# PRINCIPLES OF MEDICAL STATISTICS

## XV—GENERAL SUMMARY AND CONCLUSIONS*

IN the preceding sections I have endeavoured to make clear to the non-mathematically inclined worker some of the technique that the statistician employs in presenting and in interpreting figures. The major part of that discussion has been directed to two basic problems :—

(1) The "*significance*," or reliability in the narrow sense, of a difference which has been observed between two sets of figures—be those figures averages, measures of variability, proportions, or distributions over a series of groups ; and

(2) The *inferences* that can be drawn from a difference which we are satisfied is not likely to be due to chance.

### A Secure Foundation for Argument

The discussion of the first problem led to the development of tests of "significance"—the standard errors of individual values, the standard errors of the differences between values, and the $\chi^2$ test. The object of such tests is to prevent arguments being built up on a foundation that is insecure owing to the inevitable presence of sampling errors. Medical literature is full of instances of the neglect of this elementary precaution. Illustration is hardly necessary but I may, perhaps, give a quotation from an article published while I was preparing this section for the press : "a mere list of the treatments which have been tried in thrombo-angiitis obliterans would be of formidable length and there is little point in mentioning many of them—they have only too often fallen by the way after an introduction more optimistic than warranted by results" (*Lancet*, 1937, 1, 551). This general summary may well be written round that problem of clinical trials.

In general, worker A, who is at least careful enough to observe a control group, reports after a short series of trials that a particular method of treatment gives him a greater proportion of successes than he secures with patients not given that treatment, and that therefore this treatment should be adopted. Worker B, sceptically or enthusiastically, applies the same treatment to similar types of patients and has to report no such advantage. The application of the simple probability tests previously set out would have (or should have) convinced A that though his treatment *may* be valuable, the result that he obtained *might* quite likely have been due to chance. He would consequently have been more guarded in his conclusions and stressed the limitations of his data.

If, however, the test satisfied worker A that the difference in reaction that he observed between his two groups was *not* likely to be due to chance, then there comes the second, and usually much more difficult, problem. Were his two groups of patients really equivalent in all relevant characteristics except in their differentiation by mode of treatment ? This question immediately emphasises the importance of the initial planning of clinical trials with some new treatment or procedure, a point which was discussed in the first of these articles. The simple probability tests are *not* rules merely to be applied blindly at the end of an experiment, whether that experiment be well or badly carried out. Certainly they can tell us in either case whether certain observed results are likely or not likely to be due to chance ; equally certainly they can tell us nothing beyond that. But if the trials are well-planned then we can with reason infer that the "significant" difference observed between the groups is more likely to be due to the specific treatment than to any other factor, for such other factors are likely to be equally present in both groups in the well-planned test. If the trials are badly planned, in the sense that the groups to be compared are allowed to differ in various important respects as well as in treatment, then we can infer nothing whatever about the advantages of the specific treatment. The time to reach that very obvious conclusion is not at the end of the experiment, when time, labour, and money have been spent, but before the experiment is embarked upon. To argue at the end of a badly planned experiment that the statistical method is not applicable is not reasonable. The statistical method (like any other method) must fail if it has to be applied to faulty material ; but faulty material is often the product of a faulty experiment. Much thought, in fact, must be given to the devising of a good experiment, of really effective clinical trials, and the statistical aspect must be borne in mind from the start.

### The Problems of Clinical Trials

With methods of treatment the main questions to be settled are usually these :—

(*a*) How can the patients be effectively allocated to the two groups which are to be compared—which we can refer to as the treated and control groups.

(*b*) What criterion or criteria can be used as evidence of the effects of treatment.

(*c*) On how many patients will the trials have to be made to give reliable results.

The answers to these questions will, naturally, vary with the particular case at issue, but there may be some advantage in discussing them, briefly in general.

