1 The Aim of the Statistical Method

'Is the application of the numerical method to the subject-matter of medicine a trivial and time-wasting ingenuity as some hold, or is it an important stage in the development of our art, as others proclaim?' In 1921 that was a very reasonable question to be asked by the writer of an article on statistics in a medical journal. It could hardly be asked in the same setting today for, whatever may have been the reactions of medically qualified readers at that time, there has been in the ensuing years a very remarkable and increasing acceptance of the method. There has been a continual and substantial increase in the number of papers contributed to medical journals of which the essence is largely statistical in both presentation and analysis. Not only has there developed an enhanced knowledge and use of the national registers of life and death which the Registrars-General of the United Kingdom (and similar authorities in other countries) annually publish and analyse, but there is an increasing number of workers who endeavour to apply numerical methods of analysis to their own records obtained by observation, clinical trial or experiment. However, many such workers, and their readers, have had little, or no training in statistical method, and many of them may find the more mathematical methods of the professional statistician, as has been said, 'obscure and even repellent.' Often enough, indeed, the argument is put forward that the use of such mathematical methods is quite unjustifiable, that the accuracy of the original material is not sufficient to bear the weight of the treatment meted out to it. This assertion is not strictly logical. If a collection of figures is worth a statistical analysis at all, it is, obviously, worth the best form of statistical analysis – i.e. the form which allows the maximum amount of information to be derived from the data.

Whether mathematical statistical methods *are* the best form in particular cases, whether they are essential or may be regarded as an unnecessary elaboration must turn rather upon this question: Can we in any of the problems of medical statistics reach satisfactory results by means of relatively simple numerical methods only? In other words: Can we satisfactorily test hypotheses and draw deductions from data that have been analysed by means of such simple methods? The answer is undoubtedly yes, that many of the figures included in medical papers can by relatively simple statistical methods be made to yield information of value. Sometimes the yield may be rather less than that which might be obtained by more erudite methods which are not at the worker's command but the best should never be made the enemy of the good, and even the simplest statistical analysis carried out logically and carefully is an aid to clear thinking with regard to the meaning and limitations of the original records. If these conclusions are accepted, the question immediately at issue becomes this: Are simple methods of the interpretation of figures only a synonym for common sense or do they involve an art or knowledge which can be imparted? Familiarity with medical statistics leads inevitably to the conclusion that common sense is *not* enough. It seems that many people are not capable of using common sense in the handling and interpretation of numerical data until they have been instructed in quite elementary ideas and techniques. Mistakes which when pointed out look extremely foolish are quite frequently made by intelligent persons, and the same mistakes, or types of mistakes, continue to crop up again and again. There is often lacking what has been called a 'statistical tact, which is rather more than simple good sense.' That tact the majority of persons must acquire (with a minority it is undoubtedly innate) by a study of the basic principles of statistical thought and method.

The object of this book is to discuss these basic principles in an elementary way and to show, by representative examples taken from medical literature, how these principles are frequently forgotten or ignored. There is no doubt that the discussion will often appear too simple and that some of the mistakes to which space is given will be thought too futile to need attention. That such is not the case is revealed by the recurrence of these mistakes and the neglect of these elementary principles, a feature with which every professional statistician is familiar in published papers and in those submitted to him by their authors for 'counsel's opinion'.

Definition of Statistics

Whereas the laboratory worker can frequently exclude variables in which he is not interested and confine his attention to one or more controlled factors at a time, the worker in clinical or preventive medicine is often unable to experiment. He must inevitably use records which may be influenced by factors which he cannot control but have essentially to be taken into account. In this setting the essence of the statistical method lies in the elucidation of the effects of these multiple causes. We can define such statistics, therefore, as 'quantitative data affected to a marked extent by a multiplicity of causes,' and statistical method as 'methods specially adapted to the elucidation of quantitative data affected by a multiplicity of causes' (Yule and Kendall, *An Introduction to the Theory of Statistics*). More broadly we can regard statistics as 'the discipline concerned with the treatment of numerical data derived from groups of individuals' (P. Armitage, *Statistical Methods in Medical Research*).

