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 properly chosen sample are adequate in nearly all circumstances. In particular, sampling methods to provide vital statistics may be specially appropriate in developing countries where total information on health aspects of the population through birth and death registration may be unobtainable.

It follows that in preparing to make a survey or setting in train the collection of statistical data to illuminate some problem the first questions that the worker must ask himself are: Precisely what data do I need? Can I investigate the problem by means of a sample? If so, how shall I set about obtaining a sufficiently large and representative sample?

Drawing a Sample

Let us suppose first of all that there is a population, or 'universe,' which can be readily sampled – whether, for example, it be of institutions in a country, houses in a town, clinical records in a hospital, or medically qualified men and women on a register. Experience has shown that an apparently quite haphazard method that leaves the choice to the worker is very unlikely to be truly haphazard. He will unconsciously pick too many (or too few) houses at the corner of the street, too many (or too few) bulky clinical files, too many (or too few) surnames beginning with a particular letter. The bias may be quite unknown either in kind or degree. But it is no less likely to be there and must be avoided. It must be avoided by setting up rules of choice, to make that choice completely random and quite free of any element of personal selection.

In their simplest form the rules give everyone in the population an independent and *equal* chance of appearing in the sample. If the individual components of the population are already numbered serially, say from 1 to 970, then the required sample can be readily drawn with the aid of tables of random sampling numbers (see pp. 305 to 312). Starting randomly at, say, Set VII (p. 311), columns 14, 15, and 16, row 12, the numbers of the 'individuals' required (whether they be houses, files, or names) are 254, 479, 704, 510, 496, etc. Any number outside the range is ignored. Similarly if any number appears for a second or third time it is ignored. The process is continued until a sample of the size regarded as sufficient has been drawn, e.g. a total of 97 numbers, a sample of 1 in 10. Every number has had an independent and equal chance of appearing and thus the sample is free from bias.

It does not inevitably follow that the sample is a 'good' sample, in the sense that it is a representative cross-section of the population. The play of chance itself must, of course, sometimes produce an unusual and,

therefore, unrepresentative picture. If the sample is large (some hundreds) it is not likely to be seriously distorted; if it is very small (20 or less) it could easily be grossly in error. The solution to that dilemma must lie principally in a larger sample, but it can also sometimes be partially found by the device of ***stratification.***

The Stratified Sample

Given sufficient knowledge of the population to be sampled we may divide it into well-defined sub-groups or ***strata*** and then draw our sample of 1 in *n* from each of these strata separately. Within each stratum the choice is still entirely random, but automatically we have ensured that the final total sample includes the right proportion of each of the strata. For example, in sampling a population of children to measure their heights and weights we might first divide it into boys and girls and, within each sex, into the age groups 5–8, 9–12, and 13–15 years. Within each of these 6 groups we would then draw our sample of, say, 1 in 8 by the method given above. The total sample will clearly be 1 in 8 and at the same time it must contain the correct proportions of boys and girls of the different age groups.

Sometimes it may be better to go one stage further and to use a different sampling fraction in the different strata. Thus, suppose the population of children to be sampled was as follows:—

Years of age	*Boys*	*Girls*
5–8	156	148
9–12	624	635
13–15	49	52

If we use the same sampling fraction, 1 in 8, throughout we may have a sufficient number of the 5–8 year group, more than we need of the 9–12 year group, and too few of the 13–15 year group – particularly if we wish to examine the measurements within each group separately. We might, therefore, choose to take 1 in 5 of the 5–8 year group, 1 in 10 of the 9–12 year group, and as many as 1 in 2 of the 13–15 year group (giving 25–30 observations in each stratum). Within each stratum the choice is still random and the chance of appearing is equal. Between the strata the chance has been allowed to vary but its level is ***known*** for each component group and it can (and must), therefore, be taken into account in reaching a figure for the total sample. Thus we might reach the following results:—

