper to the absence of knowledge of probabilities *a priori* would be wrong for almost any set of such probabilities, and could in no sense be a common term in the proper inferences from all such sets.

(*d*) Variety of logical form.

A fourth feature which has emerged in the study of inductive inference is that data of apparently the same logical form, though with different mathematical specification, give rise to inferences not always of the same logical form.

For example, when in 1930 I introduced the notions of the fiducial distribution and fiducial limits I did so with the example of the sampling distribution of the estimated correlation coefficient r for various values of the true correlation ρ. The distribution of r is continuous between the limits -1 and $+1$, and for any value of P there is a value of r, which may be called $r_P(\rho)$, such that r exceeds it with frequency $1 - P$, and falls short of it with frequency P. These functions of ρ increase monotonically from -1 to $+1$ as ρ passes from -1 to $+1$. Consequently, corresponding with any observed value r, there is a value of ρ, which may be denoted as $\rho_{1-P}(r)$ such that for this value of ρ the observed value will fall short of r with frequency P and exceed it with frequency $1 - P$. In fact if P is expressed as an explicit function

$$P = F_N(r, \rho)$$

such that the distribution of r for given ρ is given by the frequency element

$$\frac{\partial F}{\partial r}\,dr,$$

then the distribution

$$-\frac{\partial F}{\partial \rho}\,d\rho$$

will be the fiducial distribution of ρ for given r, in the sense that the frequency of exceeding any chosen value of ρ is the frequency, for that value of ρ, of r being less than the value observed. The quantiles of this distribution thus give the fiducial limits of ρ at any chosen level of significance.

Had I taken a discontinuous variate, such as the number of successes observed out of N trials, and sought in terms of the observations to obtain a fiducial distribution for the true probability, (say x), it would certainly have been possible to find a value of x such that the probability of the number of successes observed, or any higher number was, let us say 5 per cent., so that smaller values of x could be rejected at least at the 5 per cent. level of significance; but this gives only an inequality statement for the probability that x is less than any given value. Neyman seems to ignore this distinction, and to speak in both cases of confidence limits. Logically, however, the form of inference admissible is totally distinct.

Equally, statements of fiducial probability in continuous cases are only proper if the whole of the information is utilized, as it is by the use of sufficient estimates, whereas for any test of significance, however low in power, it may well be possible to point to the limits outside which parametric values are significantly contradicted by the data at a given level of significance. These also should be regarded as giving only rough statements for the fiducial probability.

There are other cases in the theory of estimation in which rather similar data yield information of remarkably different kinds. Consider, for example, the case in which x and y are two observables distributed in normal distributions with unit variance in each case, and independently, about hypothetical means ξ and η. No situation could be simpler. Suppose, however, that the data contain a functional relationship connecting ξ and η. Then different cases arise from different functional forms:

(i) If there is a simple linear connection between ξ and η, so that (ξ, η) represents a point on a given straight line, then the foot of the perpendicular from the observation point (x, y) is a sufficient estimate, and the fiducial distribution of (ξ, η) on the given line will be a normal distribution with unit variance about this estimate. All possible observations on the same perpendicular are equivalent.

(ii) If the given locus of (ξ, η) is a circle, there is no sufficient estimate; the distance of (x, y) from the centre of the given circle is, however, an ancillary statistic, which together with the maximum likelihood estimate makes the estimation exhaustive. For each possible distance an appropriately oriented fiducial distribution on the circle may be specified.

(iii) In general there is a well defined likelihood function, and therefore an estimated point of maximum likelihood. It is not obvious that any general substitute can be found for the ancillary statistic, save in an asymptotic sense, or that any statement of fiducial probability is possible in general. Thus three logically distinct types of inference arise from simple changes in the mathematical specification of the problem.

(*e*) Finally, in inductive inference we introduce no cost functions for faulty judgements, for it is recognized in scientific research that the attainment of, or failure to attain to, a particular scientific advance this year rather than later, has consequences, both to the research programme, and to advantageous applications of scientific knowledge, which cannot be foreseen. In fact, scientific research is not geared to maximize the profits of any particular organization, but is rather an attempt to improve *public* knowledge undertaken as an act of faith to the effect that, as more becomes known, or more surely known, the intelligent pursuit of a great variety of aims, by a great variety of men, and groups of men, will be facilitated. We make no attempt to evaluate these consequences, and do not assume that they are capable of evaluation in any sort of currency.