For example, suppose we have a number of persons all of whom have been in contact with a case of infectious hepatitis and to a proportion of them is given an injection of gamma globulin. The others are observed as 'controls'. We wish to know whether the injection prevents the development of a clinical attack. It is possible that the risk of developing an attack after exposure is influenced by such factors as age and sex, social class and all that that denotes, duration and intimacy of contact, and so on. So far as is possible a statistical analysis necessitates attention to *all* such influences. We must endeavour to equalise the groups we compare in every possibly influential respect except in the one factor at issue – namely, the prophylactic treatment. If we have been unable to equalise the groups *ab initio* we must equalise them to the utmost extent by the mode of analysis. As far as possible it is clear, however, that we should endeavour to eliminate, or allow for, these extraneous or disturbing causes *when the observations are planned*; with such planning maybe we can determine not only whether the treatment is of value but whether it is more efficacious in one situation than another, at one age than another, etc. It is a serious mistake to rely upon the statistical method to eliminate disturbing factors at the completion of the work. *No* statistical method can compensate for badly planned observations or for a badly planned experiment. But a knowledge of it can contribute considerably to the design of an experiment.

Planning and Interpretation of Experiments

It follows that the statistician may be able to advise upon the statistical lines an experiment such as that referred to above should follow. Elaborate experiments can be planned in which quite a number of factors can be taken into account statistically at the same time (see, for example, *Experimental Designs* by W. G. Cochran and G. M. Cox and

Planning of Experiments by D. R. Cox). It is not the intention of this book to discuss these more complex methods of planning and analysis; attention is mainly confined to the simpler types of experimental arrangement which are so frequently required in medicine. Limitation of the discussion to that type must not be taken to mean that it is invariably the best form of experiment in a particular case.

The essence of the problem in a simple experiment is, as emphasised above, to ensure beforehand that, as far as is possible, the control and treated groups are the same in all *relevant* respects. The word 'relevant' needs emphasis for two reasons. First, it is obvious that no statistician, when appealed to for help, can be aware of all the factors that are, or may be, relevant in particular medical problems. From general experience he may well be able to suggest certain broad disturbing causes which should be considered in planning the experiment (such as age and sex in the example above), but with factors which are narrowly specific to a particular problem he cannot necessarily be expected to be familiar. The onus of knowing what is likely to be relevant in such a situation must rest more upon the experimenter, who is, presumably, familiar with that narrow field. Thus, when the statistician's advice is required it may be his task to suggest means of allowing for the disturbing causes, either in planning the experiment or in analysing the results, but not, invariably, to determine what *are* the relevant disturbing causes. At the same time no statistician who is wise will advise at all upon a medical problem with which he is quite unfamiliar. Successful collaboration demands that the statistician learn all he can of the problem at issue and the experimenter (clinician, community physician, etc.) all he can of the statistical approach. Without substantial knowledge on both sides the blind may well lead the blind.

The second point that must be observed as regards the equality of groups in all relevant respects is the caution that must attend the interpretation of statistical results. If we find that Group A differs from Group B in some characteristic, say an attack-rate, can we be certain that that difference is due to the fact that Group A was inoculated (for example) and Group B was uninoculated? Are we certain that Group A does not differ from Group B in some other character relevant to the issues as well as in the presence or absence of inoculation? For instance, in a particular case, inoculated persons might, on the average, belong to a higher social class than the uninoculated and therefore live in surroundings in which the risk of infection was less. We can never be *certain* that we have not overlooked some relevant factor or that some factor is not present which could not be foreseen or identified. It is because he knows a complex

chain of causation is so often involved that the statistician sometimes appears to be an unduly cautious and sceptical individual.

Statistics in Clinical Medicine

The essence of an experiment in the treatment of a disease lies in comparison. To the dictum of Helmholtz that 'all science is measurement,' we should add, as that great experimenter Sir Henry Dale pointed out, a further clause, that 'all true measurement is essentially comparative.' On the other hand there is a common catch-phrase that human beings are too variable to allow of the contrasts inherent in a controlled trial of a remedy. Yet if each patient is 'unique' it is difficult to see how any basis for treatment can be sought in the previous observations of other patients – upon which clinical medicine is founded. In fact, of course, physicians must, and do, base their 'treatment of choice' upon what they have seen happen before – whether it be in only two or three cases or in a hundred.