Years of age	(1) Number of children in the sample		(2) Mean height in the sample (cm)		(3)=(2)×(1) Sum of heights in the sample (cm)		(4) Sampling fraction		(5)=(3)×(4) Estimated sum of heights in population (cm)	
	Boys	Girls	Boys	Girls	Boys	Girls	Boys	Girls	Boys	Girls
5–8	31	30	115	113	3565	3390	1 in 5	1 in 5	17 825	16 950
9–12	31	32	138	135	4278	4320	20	20	85 560	86 400
13–15	25	26	157	159	3925	4134	2	2	7 850	8 268

The estimated mean height of all the children in the population will then be the sum of the column (5) above (222 853 cm) divided by the total population of children (1664), or 134 cm. In reaching this result we have, it will be seen, allowed for the unequal sampling fractions which were adopted. This is *fundamental*. A quite false result would be reached by putting together the samples as they stand without allowing for the different relative proportions of the total populations that they represent. Since the samples are not quite perfect fractions it would be slightly more accurate to 'weight' the sample means directly by the total numbers of children in the population, i.e.

$$[(115 \times 156) + (113 \times 148) + (138 \times 624) + (135 \times 635) + (157 \times 49) + (159 \times 52)] \div 1664 = (222\,462 \div 1664) = 134 \text{ cm}$$

If stratification is to be worth while it is clear that we must know, or have good grounds to suspect, that the strata differ appreciably from one another in the characteristic, or characteristics, in which we are interested, e.g. that men differ from women, that one age group differs from another, that doctors differ from lawyers. If the strata do not differ or, in other words, the population as a whole is relatively uniform, there is no point in dividing it into sub-groups. There can be no gain in accuracy in such circumstances. It is, therefore, necessary to think closely before adopting the more involved technique. And clearly it is impossible to adopt it if the population to be sampled is not defined in the necessary detail.

Sampling by Stages

Sometimes a strictly random sample may be very difficult indeed to draw and it may be more practicable to take the required sample in a series of stages (this is known as multi-stage sampling). Suppose, for example, we wished to learn the number of X-ray examinations made of all the patients entering hospital in a given week in England and Wales. It would be very difficult, if not impossible, to devise a scheme which would allow the total population of patients to be directly sampled. On the other hand it would be relatively simple to list the towns and rural districts of the whole country and randomly to draw a sample of these areas of, say, 1 in 12. Within this sample of areas all the hospitals could then be listed by name and a random sample of these be drawn, say 1 in 5. Within this sample of hospitals 1 in 4 of the patients entering in the given week could be randomly chosen for observation and recording. Thus by stages we have reached the required sample. If appropriate, stratification could be introduced at one or more stages, e.g., the areas could be sampled in broad regions and subdivided into urban and rural, the hospitals could be broadly classified and sampled according to their function, and the patients could be subdivided by their sex and age and then randomly selected.

Other Methods of Sampling

It will be seen that the use of random sampling numbers requires that the population involved be already numbered or, at least, be numbered as the required numbers are drawn. If that is not the case, one method of sampling that is *usually* effective is to start from a random number and then systematically take every *n*th name (or file, etc.). In this way, suppose that from a list of 1000 clinical case records 125, or 1 in 8, are to be drawn for study. For the starting-point a number between 1 and 8 is randomly selected, say 3. Every 8th file from that point is then drawn – 3, 11, 19, 27, 35, etc. This procedure is known as 'systematic' sampling.

It should be fully realised that in certain circumstances it can give a biased result. For example, every fiftieth house in a series of streets might conceivably produce a sample with too many corner houses and too few in the centre of the street. More generally, the population to be sampled may have some periodicity in its characteristics. The fixed interval method of sampling may then produce relatively too many high (or too many low) values according to where the interval happens to fall in relation to the periodicity.

