METHODS OF SAMPLING PATIENTS IN STUDIES OF GENERAL PRACTICE

G. J. DRAPER, M.A.
Department of social medicine, University of Oxford

IT IS OFTEN desirable to obtain a representative sample of the patients on a general practitioner's list either to gather information about the patients or to ascertain their opinion on various matters affecting the practice. The number of such studies being carried out by general practitioners and others appears to be increasing, and the object of this paper is to give a brief discussion of methods of obtaining statistically valid samples. It is assumed that interest centres on one practice or a number of practices. If a sample of the whole population of Great Britain or, say, that of a given county or town is required, the problems involved are quite different and are discussed in some detail in e.g. Moser (1958).

Only elementary sampling concepts and the simplest designs and methods are discussed; a few additional definitions and a description of methods and problems in obtaining samples of general practitioners are given in Bevan and Draper (1965). For a discussion of more complex ideas and sampling in general the reader is referred to Moser (1958) or Sampford (1962).

This paper is mainly concerned with the practical problem of drawing samples from various listings of patients, both those which the doctor himself may have and those which the executive council has. The executive council lists may not of course be available to the person wishing to draw the sample.

In order to make this discussion as self-contained as possible a few definitions are given before coming to the practical problems of sampling.

Definitions

The aggregate of individuals or objects from which the sample is to be drawn is referred to as the *population* (or *universe*). This population is to be regarded as subdivided into *sampling units.*

In the present case sampling units might be individual patients or families. The population might be all the patients on the lists of a partnership or all the families represented on the list of a given doctor.

Sampling frame

A *sampling frame* is a list of all the sampling units in the population. This might be the collection of medical record envelopes for all patients in the practice, or an age-sex register or a specially compiled list of all families represented on a doctor's list.

Sampling fraction

The *sampling fraction* is the proportion of sampling units contained in the population which is drawn into the sample. In *stratified sampling* (see below) this may vary between strata.

Types of sample

Only the two simplest types of random sampling will be considered *viz. simple random sampling* and *stratified random sampling.*

The defining characteristic of the former is that if a sample of size S is drawn any set of S units has the same chance of being selected; this implies, in particular, that every sampling unit has an equal chance of being chosen. The sample is fair or unbiassed in an obvious intuitive sense.

In *stratified sampling* the population of units is first divided into a number of sub-populations or *strata* such that the individuals or units in each stratum are similar to others within that stratum. Thus, for instance, a practice list may be subdivided into strata each consisting of patients in a given age-group. A simple random sample is then drawn from each stratum.

For technical reasons, as well as the obvious one that some strata may be of more interest than others, the sampling fraction will often vary between strata, i.e. sampling units in some strata may have a greater probability of entering the sample than those in others. If this is done some care has to be taken in estimating averages and percentages from the sample results since weights will have to be applied to allow for the different probabilities of selection in the different groups. Even if the sampling fractions are identical special formulae will be needed in computing, e.g. standard errors. For further details of definitions and formulae see for instance Moser (1958), Sampford (1962).

Size of sample

Basically the size of the sample required depends on two factors:

(1) The variability, in the population studied, of the values of the quantities being measured. The greater this variability the larger is the sample required to make estimates with a given degree of precision.

(2) The number of sub-classes into which the sample is to be divided in

presenting the results, e.g. two sexes, or say, six age-groups or five social classes. It is obvious that one ought to have reasonably large numbers in each individual sub-group of interest.

In general very little can be said about (1) unless information is available from previous surveys or from a pilot survey. (The way in which such information is used in determining sample size is discussed in standard texts.)

As regards (2), again no very general rules can be laid down; a very crude but reasonable rule of thumb might be that for most purposes, if one is trying to estimate the proportion of the sampled population having a given characteristic, the sample size should be at least 100 and that any sub-class should have at least about 20 in it. Since some people will be unwilling or unable to provide information the final sample achieved will almost invariably be smaller than the one intended. It seems reasonable to aim at a response rate of at least 80 per cent.

Drawing the sample

The most rigorous method of sampling is to number each sampling unit in the sampling frame and then to use a table of random numbers, as described e.g. in the books by Moser and Sampford already mentioned, to determine which units are to be included in the sample. This can be fairly time consuming and in practice *systematic sampling* is often used. Suppose that the sampling fraction is $1/r$. Then the sample is selected by counting through the list of sampling units and choosing every rth unit. r is referred to as the *sampling interval.* The starting point is determined by choosing at random a number between 1 and r. If the criterion by which the list is arranged is completely unrelated to the characteristics being studied, i.e. if the list is effectively random (this would normally be the case for instance in a list arranged alphabetically by surname, but even here there might be a tendency for certain nationalities to occur in particular parts of the list. This is unlikely to be important in most cases), then systematic sampling is virtually equivalent to random sampling. If, however, there is a trend in the arrangement of the list and this is related to the characteristic being studied, e.g. if the list is arranged in order of age, and frequency of attendance is being studied, then systematic sampling will lead to a more representative sample than would a random sample. From some points of view this may be thought desirable but it should be noticed that this vitiates the computation of measures of variability, e.g. standard errors, and inferences which use them. If for any reason there was a periodicity in the list and this was related to the sampling interval a grossly biassed sample could result. This is not likely to happen but an instance where it could is given below in the section on the use of age-sex registers.