However, though, broadly speaking, human beings are not unique in their responses to some given treatment, there is no doubt that they are likely to be variable, and sometimes extremely variable. Two or three observations may therefore give, merely through the customary play of chance, a favourable picture in the hands of one doctor, an unfavourable picture in the hands of another. As a result, the medical journals become an arena for conflicting claims – each in itself, maybe, perfectly true of what the doctor saw but insufficient to bear the weight of the generalisation placed upon it.

Far, therefore, from arguing that the statistical approach is impossible in the face of human variability, we must realise that it is *because* of variability that it is often essential. It does not follow, to meet another common criticism, that the statistical approach invariably demands large numbers. It may do so; it depends upon the problem. But the responses to treatment of a single patient are clearly a statement of fact – so far as the observations were truly made and accurately recorded. Indeed that single case may give, in certain circumstances, evidence of vital importance.

If, for example, we were to use a new drug in a proved case of acute leukaemia and the patient made a complete, immediate and indisputable recovery, we should have a result of the most profound importance. The reason underlying our acceptance of merely one patient as illustrating a remarkable event – not necessarily of cause and effect – is that long and wide experience has shown that in their response to acute leukaemia human beings are ***not*** variable. In spite of modern drugs and temporary remissions they one and all fail to make complete, immediate and indisputable recoveries. Therefore, although it would clearly be most unwise upon one case to pass from the particular to the general, it would be sheer madness not to accept the evidence presented by it.

If, on the other hand, the drug were given to a patient suffering from acute rheumatic fever and the patient made a complete, immediate and indisputable recovery, we have little basis for remark. That recovery may clearly have followed the administration of the drug without the slightest probability of related cause and effect. With this disease human beings ***are*** variable in their reactions – some may die, some may have prolonged illnesses but recover eventually with or without permanent damage, some may make immediate and indisputable recoveries – whatever treatment we give them. We must, therefore, have more cases before we can reasonably draw inferences about cause and effect. We need a statistical approach and a designed experiment (the details are discussed in Chapter 20).

While, therefore, in many instances we do need larger numbers for a sound assessment of a situation, it certainly does not follow – as is sometimes asserted – that the statistician would have rejected some of the original and fundamental observations in medicine on the grounds of their small number. To take a specific example, *fragilitas ossium* was originally described on two cases and this, a later writer said, statisticians would regard as useless evidence. But why should they? If exact descriptions and illustrations were given of these two cases, then, of course, they form part of the body of scientific knowledge. They are undeniable evidence of an occurrence. What ***can*** happen, what ***does*** exist, quite regardless of the *frequency* of occurrence and irrespective of ***causation*** or association, may be observed, as already stated, even on a sample of one. It can only be in relation to an appeal from the particular to the general that a statistician – and, equally, any trained scientific worker – could object. If on the basis of the two cases the clinician, in practice, let us say, near the London meat market, should argue that the condition was specific to butchers, then one might suggest that the experience was too limited in size and area to justify any such generalisation.

In short, there is, and can be, no magic number for either clinician or statistician. Whether we need one, a hundred, or a thousand observations turns upon the setting of our problem and the inferences that we wish to draw.