#### (*a*) ALLOCATION TO GROUPS

By the allocation of patients to the two groups we want to ensure that these two groups are alike except in treatment. It was pointed out in the first section that this might be done, with reasonably large numbers, by a random division of the patients, the first being given treatment A, the second being orthodoxly treated and serving as a control, the third being given treatment A, the fourth serving as control, and so on, no departure from this rule being allowed. It was also pointed out that this method could be elaborated, the groups being made equal in such well-defined characteristics as age and sex, and then randomly composed in other respects (and, of course, more than one form of treatment could be brought in). While the treatment to be tested has only an empirical basis—as it must have before it has been adequately tried out—there can be no serious moral objection to this procedure, though practical difficulties of administration may well arise. On the other hand, once there is evidence that one treatment gives better results than another (even though the evidence is slender) the moral problem becomes acute. One cannot treat human beings like laboratory animals and to withhold from a patient a treatment which is likely to benefit him is impossible. All the more important, therefore, is it to secure reliable evidence of the effects of a form of treatment before that

* In Sections IV and X I discussed the meaning and use of the standard deviation and the coefficient of correlation. I have been asked to show how in practice these two statistical values are calculated. I propose to do this in two further sections which will follow this concluding summary.

position arises. In the early days of a new treatment there are also likely to be some workers who regard it favourably, and others who distrust it. If a random division of patients is objected to, or is administratively impossible, it should be possible at this stage to make comparisons between similar types of patients to whom worker A is giving the treatment and worker B is not. For example, in the treatment of pulmonary tuberculosis by collapse therapy there are physicians who now believe that an artificial pneumothorax should be induced at an early stage; there must have been, and no doubt still are, many patients of similar types to whom that treatment has not been applied, who would serve as an effective standard of comparison. The difficulty is that usually any one worker's field of observation is too limited to give a convincing result, while a prolonged period of observation of each patient is also a necessity and difficult to secure. Organisation is required so that patients may be classified on a uniform system, and the results collated and judged by identical criteria. In the long run it is probable that useless forms of treatment will be discarded and the good will survive, but it may be an unfortunately long run which carefully controlled trials would have effectively shortened.

*The advantage of recording limited data.*—Even the smallest amount of data has its advantage, if collected on some uniform system and clearly defined. In some instances it is only by the accumulation of such data that an answer to a problem can be reached. For example, there is some evidence that epidemics of milk-borne and water-borne enteric fever differ in the sex- and age-incidence of the persons attacked, the former attacking women and children—the larger consumers of milk—with proportionately greater frequency. The problem cannot be settled by the evidence from any one epidemic; it requires the accumulation of data from a series of epidemics of the two types. The field of observation of any one worker is insufficient, but if uniform data of the sex and age of patients are systematically collected and published reliable evidence will eventually be reached.

*The problem of classification.*—In that particular instance the criteria for classification of patients, namely, age and sex, are simple; in grouping *types* of patients, given or not given a specific form of treatment, the task may be very much more difficult. No purely objective criteria may be available and subjective factors, variable from one worker to another, may enter in—for instance in classifying patients with cancer or pulmonary tuberculosis to the stage of disease. Can any system in each case be devised which with any worker ensures that like is being put with like, at least in broad categories? It is often said that it cannot be done, that particular problems are not susceptible to statistical analysis because patients cannot be efficiently classified before and after treatment. It is true that there are sometimes very serious difficulties in making such objective classifications but these difficulties must be faced if the problem is important. Can a clear-cut answer be reached in any other way to the fundamental questions "is this treatment of value, of how great a value, and with what types of patients?" In the large majority of cases it is difficult to see how it can. Even if the treatment is not of general value but of apparently great benefit in relatively rare isolated cases, satisfactory evidence of that must lie in statistics—viz., that such recoveries (however rare) do not occur with equal frequency amongst equivalent persons not given that treatment. Sooner or later the case is invariably based upon that kind of evidence, but in the absence of planned trials it is often later rather than sooner. If it be maintained merely in general terms that a particular type of patient fares much better under such-and-such a form of treatment, then two queries arise. If the patient can be thus defined as of this particular type why cannot he be classified and compared with the patients of similar type not specifically treated? To reach the conclusion that he has benefited from treatment he must have been compared at least mentally with his untreated prototype, and the conclusion is itself based upon statistical though unrecorded evidence. The difficulty does not seem to lie, in that case, in classifying (for the clinician has done that in drawing his conclusion) but rather in the small field of observation of any one worker and in the lack of organised trials in the earlier days of a form of treatment. It may of course be said with truth that no two patients are alike in all respects; but if that is a logical objection to classifications it is equally a logical objection to treating any patient on the basis of past experience. In medical statistics, moreover, we are not usually comparing the reactions of individuals but of broadly similar groups of individuals, and in comparing randomly chosen groups, or groups representative of a type, we can reasonably presume, if the groups are fairly large, that the distribution of unknown characters which may influence the issue is likely to be equivalent.

