Brit. J. prev. soc. Med. (1955), **9**, 135–139.

STATISTICAL EXPERIMENTATION IN THERAPEUTICS

BY

RAYMOND WRIGHTON

Department of Social and Industrial Medicine, University of Sheffield

The issue I propose to examine is whether, or in what circumstances, the statistical approach to the assessment of curative measures is legitimate. This issue is straightforward, since, unless the mere enumeration of cases is regarded as a statistical procedure, the distinction between the statistical and the non-statistical approach to a problem is absolute; an investigator must inevitably be aware that he is making a clearcut and significant decision when he choses to adopt one method rather than the other.

The conclusion I shall draw is that in therapeutics —and analogously in other branches of inquiry— the method of statistical experimentation is largely invalid. This is not to say that the accumulation of statistical material in medicine, or in other fields, is futile, or that the analysis of such material may not provide valuable guidance for the research worker and essential information for the administrator. It is merely to assert that the statistical experiment is not a satisfactory exploratory tool in the hands of the research worker confronted with a specific problem.

We can, I believe, arrive at this conclusion in two ways. Either we may identify ourselves with the experimentalist, who is seeking from the laboratory or clinic to promote a rational basis for therapeutics, and argue that the use of the statistical method is incompatible with the way of thought which ordinarily governs the activity of the scientist. Or, adopting the viewpoint of the mathematician or logician, we may argue that existing theories of statistical inference are inadequately grounded, and that until this matter is rectified we cannot help but accede to any objection on extra-logical grounds which is made to the use of the statistical techniques promoted by these theories; alternatively,—and this is the writer's own standpoint—we may feel ourselves able to argue in this capacity that a consistent theory is possible but that it does not elucidate the problems with which the physician or experimentalist is ultimately concerned.

We have therefore to consider two approaches to the question side by side; and if objections on general grounds to the statistical method are admissible, we must take it that the confusion which characterizes fundamental discussion of modern mathematical statistics will remain until the mathematical statistician is prepared to throw overboard so much that he abrogates in effect his right to advise on the logical structure of biological or other experimentation; on the other hand, if *logical* objections are valid, we must re-examine the position of statistical inference *vis-à-vis* general considerations in the philosophy of science.

Let us first consider objections which can be directed *on general grounds* against the use of statistical experimentation in medicine. Since, by reason of its simplicity and directness, the idea of a statistical approach to therapeutics very readily leaps to the mind, these have often been stated, and at different epochs. First, under the influence of Laplace, when Pierre Louis put forward his *numerical method* for medicine and pathology. Again, when the influence of Quetelet began to be felt in medicine; and later, in Great Britain, when Karl Pearson reintroduced and extended Quetelet's ideas. This last phase is the most interesting from the methodological point of view, since the prospect of unprecedented advances in prophylaxis and therapeutics disclosed by the contemporaneous rise of bacteriology reinforced the discussion with realistic anticipations. Opposition to Pearson's intrusion into the medical field came chiefly from Sir Almroth Wright as a serologist. The immediate point at issue related to Wright's proposals for anti-typhoid inoculation which Pearson publicly opposed on statistical grounds (Wright, 1904; Pearson, 1904). His objections were overruled, anti-typhoid inoculation as recommended by Wright was used amongst British troops in the first world war from the outset; and as a result of the controversy, Wright published several notable contributions to the analysis of the credentials of the statistical method (Wright, 1912, 1921, 1953). It will be convenient to summarize these contributions—with which the writer is in broad agreement—and refer the reader to the original papers for a fuller discussion.

When considering Wright's argument there are two points to be borne in mind. Firstly, most modern statisticians would be amongst the first to agree with some of his criticisms of the statistical procedures of his time and in particular with his strictures with respect to clinical trials with control and treated groups not strictly comparable. Secondly he was as much concerned, in his polemic, to establish the claims of the laboratory scientist against the Harley Street clinician, as to resist the threatened encroachments of the statistician. We may ignore these aspects of his argument and enumerate his fundamental objections to the statistical method. I use, where possible, his own words:

"It is commonly asserted that 'Science is measurement'—*measurement* being understood to comprise the achieving of numerical results by *enumeration weighing*, and quantitative methods, generally. That is quite erroneous; and consideration will show that it is no particular good to anyone to be able to rehearse statistical or other figures. The essence of science is certainty and not measurement."