It must be clear, too, that almost without statistics, and certainly without accurate measurement, the mental, or quite rough, contrasting of

one treatment (or some other course of action) with another, will give a truthful, if not precise, answer, *if* the treatment has a very real and considerable effect. Without a strictly controlled trial the merits of penicillin could not fail to come to light. Its effects would have been incompatible with past experience. With such 'winners' it is easy for the critics of the often relatively slow statistical approach to be wise after the event, and to say that the general evidence available at the start of a long, and perhaps tedious, trial made it unnecessary or pedantic. They forget the many occasions when the trial has shown that a vaunted treatment has little, if any, value – in spite of all the general 'evidence' that was available. Without a trial it might well have lingered on, to the detriment of patients. Further, it is difficult to determine through general impressions whether some drug is quite useless or of some slight but undoubted value – and to reduce, say, a relapse rate from 6 to 3 per cent would not be unimportant. It is even more difficult to determine with uncontrolled and unco-ordinated observations whether one powerful drug is more valuable than another in particular situations. Only a carefully designed clinical trial is likely to serve this purpose. But that is not by any means to say that the statistically guided experiment is the ***only*** profitable means of clinical investigation or invariably the best way of advancing knowledge. It is merely one way.

One difficulty of the variability of patients and their illnesses, is in classifying the patients into, at least, broad groups, so that we may be sure that like is put with like, both before and after treatment. But unless this problem can be solved in specific instances (and particularly with mental illnesses), clear-cut answers cannot be reached to the fundamental questions 'is this treatment of value, of how great a value, and with what types of patients?' Even if the treatment is not of general value but apparently of benefit in relatively 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 a planned approach it is often later rather than sooner. As Lord Platt has emphasised, in a more general setting, records in clinical research are likely to be disappointing 'unless they have been kept with an end in view, as part of a planned experiment. . . . Clinical experiment need not mean the subjection of patients to uncomfortable procedures of doubtful value or benefit. It means the planning of a line of action and the recording of observations designed to withstand critical analysis and give the answer to a clinical problem. It is an attitude of mind.'

Returning to the problems of classification, by the statistical process of condensation of the individual items of information into a few groups, and, further, into average and other values briefly descriptive of the data, we are clearly sacrificing some of the original detailed information. We must be particularly careful, therefore, that we sacrifice nothing relevant to the issues or more than is essential to clarity and ease in handling, interpreting and presenting the data. It is rarely feasible in practice, however, to publish the full case histories of a large number of patients specially treated and similar details of an equally large control group; and even were it feasible, that material alone cannot supply the reader (or the writer) with the information he needs until it has been appropriately condensed. The question is, was the special treatment of value, i.e. the elucidation of cause and effect? That elucidation must normally be achieved by the construction, from the original mass of recorded data, of relatively short tables, and statistical values based upon them, to show the relevant position before and after treatment of the specially treated and orthodoxly treated groups.

Other examples of the application of statistics in clinical medicine are: (*a*) observations on the natural history of a disease – what are its presenting signs and symptoms, what is its course, how variable is it in its manifestations from patient to patient, with what characteristics, e.g. age or sex, is it associated?; (*b*) the follow-up and assessment of patients treated in some particular way – particularly, perhaps, in surgery; (*c*) the definition of 'normal' – at what point does the measure of some bodily characteristic become pathological?; (*d*) the accuracy of laboratory procedures constantly used in clinical medicine, e.g. blood-counts.

In short, in statistics must lie the answers to many of the fundamental questions posed by clinical medicine – of diagnosis, of treatment and of prognosis. We need the careful collection of statistical information, its analysis by appropriate methods and its presentation in the literature. Initially it calls for precise and accurate observation of the patient. By such means the general body of medical knowledge is built up.

Statistics in the Field of Public Health

In public health work we may sometimes be concerned with a similar planning of experiments and the analysis of their results, e.g. in a test of the efficacy of a vaccine as a means of preventing an attack from, say, whooping-cough; or in measuring the effects on mother and infant of supplementing the diet of pregnant women in some particular way. Frequently, however, we have to deal in this field with statistics that

come from no deliberate experiment but that arise, and are collected, from a population living and dying in an everyday course of events. Thus we have the general death-rate of the population in a given period of time; its death-rates at particular ages – in infancy, in childhood, in the prime of life, and in old age; its death-rates from particular causes – respiratory tuberculosis, cancer, violence, etc. For some diseases – the infectious notifiable diseases – we may have figures relating to the number of attacks occurring from time to time. And so on.