### (*b*) ASSESSMENT OF THE RESULTS OF TREATMENT

The second query that arises from our general statement is *how* much better do the patients fare under the particular form of treatment? How can the advantage be qualitatively or quantitatively assessed? For that purpose the criterion of success or failure must be defined, and clearly the more objective it can be made the better it will be. The criterion must, of course, vary with the problem. It is useless to use the survival-rate as an index with a disease that has an extremely low fatality-rate. Speed of recovery may be an appropriate test in one case, incidence of complications in another, absence of remission in a third, structural change in yet another, and so on. The choice of criterion and the way in which it is to be measured or defined are inherent in the question at issue and an essential part of the planning of the experiment, the clinical trials, or whatever is under discussion. The way in which it is to be recorded, the means of securing uniformity if different workers are involved, and the steps to be taken to avoid the omission of necessary items of information, must all enter into this plan in its initial stages. Team-work is often requisite and in that team I suggest (at the risk of being accused of over-emphasising the importance of my own subject) the medical statistician ought to be represented. His inclusion should have two advantages. He should be able to advise on the statistical aspects of the inquiry at its inception, and secondly, and equally important, he will learn at the start the details of the problem, the difficulties of solving it, and the factors that may complicate it. If his task is only to come in at the end, merely to make a technical analysis, he may be faced not only with material that is not capable of answering the questions posed but also with material which he may imperfectly understand, having had no previous association with it, and therefore be liable to misinterpret.

### (*c*) THE NUMBERS REQUIRED

Finally a question very frequently put to the statistician relates to the size of the sample that is

necessary to give a reliable result. To that there is usually no simple answer. If two groups are to be compared, a treated and a control group, then the size of the sample necessary to "prove the case" must depend upon the magnitude of the difference that ensues.

If, to take a hypothetical example, the fatality-rate (or any other selected measure) is 40 per cent. in the control group and 20 per cent. in the treated group, then by the ordinary test of "significance" of the difference between two proportions, that difference would be more than is likely to occur by chance with 42 patients in each group (taking twice the standard error as the level). In other words, with those fatality-rates we should have to take at least 42 patients in each group to feel at all confident in our results. If there were 50 patients in each group and 20 died in the control group and 10 in the specially treated group that difference is (on the criterion of "significance" adopted) more than would be likely to occur by chance. If, on the other hand, the improvement was a reduction of the fatality-rate from 40 to 30 per cent. we should need at least 182 patients in each group. If we had 200 in each group and 80 died in the one and 60 in the other, that difference is more than would be likely to occur by chance. Finally, if the fatality-rate was only 4 per cent. in the control group and 2 per cent. in the treated group we should require as many as 600 patients in each group to be able to dismiss chance as a likely explanation. With that number in each group there would be 24 and 12 deaths, and a difference of this order on smaller numbers might well be due to chance. (In such a case the fatality-rate, of course, might not be the best measure of the advantages of the treatment.)