Occasionally some other simple method may present itself. For instance, every man serving in the Royal Air Force has a service number allotted to him at entry. It would be proper to choose a 1 in 10 sample by selecting all men whose number ended in, say, an 8, or a 1 in 1000 sample by selecting all men whose number ended in, say, 345. On the other hand, to choose all the men whose surnames began with certain letters is open to grave objection. Suppose we we take the letters M, J, W, and O. In Great Britain the sample will certainly include unduly large numbers of the Scots (Mackintosh, etc.), the Welsh (Jones and Williams), and the Irish (O'Brien, etc.). Yet if we deliberately leave out the letters M, J, W, and O we shall have too few of these nationals. The method is not a good one and should be rejected.

Another procedure that is likely to be quite satisfactory in many situations is to select persons born on specified dates in any month in any year, e.g. on the 9th, 16th, and 27th, giving a sample of almost 1 in 10 (36 in 365).

With such special methods we have constantly to reflect on this question: 'In relation to the observations we seek to make, will this mode of choice bring one kind of person rather than another into the net? In short, will it result in a biased picture?' Much thought must be given to that before embarking upon a sampling scheme and particularly upon one of an unorthodox nature.

Care will be especially needed when the characteristic being measured varies in time. For instance, the prevalence of sickness, e.g. influenza, varies seasonally. We shall therefore reach a quite false answer if our sample observations do not cover the whole year. This feature of *time* often calls for most careful thought.

In conclusion, whenever a worker has adopted a sampling method to derive his observations and is presenting the results of his work, he should give the reader an exact account of how he went to work. He must state briefly but comprehensively the sampling techniques that he adopted and the degree to which he was successful in applying them (including the incidence of non-response discussed below). Without this information the reader cannot judge whether the sample is likely to be a valid one, i.e. representative and unbiased.

Non-response

One of the most difficult problems that will arise in working with a random sample is that of 'non-response.' Some of the persons included in the sample may refuse to be interviewed; some may be too ill; some, perhaps, cannot be traced; some of the children, in the example above, may be absent when we visit the school to take the measurements; even when dealing with a file of clinical case histories the information required may be missing from some proportion. Every missing 'individual' (person or item of information) detracts from the randomness of the sample. We do not know, and usually cannot know, that the individuals that we *can* include give a true picture of the total population. The absentees, whatever the cause of their non-response, may have different characteristics from those who are present. In other words the sample observed has thus become a biased sample and if the number of missing items is large it may be very seriously biased. It is for this reason that every possible effort must be made to gather into a drawn sample all those originally included in it. Indeed, one should remember that the best-laid sampling scheme is quite meaningless unless this effort is made. If there are missing individuals (and almost always there are some) then much thought must be given to them, as to whether their absence is likely to distort the sample in relation to the particular facts under study.

Sometimes, inevitably, the missing items may be numerous and it may be worth while drawing a random sub-sample upon which more intensive efforts can be made to draw in 100 per cent of the required individuals. The complete, or nearly complete, sub-sample can then be compared with the less satisfactory main sample to measure the amount of bias if any, that may exist in the latter. To give a specific example, in one inquiry into the earnings of doctors, before the advent of the National Health Service in Great Britain, nearly 6000 medically qualified men and women were approached for information. It was realised that the non-response rate would almost certainly be high – in fact it proved to be as much as 27 per cent. A small random sub-sample of 1 in 10 (or 600) was, therefore, specially drawn. This much smaller number could then be more extensively studied from available records (e.g. the nature of their speciality) and more assiduously pursued for a reply. The results in this sub-sample strongly indicated that the more extensive income figures derived from the larger but incomplete group could not be seriously at fault.