Insofar as the statistical method is explicitly concerned to repudiate this last proposition, this is the fundamental antithesis on which Wright bases his case. Any science such as medicine relies in its application on firmly established facts which "like fixed points in a trigonometrical survey supply the foundation for mapping out the rest of the territory". In practical affairs, so-called rational decisions are made with the help of such a mapping, though the crudity which results from a limited number of fixed points being available may well be very great. It is the pretence of the method of statistical experimentation to provide direct answers to ill-defined practical questions and, in doing so, to supersede this procedure, which Wright refers to as *contesseration*. Contesseration is the fundamental inductive process. Its mode of operation is tentative and it is liable to unavoidable error; such error, however, admits of correction as the body of our knowledge grows. We delude ourselves if we suppose that there can be any shorter cut from inadequate to adequate knowledge.

"Whenever medicine has progressed it has progressed by making a new departure suggested by a chance observation or a new departure based upon a crucial laboratory experiment. . . . It has not gone forward by refining upon such discoveries by the aid of cumulative experiments." With the cumulative approach, as the physical and moral difficulties increase the intellectual harvest from the work becomes so patently unremunerative that every reasonable man will cry off from statistical inquiry."

We have only to think of the range of substances whose therapeutic value is explored in modern chemotherapy to appreciate the common sense behind this remark. It is not, however, only the barrrenness of the statistical method which is to be deplored. It is possible that, if taken seriously, it will serve to inhibit the making of that type of observation which does lead to fruitful advances.

The true man of science, according to Wright, has a conception of the complexity of nature which demands that he should wish to progress only from step to step. "The man who lives no life within the brain and has no conception of the complexity of nature . . . is always hankering after . . . the *saltus empiricus*. He tells us . . . that this procedure should, whenever possible, be employed and that it is nothing but simple good common sense to eliminate from consideration intermediate links in a chain of causation." "Your statistician," adds Wright, "aids and abets him."

Finally, we should be suspicious of the validity of the statistical method because it is so very easily employed. ". . . a man who proposes to carry out a crucial experiment should possess some native experimental ingenuity; whereas a man requires absolutely none for carrying out a statistical experiment." If the statistical method were novel this consideration would carry little weight. But in fact it has been advocated and applied for over a century. No important advance has resulted in any biological field.

We have discussed the informal aspect of the problem principally from the standpoint of the scientist. Alternatively, we can briefly consider it from the point of view of the physician. Let us suppose that Treatment *A* can be taken to cure 60 per cent. of patients and Treatment *B* to cure 40 per cent. A statistical trial cannot yield a type of information of higher order than this; its outcome will usually be less clearly defined, both in terms of stated percentages and in terms of what is meant by *cure*. What help does this information give the physician? When he has no knowledge at all of his patients and no collateral physiological knowledge of the mode of action of either of the two treatments, it tells him that it will be better, in a statistical sense, for him to use Treatment *A* rather than Treatment *B*. If he does possess any further knowledge under these two headings—and in any practical situation he certainly will—then he must inevitably ask: "May it not be that by judiciously allocating Treatment *A* and Treatment *B* between different patients I can ensure that all recover, or certainly more than the 60 per cent. corresponding

to the indiscriminate use of the 'better' treatment?" The person who has organized the trial cannot answer this question without going beyond his terms of reference, that is beyond his claim to be a fact-producer and an impartial arbiter. Viewed in this light, "The method of statistical experimentation can be used for testing the value of prophylatic measures but not for testing that of therapeutic measures."* The *statistical* trial can at the most be regarded as an adjuvant to *statistical* medicine.

In the second place the physician may ask the investigator a more sophisticated, but equally cogent question: "If you possess so little knowledge that you have to resort to the statistical method, how can you guarantee that my patients do not differ from the subjects used in your experiment in respects which are relevant to the outcome of treatment but of which you are unaware?" To this again the investigator can reply only with assurances which do not derive their sanction from the statistical information formally yielded by his enquiry.

The strongest statement of the case against the use of the method of the statistical experiment in therapeutics will therefore include the assertions:

(*a*) that the method is not in accord with ordinarily accepted scientific canons.