The determination of the numbers required is based, it will be noted, upon the difference observed between the groups. In practice we often do not know what that difference is likely to be, until at least some trials have been made. There can be no answer given in advance to the question "how many observations must be made." Unless there is some indication from past experience as to the kind of difference that may result, or unless we can argue on a priori grounds, we must confess ignorance of the numbers required to give a convincing result.

### Common Sense and Figures

Apart from these problems of the errors of sampling, much of my discussion of the interpretation of figures has centred, it will have been noted, not so much on technical methods of analysis but on the application of common sense to figures and on elementary rules of logic. The common errors discussed in previous sections are not due to an absence of knowledge of specialised statistical methods or of mathematical training, but usually to the tendency of workers to accept figures at their face value without considering closely the various factors influencing them—without asking themselves at every turn "what is at the back of these figures? what factors may be responsible for this value? in what possible ways could these differences have arisen?" That is constantly the crux of the matter. Group A is compared with Group B and a difference in some characteristic is observed. It is known that Group A differed from Group B in one particular way—e.g., in treatment. It is, therefore, concluded too readily that the difference observed is the result of the treatment. To reject that conclusion in the absence of a full discussion of the data is *not* merely an example of armchair criticism or of the unbounded scepticism of the statistician. Where, as in all statistical work, our results may be due to more than one influence, there can be no excuse for ignoring that fact. And it has been said with truth that the more anxious we are to prove that a difference between groups is the result of some particular action we have taken or observed, the more exhaustive should be our search for an alternative and equally reasonable explanation of how that difference has arisen.

It is also clearly necessary to avoid the reaction to statistics which leads an author to give only the flimsiest statement of his figures on the grounds that they are dull matters to be passed over as rapidly as possible. They may be dull—often the fault lies in the author rather than in his data—but if they are cogent to the thesis that is being argued they must inevitably be discussed fully by the author and considered carefully by the reader. If they are not cogent, then there is no case for producing them at all. In both clinical and preventive medicine, and in much laboratory work, we cannot escape from the conclusion that they are frequently cogent, that many of the problems we wish to solve *are* statistical and that there is no way of dealing with them except by the statistical method.

A. B. H.

## AN ORTHOPÆDIC NURSING CERTIFICATE

THERE are at the present time in this country some thirty orthopædic hospitals most of which issue certificates of proficiency to their nursing staff on completion of their training. These certificates lack uniformity and offer no accepted standard when applications for other posts are being considered. The Central Council for the Care of Cripples, which, since its inauguration in 1920, has acted as a coördinating body in matters concerning the welfare of cripples, now proposes an orthopædic nursing certificate based on a uniform syllabus. In consultation with the principal orthopædic hospitals a scheme for such a certificate has been drawn up and the rules and syllabus have been issued in pamphlet form [1] by the Council. The certificate will be awarded as the result of tests held at the end of the first and second years of training respectively, but probationer nurses who have passed the preliminary State examination will be exempt from the earlier test. The first, which includes anatomy, physiology, hygiene, and practical nursing, both written and oral, will be taken at the training hospital in May or November. The second test on orthopædic conditions and their nursing will be taken partly at the hospital, but for the practical and oral portion examinees will generally be asked to attend at a centre—London, Bristol, Newcastle, or Birmingham—again in May or November. General State-registered nurses will be allowed to sit for the final examination at the end of one year's training. The entrance fee will be 10*s.* 6*d.* for the first test and one guinea for the second. There are five orthopædic surgeons on the executive committee of the Central Council one of whom, Mr. E. S. Evans, has acted as chairman of the subcommittee which drew up the scheme. Dame Agnes Hunt, who is president of the Council, expresses the hope that every orthopædic hospital which offers training to probationers will adopt the certificate, so that its possession may be generally accepted as evidence of sound training in the elements of orthopædic nursing. It is proposed to hold the first examination in November, 1937.

[1] From the secretary of the Council, 34, Eccleston-square, London, S.W.1. 2*d.*, post free.