When decision is needed it *is* the business of inductive inference to evaluate the *nature* and *extent* of the uncertainty with which the decision is encumbered. Decision itself must properly be referred to a set of motives, the strength or weakness of which should have had no influence whatever on any estimate of probability. We aim, in fact, at methods of inference which should be equally convincing to all rational minds, irrespective of any intentions they may have in utilizing the knowledge inferred.

We have the duty of formulating, of summarising, and of communicating our conclusions, in intelligible form, in recognition of the right of *other* free minds to utilize them in making *their own* decisions.

References

BARNARD, G. A. (1945), "A new test for 2 × 2 tables", *Nature*, **156**, No. 3954, 177.
—— (1946), "Sequential tests in industrial statistics", *J.R. Statist. Soc.*, *Supp* . **8**, 1–21.
—— (1947*a*), "Significance tests for 2 × 2 tables", *Biometrika*, **34**, 123–138.
—— (1947*b*), "The meaning of a significance level", *Biometrika*, **34**, 179–182.
—— (1947*c*), Review: Sequential Analysis. By Abraham Wald, *J. Amer. Stat. Ass.*, **42**, 658.
—— (1949), "Statistical inference", *J. R. Statist. Soc.*, *B*, **11**, 115–139.
COCHRAN, W. G., & COX, G. M. (1950), *Experimental Designs*. New York: Wiley. London: Chapman & Hall.

Davies, O. L. (ed.) (1954), *The Design and Analysis of Industrial Experiments*. London & Edinburgh: Oliver & Boyd.

Finney, D. J. (1952), *Statistical Method in Biological Assay*. London: Griffin.

Fisher, R. A. (1922), "The goodness of fit of regression formulae, and the distribution of regression coefficients", *J.R. Statist. Soc.*, **85**, 597–612.

—— (1930), "Inverse probability", *Proc. Camb. Phil. Soc.*, **26**, 528–535.

—— (1933), "The concepts of inverse probability of fiducial probability referring to unknown parameters", *Proc. Roy. Soc.*, *A*, **139**, 343–348.

—— (1934), *Statistical Methods for Research Workers*. (5th ed. and later.) London & Edinburgh: Oliver & Boyd.

—— (1941), "The interpretation of experimental fourfold tables", *Science*, **94**, No. 2435, 210–211.

—— (1945), "A new test for 2 × 2 tables", *Nature*, **156**, No. 3961, 388.

Goulden, C. H. (1939 and 1952), *Methods of Statistical Analysis*. New York: Wiley. London: Chapman & Hall.

Neyman, J. (1938), "L'estimation statistique traité comme un problème classique de probabilité", *Actualités Scientifiques et Industrielles*, No. 739, 25–57.

—— & Pearson, E. S. (1933*a*), "The testing of statistical hypotheses in relation to probabilities *a priori*", *Proc. Camb. Phil. Soc.*, **29**, 492–510.

—— (1933*b*), "On the problem of the most efficient tests of statistical hypotheses", *Phil. Trans. Roy. Soc.*, *A*, **231**, 289–337.

Pearson, E. S. (1947), "The choice of statistical tests illustrated on the interpretation of data classed in a 2 × 2 table", *Biometrika*, **34**, 139–167.

"Student" (1908), The probable error of a mean, *Biometrika*, **6**, 1–25.

Venn, J. A. 1876), *The Logic of Chance* (2nd ed.). London: Macmillan.

Wald, A. (1950), *Statistical Decision Functions*. New York: Wiley. London: Chapman & Hall.

Wilson, E. B. (1941), "The controlled experiment and the fourfold table", *Science*, **93**, No. 2424, 557–560.

—— (1942*a*), "On contingency tables", *Proc. Nat. Acad. Sci.*, **28**, No. 3, 94–100.

Worcester, J. (1942*b*), "Contingency tables", *Proc. Nat. Acad. Sci.*, **28**, No. 9, 378–384.

Yates, F. (1949), *Sampling Methods for Censuses and Surveys*. London: Griffin.