(*b*) that the consistent use of the method is not in accord with the aims of medical practice as commonly conceived.

Against this case lie arguments of which everyone must be fully aware. The statistical method militates against the making of inferences from an insufficient foundation in fact; it guarantees that experience of untreated cases is placed under review; it compels the systematic classification of observational material; it ensures that an investigator does not overlook cases which do not fit in with a preconceived thesis; and it seeks to present the outcome of experience in a form which is pre-eminently communicable. Does, however, the statistical method in any situation provide the *only* means for guaranteeing the intellectual integrity of the investigator? Can it not moreover perhaps introduce fallacies of its own? We may suspect on general grounds that it does. It is certainly desirable to examine the logical foundations of the method and obtain a more precise measure of these disadvantages.

In this matter Wright does not seem to exaggerate; "If, like an aphasic patient, or the ordinary man endeavouring to express himself in a foreign tongue, [we] have to struggle with a terminology which fails to provide words for quite fundamental notions, it is safe to assume that we are dealing with a science which has attracted to its service only men of mediocre ability. . . . The vocabulary of *Statistical Experimentation* is, as the reader who has any acquaintance with it does not require to be told, distressingly defective."

Wright's biographer remarks that he probably made too little allowance for recent progress in the development of statistical technique (Colebrook, 1954). This, however, is beside the point. Any advance in theoretical statistics can be of sufficiently radical importance to overturn Wright's main thesis, only if it involves a clarification of the disputed vocabulary of the subject. The majority of modern handbooks of statistical methodology contain no clearer definitions of such terms as *random*, *probable*, and *significant* than are implicit in Venn's "Logic of Chance", the first edition of which appeared in 1866. Judged by this criterion, progress has occurred in the mathematical but not in the logical structure of theoretical statistics. Whatever fundamental advance has been achieved does not obtrude into the practical handbooks.

The generalized statistical experiment may be treated formally as a sampling procedure in which balls are drawn haphazard from an urn in order to assess relations between the original proportions of different known types of ball within the urn (Wrighton, 1953). The fundamental problem of statistical inference therefore lies in giving precision to the type of inference which can be made in formal situations of this type. The classical approach to this problem largely persists and presupposes that it is in the following form that the problem is to be attacked. An urn contains unknown proportions of classified types of ball; a number of balls has been withdrawn at random, yielding a sample whose constitution is to be analysed in such a way as to tell us what can be inferred about the original contents of the urn. We may refer to this approach as *retrospective* or *analytical*.

Granted this approach we may adopt one of several procedures. We may perhaps, on the basis of the fixed body of data with which we are presented, attempt to make a statement about the state of mind which should be induced by an observed discrepancy between sample proportions. This we do if we operate a significance test. Again, we may seek to pass *exact* judgement on the true value of a population proportion. This we do if we adopt a point estimation procedure. Again, we may employ

* Wright (1953), p. 203.

a decision test and seek thereby to recommend that a particular course of action should be taken as a result of the sample observation. Lastly, we may seek to erect a scale of "probabilities" related to statements, each of which can be regarded as *a priori* possible, concerning the contents of the urn.

All these procedures have one significant feature in common. In each case we ignore collateral evidence and attempt to pass a judgement with putative practical significance upon evidence provided by a closed body of data. In Wright's phraseology, we ignore the requirement that judgements formed from observational data should be made by contesseration with other information available in the field. Except in special and uninteresting cases this formal neglect of collateral evidence must certainly render any calculus with pretences to the weighing of evidence nugatory in practical application. Can, however, such a calculus be developed, which is even consistent, in the sense that different judges may be drawn by it to identical numerical conclusions in respect of the implications of particular closed bodies of sampling data? On general grounds we cannot reasonably anticipate that the answer to this question can be *yes*; and experience shows that all theories so far elaborated, and however plausible initially, collapse on analysis. The ultimately operative charge against them is that of equivocality. The Bayes-Laplace calculus has so often been refuted, not only because of justifiable reluctance to accept the concept of prior probabilities and the postulate of their equality—these in themselves need afford no deterrent to the mathematician—but also because it is impossible to distribute these equi-probabilities unambiguously even in formal situations. Similarly, the initially intuitive approach to significance testing must be rejected, not merely because of its inherent irrelevance in practice, but also because (as J. Neyman and E. S. Pearson have shown) it is impossible, except in artificial and trivial situations, to assign appropriate regions of rejection unequivocally. We might comment similarly on other forms of retrospective calculus.