Usually one will be given the sample size required rather than the sampling fraction. Suppose that a sample of size S is to be drawn from a population containing N units. Then the sampling interval r is N/S, e.g. if a sample of 100 patients is to be drawn from a list of 2,500 then every 25th name should be chosen. If N/S is not an integer then r is taken as the nearest integer to (or below) N/S, e.g. if a sample of size 80 is required from a list of 2,435 every 30th name should be chosen.

Use of different lists

In this section the application of the above ideas to five different types of list is discussed.

1. *Medical record envelopes arranged in alphabetical order*

If the records for a single doctor or a whole partnership are arranged in one continuous sequence in alphabetical order a simple random sample of the complete practice is most easily obtained by systematic sampling, the sampling interval r being computed as described above. If the list to be sampled is in several distinct sections (e.g. because each doctor's list is filed separately or because parts of the list are kept in branch surgeries) the sampling interval is calculated by taking N to be the grand total of patients and S to be the total sample required. A systematic sample is then taken from each component list, using this sampling interval.

A slightly more complicated problem is that in which a sample of size S is required but, for instance, children and the very old are to be excluded. One approach would be to go through the list removing cards belonging to these patients and then to sample as described above. A simpler method is as follows.

Estimate the number of patients who *are* eligible for the sample. This is now the population, of size, say, N. For a sample of size S the sampling interval r is, as before, the nearest integer to N/S. Take a systematic sample as before from the complete list of eligible and non-eligible patients, but simply discard all non-eligible names.

This will lead to a sample which should be of approximately size S but may be appreciably different from it. (If it is important that the sample should not be very much less than S the sampling interval could be taken as, say, two-thirds of N/S. This will probably lead to a sample larger than that required. The excess names can be discarded either randomly or by deleting every qth name—q being chosen so as to bring the sample down to size S.)

In a similar way a sample of *families* may be drawn by 'representing' each by the head of the household and regarding only these names as eligible for the sample.

Note. In sampling problems such as these, methods of dealing with non-eligible names which involve substituting the nearest eligible one or, for instance, taking the mother of a child as replacement for the child itself when the latter falls into the sample, will lead to biassed samples.

Stratified sampling from an alphabetically arranged practice list may most easily be carried out in two stages and is best illustrated by an example. Suppose that a practice consists of 1,200 men and 1,400 women. (It is assumed that children are not to be included in the sample). One in 30 men and one in 20 women are to be chosen for the sample, i.e. a total of 40 men and 70 women. A systematic sample is taken throughout the list using a sampling interval which will ensure that *at least* 40 men and 70 women are included. In the present case if every 15th name were chosen, children being ignored, adequate numbers should be obtained. The sexes should then be re-sampled separately so as to achieve the required sample sizes.

The methods outlined above can be adapted and applied to sampling from the other lists described below. The special features and uses for these will be considered in turn.

2. *Medical record envelopes arranged by family*

If the notes for each family are kept together the sampling unit may be taken as the set of records relating to a family. The methods described in the last section may then be applied to these sets of records to obtain a sample of families. It should be noted that this is in fact a sample of families with at least one member on the practice list. If the sample is confined to families *all* of whose members are on the list then some sets of records will not be eligible for the sample.

If a sample of individuals is required the methods are as described in Section 1, the separate record envelopes being counted as such when carrying out the sampling. (*See also* the discussion at the end of this paper.)

3. *Age-sex registers*

A simple random sample can easily be drawn from an age-sex register by systematically sampling every rth name throughout the list. Two points already mentioned under the heading of systematic sampling should be noticed here. First, the sample will be more representative than a truly random sample and hence standard errors will be over-estimated. Secondly, and more important in practice, there is one pitfall to be avoided where both sexes are to be included in the sample. Suppose that male and female patients born in a given year are arranged in two parallel columns on the same page. Then, if the sampling interval is even, it is essential to count down one column then down the other, for if the two patients on one line follow each other in the counting and these are followed by the next two and so on it is easy to see that successive patients included in the sample are likely to be of the same sex.

If a sample excluding children and old people or, say, composed entirely of women is to be drawn this is very easy with an age-sex register since only the appropriate parts of the register need to be sampled.

Similarly if a sample stratified by age and (or) sex is required then the appropriate parts of the register forming the various strata can be separately sampled.

The fourth and fifth lists from which samples may be drawn are those held by the executive councils. Each executive council has two lists of all the patients in its area.

The medical registers list the patients according to the doctor with whom they are registered.

The nominal register is a complete list of all patients in the area.