Another procedure that may sometimes reduce the tendency of non-response lies in brevity and simplicity in one's requirements. Too many, and too difficult, questions do not encourage co-operation – particularly in the approach by questionnaire. There is always a desire to learn many things at the same time and by giving way to it one may end in learning nothing because of the resulting excessively high incidence of non-response. Once again a partial solution may sometimes be sought by using sub-samples for different questions. Thus in studying in Great Britain the services given by general practitioners to their patients over a calendar year a sample of 6000 doctors was drawn. To reduce substantially the amount of work required of each, they were allocated randomly in sub-samples of 500 to one month of the year. During that one month only they were asked to keep a complete record of the number of attendances by and visits to patients. Thus for every month a fairly large sample of the total population of practitioners was available to give a measure of the services rendered in that month and a summation of the sample values would give the figure for the year. Further, each doctor was asked to carry out a relatively small task although an appreciable proportion (one-third) of the total number was used. In addition, information was sought on five other matters by randomly dividing the sub-samples of 500 into five further sub-samples of 100 each. One such group was asked to record the number of operations performed, a second group the number of injections given, a third the number of night visits paid, and so on.

Thus the demands on any individual were carefully restricted and the non-response rate proved to be very satisfactorily small (2–3 per cent). The original random sampling scheme was thereby maintained practically unimpaired. (For a full discussion of these sample surveys of the 'doctor's day and pay' made by the author see the *Journal of the Royal Statistical Society*, 1951, Series A, **114**, 1–36.)

Confidence Intervals

The methods of random sampling described above have one further advantage. When an average value or a proportion is calculated from the sample we can estimate with a given degree of probability what that value or proportion must be in the population sampled. More strictly speaking we can calculate the interval in which we can be fairly confident that the population value will lie. Thus if in a random sample of 100 observations we find the proportion of persons hard of hearing to be 18 per cent, we can be fairly confident that the proportion of such persons in the population sampled lies between 10 and 26 per cent. If that range is too wide for our purpose, then the only solution is to take a larger sample of observations. More attention will be paid to these aspects of samples in later chapters.

In conclusion, the importance of this concept of random sampling could not be more clearly emphasised than in the illuminating comment once made by a newcomer to that field:—

'The necessity of using a true random sample of the population in a survey of this nature is well known and needs no emphasis; nevertheless, it may be added that contact with such a sample provided a new experience. The actual practice of medicine is virtually confined to those members of the population who either are ill, or think they are ill, or are thought by somebody else to be ill, and these so amply fill up the working day that in the course of time one comes unconsciously to believe that they are typical of the whole. This is not the case. The use of a random sample brings to light those individuals who are ill and know they are ill, but have no intention of doing anything about it, as well as those who never have been ill, and probably never will be till their final illness. These would have been inaccessible to any method of approach but that of the random sample. Perhaps one of the deepest impressions left in my mind after conducting the survey is the fundamental importance of the random sample – unusual as it is in most medical work. It does not make for ease of working: all sorts of inaccessible personalities may be encountered, and it is more time-consuming; but the degree of self-selection imposed by the population on itself in regard to its approach to doctors inevitably gives anything other than a random sample a considerable bias. It has, however, one disadvantage in that the percentage of refusals may be high.' ('The Social Medicine of Old Age.' The Report of an inquiry in Wolverhampton made in 1948 by Dr. J. H. Sheldon, C.B.E., F.R.C.P.)

Summary

In statistical work in the different fields of medicine we are constantly studying samples of larger populations. Sometimes we shall wish deliberately to draw such a sample from the population. Although in so doing we shall introduce a sampling error (which can be estimated), we shall nevertheless often gain in precision by the greater and more skilful attention that can be given to the collection of a smaller amount of data.

The sample should be drawn by some strictly random process that gives every individual in the parent population an equal chance (or known chance) of appearing in the sample. Random sampling numbers provide such a process. If groups within the population vary widely in their relevant characteristics, it may well be advantageous first to divide the population into those groups, or strata, and then to draw a sample from each stratum appropriately. Sampling by stages may sometimes be necessary, e.g. by selecting towns within a region, hospitals within the selected towns, and patients within the selected hospitals. When these methods are impracticable a sample may sometimes be effectively derived by taking every *n*th name, or unit, from a list. This method, however, calls for careful thought.

Every effort must be made to keep non-response (or missing items) to a minimum. The most careful sampling scheme is of no value if a large proportion of the required data is not in the end obtained.