What holds for these inverse theories holds also for Wald's theory of Decision Functions, initially advanced to supply a logical foundation for recently elaborated statistical techniques* in the same way as, earlier, Jeffreys put forward a form of the classical inverse calculus to explain techniques associated with the name of R. A. Fisher. It is often said that Wald and Jeffreys stand for ways of thought which are completely antithetical. This seems to me to be a gross misunderstanding. The dividing line between a calculus which aims to provide numerical judgements which may have bearing on the making of a decision and one which actually dictates the decision is so vaguely defined that discussion of objections to Wald's theory must largely rehearse the case against that of Jeffreys. Wald's minimax principle is certainly unambiguous in a way in which Bayes' postulate is not; but it is equally arbitrary. Collateral evidence is allowed for in the requirement that a whole range of assessments of conditional risks must be made. But these assessments are to be in numerical terms and could not be made in any situation conceivable in medical research. Moreover, could they be made, the investigator would be in possession of so much knowledge that the initial undertaking of inquiry at a statistical level would be superfluous.

The only conclusion we can therefore draw from recent studies is that, on purely formal grounds, it is futile to attempt to develop a calculus of judgement or of decision. This is hardly compelling, since the plain man will have reached the same conclusion in the absence of intellectual effort. We are left with a line of development opened up by J. Neyman. To Neyman is due the notion of interval estimation by direct probabilities. This approach holds out more promise of success, since it suggests that we can hope to generate a class of statistical facts which can be viewed on almost equal terms, and therefore in effect contesseratively, with related non-statistical facts. Here we have a complete break with tradition. The retrospective approach is seemingly abandoned and the unattainable goal of judgement is ignored. But Neyman seems to have regarded his notion as purely mathematical, and has attempted to use it, not to throw light upon the nature of statistical reasoning, but to develop a retrospective calculus which is almost conventional. R. A. Fisher pointed out very early that the resulting *theory of confidence intervals* falls down because it does not admit of consistent development and he himself advanced contemporaneously the theory of *fiducial probability*, which uses the same basic notion in an inverse calculus applicable, however, only in a limited number of ideal situations. These theories do no more than provide further examples of the futility of the analytical approach.

I have put forward elsewhere (Wrighton, 1953) what I believe to be the correct explanation of the central paradox of the theory of confidence intervals and a suggestion with respect to its resolution in the context of the therapeutic trial. If we try to resolve the paradox on these terms we deduce a *prospective* calculus, and are led to an approach to statistical

* *e.g.* Sequential Analysis.

inference which drastically limits its practical scope and its mathematical interest. We are forced to admit that a primary problem in the comparative trial is the specification of the number of subjects which must be used, and to agree with the older school of statisticians that, for any results of conceivable value to accrue this number must be very large.

In the light of our earlier discussion this is not surprising. But it would be wrong to leave the impression that our attitude towards the place of statistics in therapeutics should be defined purely as a result of an elucidation of the logical credentials of its methodology. Were arbitrarily large samples available, the problem of statistical inference would become unimportant and we could reasonably proceed by identifying sample values with population values. But some of our earlier strictures on the utility of the statistical method would still remain operative. To this extent, therefore, it is these which constitute the more fundamental aspects of the case against statistical experimentation.

REFERENCES

Colebrook, L. (1954). "Almroth Wright". Heinemann, London.
Pearson, K. (1904). *Brit.med.J.*, **2**, 1243, 1432, 1542, 1667, 1775.
Venn, J. (1866). "The Logic of Chance". Macmillan, London.
Wright, A. E. (1904). *Brit.med.J.*, **2**, 1343, 1489, 1614, 1727.
—— (1912). *Lancet*, **2**, 1633, 1701.
—— (1921). Preface to 2nd ed. of "Technique of the Teat and Capillary Glass Tube". Constable, London.
—— (1953). "Alethetropic Logic", p. 203. Heinemann, London.
Wrighton, R. F. (1953). *Acta genet.* (*Basel*), **4**, 312.