The object of these statistics and the statistical methods applied to them may be regarded as twofold. On the one hand we shall use them as simple numerical assessments of the state of the public health, to show by contrasts between one place and another, or between one period of time and another, whether the death-rates of the population, for example, are relatively high or low. It is only on the basis of such evidence that we can effectively consider the problems with which preventive medicine is faced and where and when remedial measures are most needed. For instance, we find that the frequency of death in the first year of life, i.e. the infant mortality rate, becomes considerably greater as we pass down the social scale from the professional classes to the general labourer. Regardless of the factors that lead to this result, we know at least that such a problem exists and needs attention. Again, we may observe that cases of typhoid fever are more frequently notified from one type of area than from another. The cause is unknown but a problem is defined. Or, finally, the records may show that the death-rates of young children from the common infectious diseases are higher in the more crowded urban communities of a country than in the less crowded. Can we counteract that unfavourable experience of the over-crowded areas?

The initial use of such statistics, as accurately and completely collected and compiled as is possible, is therefore to *direct attention* to the problems of health or ill-health presented by the population under study. Without such figures we can have but little knowledge of the most important fields for action, and the collection and tabulation of vital statistical data are, therefore, fundamental to public health work. In other words, the certificates of death and sickness that the doctor in his daily work is required to complete are not merely ephemeral bits of paper to satisfy legal demands. They may well make serious contributions to the problems of preventive medicine. Needless to say, there will be difficulties. Diagnosis and accurate certification are not easy tasks. The aids to them and their resulting accuracy must change from time to time and vary from place to place. Every medical statistician must be aware of that. But if we await perfection we shall wait for ever. So long as we are not ignorant of the imperfections much can be learnt from these imperfect records of mortality and sickness. For instance, in spite of some undoubtedly wrong diagnoses we are perfectly well aware of the marked seasonal distribution of influenza in Great Britain and the problem that that distribution raises in the epidemiology of this disease. In spite of the errors in determining the cause of death we are aware that relatively more men than women die of cancer of the lung (in the U.K. at this time). And so on.

The second, and of course closely associated, object in the collection of such figures is the determination of the basic *reasons* for the contrasts observed. For unless we can determine those reasons, the development of effective preventive measures must obviously be hampered and may be misdirected. Why, to take the examples given above, do infants of the more impoverished classes die at a higher rate than those of the wealthier? To what extent (if any), for instance, is it due to the malnourishment of the mother and child, to what extent to overcrowding in the home and a more frequent risk of specific infections, to what extent to ignorance of how to care for the infant, or even, sometimes, to frank neglect of it?

Is a higher incidence, or an epidemic, of typhoid fever in a particular area or type of area due to a defective water supply, to milk-borne infections, or to some other form of transmission? Does the pre-school child die more readily under conditions of overcrowding because such conditions expose it to a greater risk of infection early in life, or does it succumb more easily to the infection it has acquired through factors associated with overcrowding, such as malnutrition or, possibly, lack of skilled attention in the early stages of illness?

Here the statistical method comes into play, endeavouring to disentangle the chain of causation and allowing us, sometimes, to determine the most important factors in need of correction. Since we are dealing with uncontrolled observations, often liable to errors, the task may be very difficult; the effects of the 'multiplicity of causes' often cannot be completely distinguished, e.g. the effects on health of overcrowding *per se* as apart from the features of poverty which invariably accompany it. But the original vital statistics having indicated the problem and their analysis having, at least, suggested a cause and effect, we may be able to progress further by a more deliberate collection of additional data or sometimes by a specifically designed experiment.

However that may be, good vital statistics must be the essential forerunner of the development of preventive measures designed to promote the health and well-being of any population, or of some particular fraction of it, and must serve as one of the fundamental 'yardsticks' for determining the success or failure of such measures. They are fundamental to the study of epidemiology in the modern sense of that word, i.e. the community characteristics of every disease. If such statistics are lacking, or too inexact or too incomplete to be useful, then it may be essential to seek them in special field studies on a sampling basis.

The Use of Computers

Since the mid-1950s the automatic electronic computer has taken over more and more of the calculations involved in the statistical analysis of data. It can thus relieve the worker (with or without a calculating machine) of a heavy task. This development has both advantages and disadvantages.