The method of arranging these lists varies from one area to another. One method is described in parts 4 and 5 of this section. The basic principle here is that patients are arranged alphabetically; however it may happen that the arrangement is according to N.H.S. numbers, and in at least one case male and female patients are segregated. It may happen in future that some medical registers are arranged in order of date of birth.

4. *The executive councils' medical registers*

The medical registers consist of a separate card index for the patients of each doctor. These are arranged in alphabetical order, except that patients living in institutions, e.g. hostels, boarding-schools, may be indexed at the end of each file under the name of the institution. For any doctor the list held by an executive council includes only the patients living in that council's area; patients registered with the same doctor but living in a different area will be indexed only in the medical register for *that* area.

The method of sampling from the medical registers is obviously the same as sampling from a doctor's alphabetically arranged medical record envelopes. The decision as to whether or not to include patients living in institutions (if these are filed separately) will depend on the purpose for which the sample is being drawn (as, of course, it will whichever list is being used). If sampling is continued straight through the complete card index the proportion of such patients falling into the sample will be more or less the same as for the rest of the practice. If it is desired to exclude such patients from the sample this should be taken into account when computing the number of patients who *are* eligible for inclusion; the sampling interval is based on this number, and sampling ceases when the lists of patients living in institutions are reached.

If, as often happens, a doctor has patients in several different executive council areas this procedure has to be carried out separately for each area, though the sampling interval is calculated on the basis of overall totals.

Stratified sample could be obtained using the method described in part 1 of this section.

One disadvantage in using these lists is that patients who have only recently registered with the doctor may be missed completely and that those who have just died or moved may be wrongly included.

5. *The executive councils' nominal registers*

The nominal register held by each executive council is a complete alphabetical list of all patients living in its area. (Thus the two types of list kept by each council contain different arrangements of the same collection of names.)

The nominal registers would be of use if one wanted a sample of people registered with *any* doctor and living in the area of a particular executive council. The methods described in part 1 again apply.

Discussion

Finally it may be useful to indicate why the above methods are preferred to other more obvious ones and to mention some of the pitfalls in each.

Alternative sampling methods

The simplest method of sampling is to choose from patients visiting the surgery either, say, 100 consecutive patients or perhaps every fifth patient until the requisite number is obtained. Either method will exclude patients who never come to the surgery. For some surveys this may be considered an advantage, but what is more easily overlooked is that frequent visitors to the surgery and patients likely to be attending at the time the survey is carried out will have an increased chance of being included in the sample or may be included more than once, thus creating a bias which may be undesirable and is certainly almost impossible to allow for in the analysis.

For similar reasons a survey based on a volunteer sample is even more to be avoided.

Difficulties in achieving a random sample

In using the lists described above two possible sources of error may be distinguished.

(1) *Inadequacies in the lists*

The lists may be incomplete or may contain names of patients who have died or moved. (A patient who has just registered may have a continuation sheet or initial acceptance card in place of a medical record envelope.) If some records have been temporarily removed because for instance the patients concerned are being visited daily they should be replaced when sampling or else sampleɑ separately. Again if a patient visits two branch surgeries and has a record at each he has an increased chance of being included in the sample, though this may not be very important in practice.

(2) *Errors in sampling method*

In part 1 of the last section a method of dealing with patients not eligible for the survey was described. An alternative method is to choose a sample in which non-eligible patients are replaced say by the next patient on the list or by another member of the same family. The second of these, and to some extent the first (which may amount to the same thing) can lead to badly biassed samples, members of large families obviously having an increased chance of being included. Even worse would be to obtain a sample of families by drawing a sample of individuals and then including in the survey any family so represented, this would grossly bias the survey towards large families. For similar reasons it would be wrong to sample individuals from a list of families by first sampling the families then choosing at random one individual from each, since members of large families then have a lower chance of being included.

(3) *Non-respondents*

In any survey some individuals will be unwilling or unable to provide information. Obtaining answers from a substitute is not a satisfactory method of dealing with non-response. Every effort should be made to obtain answers from as large a proportion of the sample as possible. It cannot be too strongly emphasized that, for example, an 80 per cent response rate from a sample of 100 will almost certainly be preferable to a 20 per cent response rate from a sample of 600, even though the latter results in a greater total of respondents.

Summary

For certain investigations in general practice a representative sample of patients is required. Some elementary sampling concepts and their application to this situation are outlined. The practical problems of drawing various types of sample are discussed.

Acknowledgements

I am grateful to Dr N. H. Silverton and Dr R. P. C. Handfield-Jones for commenting on an earlier version of this paper. My thanks are also due to the Nuffield Foundation and the Nuffield Provincial Hospitals Trust for support for research from which this paper arose.

REFERENCES

Bevan, J. M. and Draper, G. J. (1965). *Med. care,* **3,** 168.

Moser, C. A. (1958). *Survey methods in social investigation.* London. Heinemann.

Sampford, M. R. (1962). *An introduction to sampling theory.* Edinburgh. Oliver and Boyd.