Among the advantages are the accuracy and reliability of computer calculation. In spite of the impression often given (gleefully) by the popular press, the computer is a very reliable machine. But just as a car, no matter how reliable, cannot be expected to behave sensibly when in the hands of an incompetent driver, so the usefulness of the results from a computer depend critically upon the competence of the programmer.

When programs have been written for specific uses and stored in a computer library, it is possible for a would-be user to request a particular program and, then, by merely supplying his numerical data, to be presented with the results of using it – even though he may not have known how to calculate those results. This is a reasonable thing to do *provided that the user has sufficient knowledge of what problem he is himself trying to solve, and, equally, of exactly what problem the computer program was devised to solve*. In short, he must have sufficient knowledge to be sure that the one is precisely relevant to the other. The use of an unsuitable program for a particular task is, alas, all too common.

The computer has also made possible the use of statistical methods that were virtually impossible in the days of hand computation – simply because of the sheer quantity of calculation involved. Such methods are, in general, beyond the scope of this book.

Perhaps the main disadvantage of computer methods is that the computer can do no more than that which it has been instructed to do. The human mind, on the other hand, when working through a collection of facts and figures, has an astonishing ability to detect some totally unexpected combination of events. By looking at, and brooding over, basic data the worker may, with a flash of inspiration, emerge with some idea or some discovery which bears no relation whatever to the original object of the work. At present nobody has any idea of how to program a computer to do that, to achieve such inspiration. Maybe this will always be so – but it is dangerous to prophesy. But the importance of studying one's basic data cannot be over-emphasized.

Summary

The statistical method is required in the analysis and interpretation of figures which are at the mercy of numerous influences. Its object is to determine whether these individual influences can be isolated and their effects measured. The essence of the method lies in the determination that we are really comparing like with like, and that we have not overlooked a relevant factor which is present in Group A and absent from Group B. The variability of human beings in their illnesses and in their reactions to them and to their treatment is a fundamental reason *for* the planned clinical trial and not *against* it. Large numbers are not invariably required and it is clear that in particular circumstances even one or two cases well observed may give information of vital importance.

Vital statistics and their analysis are essential features of public health work, to define its problems, to determine, as far as possible, cause and effect, and to measure the success or failure of the steps taken to deal with such problems. They are fundamental to the study of epidemiology.

2 Collection of Statistics: Sampling

Present-day readers of the early volumes of the *Journal of the Royal Statistical Society* would be struck by one marked characteristic. In their surveys of the state of the housing, education, or health of the population in the 1830s, it was the aim of the pioneers of that time to study and enumerate *every* member of the community with which they were concerned – the town in Lancashire, the borough of East London, the country village, whatever it may have been. That aim was frequently brought to nought by the very weight of the task. Sometimes the collection of the data was beyond their capacity in time, staff, and money; sometimes, having done their best to collect them, they were weighed down by the statistical analysis that the results demanded. In contrast, the worker today would (or should) instinctively reflect on the possibility of solving such a problem by means of sampling.

By the method of sampling he may make these, and many other, tasks not only practicable in terms of cost, personnel, speed of result, etc., but will also, quite often, render the results more, rather than less, accurate. He will, of course, be introducing an additional error, the sampling error due to the fact that he has studied only a proportion of the total. However perfect the sample may be, that is inevitable. But owing to the fact that the work of observation and recording is made so much lighter, it may well be that it can be carried out with more precision and more uniformly by a smaller number of workers and, perhaps, by more highly skilled workers. Further, with a sample of say, 1 in 10 it may be possible to pursue and complete the records for all, or very nearly all, the persons included. The attempt to enumerate the whole population may lead, through the practical difficulties, to a loss of an appreciable number of the observations required. With such an *incomplete* 'whole' population we are then, in fact, left not only with a sample but with one that raises doubts that we cannot resolve as to whether it is representative. With the completed random sample of 1 in 10 we can, on the other hand, justifiably infer the values that exist in the whole population – or, more strictly, the limits between which they are likely to lie. These estimates from a