A SIMPLIFIED GENERAL METHOD FOR CLUSTER-SAMPLE SURVEYS OF HEALTH IN DEVELOPING COUNTRIES

Steve Bennett,* Tony Woods,b Winitha M. Liyanagea & Duane L. Smithd

1. Introduction

In order to monitor the health status of the population and to evaluate the use and effectiveness of disease protection and control measures, up-to-date information is required. In developing countries in particular, the information needed is often provided by means of cross-sectional surveys. An example of such a survey is that developed by the Expanded Programme on Immunization (EPI) of the World Health Organization (WHO) (1, 2) to estimate the vaccination status among young children. This scheme is a type of cluster sampling, in which a sample of 30 clusters (villages or the like) is selected and 7 children of the required age are selected in each cluster. The scheme was designed to allow the estimation of vaccination status to within ±10 percentage points and achieves this aim very well (1, 3). It has been used for its intended purpose of estimating vaccination coverage in many parts of the world (1).

Such a cluster-sample design is the only practical solution for most surveys, where the idea of taking a simple random sample of individuals across the country would be impractical or impossible. The EPI design is appealing in its simplicity, and has been extended to other health surveys, where the aims were different. Sometimes the cluster-sampling scheme or the sample size have been modified to take account of the objectives of the new survey (4) but at other times the “30×7” design has been adopted uncritically. A sample size which is adequate to estimate vaccination status to within 10 percentage points will not be adequate if a more precise estimate is needed, or if a comparatively rare event like mortality is being studied. Single-stage cluster sampling may be quite unsuitable for a survey in which estimates are required for separate regions of the country.

A need for “further research into possible alternatives to the currently-used 30×7 EPI survey” has been expressed (2) and the aim of this article is to present a more general approach to the design of cross-sectional health surveys, while retaining as far as possible the simplicity of the EPI strategy.

We shall consider the sampling and statistical aspects of such surveys: the sample design and selection method, the size of the sample and the estimation of standard errors. There are many excellent textbooks which describe complex designs and appropriate formulae for their analysis (5, 6), but a certain level of expertise is needed to make the most of these, and this is often not available to workers in the field. Many of the ideas in this article have been discussed in the context of EPI surveys (2) and have been used in guidelines produced for particular surveys by WHO and other organizations (7, 8), but these may not be readily available. The monograph by Lemeshow et al. (9) covers some of these issues in detail, and many of the points made here have also been discussed recently, by Freirichs & Tar (10) and Freirichs (11), who present a practical scheme for a rapid health survey making use of microcomputers, with a more specific sampling design. Details of other practical aspects of survey methodology such as field organization, questionnaire design, etc., may be found in a number of books (12, 13).

In Section 2, we outline some of the concepts used in this article. Section 3 describes the selection of the sample and Section 4 discusses criteria of sample size. The analysis of data is described in Section 5 and some extensions to the basic design are considered in Section 6.

2. Aims and concepts

It is important in any survey to set out clearly in advance the aims of the investigation. This is particularly important in deciding the sampling strategy and the size of sample to be taken. The principal aim of the study will implicitly define the basic sampling unit or BSU (also known as the ultimate sampling unit (7), or listing unit (5)). For example, in an EPI survey the principal aim may be to measure the vaccination status of children aged between 12 and 23 months. In this case the BSU is the child aged 12-23 months: the sample size is determined in terms of numbers of these “index” children. Interviewers are instructed to visit sufficient households to achieve this number, and only to carry out interviews in households in which an index child is found. This is fine as long as the study is restricted to matters directly concerning children aged 12-23 months, but if the purpose of the survey is expanded to also ascertain for example the use of oral rehydration therapy for children aged 0-5, then the sample of such children may be unrepresentative because it will only comprise those who live in households containing a child aged 12-23 months.

Most surveys have multiple aims, and for this reason should be expected to use the household as the BSU. The only exception to this would be surveys which clearly are focused only on one specific type of individual, and do not involve other members of the household, except as they affect the individual under study. Even when this is the case, there are good reasons why the BSU should still be

---

*a Lecturer, Department of Epidemiology and Population Sciences, London School of Hygiene and Tropical Medicine, London, United Kingdom.

b Partner, LT Consultants, Reading, United Kingdom.

c Statistician, Pharmakopius International Health Care Consultancy, Goring-on-Thames, United Kingdom.

d Assistant Director for Health Programmes, Aga Khan Foundation, Geneva, Switzerland.

the household. Sample-size calculations may be carried out in terms of the number of individuals of a particular type needed, and then translated into an approximate number of households. The term “household” may be interpreted according to local conditions; a convenient definition may be “those whose food is prepared by the same person”.

For households there may exist a sampling frame, or list from which the sample may be drawn. If one does not exist, some acceptable method can usually be established for choosing households one by one. Such a sampling frame is likely not to exist for BSUs other than households. It would be rare to find health records which are complete and up-to-date that they contain the current population of children aged 12-23 months for example.

A survey will collect data on many different items, and most frequently its results will be presented in terms of rates which are the ratio of two counts. An example of this would be the estimation of usage of a health centre by children aged 5-14, which might be estimated in an appropriate sample by:

Number of children aged 5-14 in sample who have visited a health centre in the past month

Number of children aged 5-14 in sample

In a survey in which the household was the BSU, not only the numerator of this ratio (the number of children who have visited a health centre), but also the denominator (the number of children aged 5-14 in the sample), would be an unknown quantity until the survey had been carried out. Both would be different if a different sample of households had been selected.

Finally, it should be noted that we shall use the term cluster in its standard sampling sense to mean a natural grouping within the population, such as a village, district or other community, from which a subsample may be selected, and not in its EPI usage as that subsample itself. Although we talk in terms of “communities” the reader may interpret this as villages, urban blocks or enumeration districts or whatever grouping is appropriate.

3. Selecting the sample

Selection of the sample may be done in several stages: for example a country may be split into regions, a number of districts chosen from each region, a few communities from each district and a number of households from each community. However, the basic principles for deciding sample size and structure and the methods for estimating rates and their standard errors are the same. They will be demonstrated first for the simplest situation where a selection of communities is made directly within some country (or region), and estimates are obtained for that country.

The extension to several stages of sampling is straightforward and is described in Section 6. The number of communities and households to be chosen will be discussed in Section 4. Here we only discuss how the selection should be made.

Selection of clusters

The strategy used for the selection of communities is the same as that used in the EPI method. It will be necessary to have a list of all the communities in the region where the survey is to take place. Some approximate measure of the number of households in each community is also necessary. If one can assume that the mean size of household will not vary greatly from one community to another, then any general measure of community population size will do. The relative size of the communities is more important than their absolute size, so even an out-of-date census will be adequate if some allowance is made for known variations in population growth rate since then. If some communities are too small to provide an adequate sample of households, they should be combined with other neighbouring communities before making the list.

Selection of a sample of communities is then performed by sampling with probability proportional to size (PPS). As in the EPI methodology, this is carried out by creating a cumulative list of community populations and selecting a systematic sample from a random start. For example, suppose it is required to take a sample of three communities from the list of 10 communities shown in Box 1. Divide the total population of the communities (6700) by the number of communities to be selected (3) to obtain the sampling interval (6700/3 = 2233). Choose a random number between 1 and 2233. Suppose this number is 1814. This should be fitted into position in the list to identify the first community in the sample. Since 1814 lies between 1601 and 1900, community 4 will be chosen. Now add the sampling interval to the initial random number: 1814 + 2233 = 4047, and so community 6 is chosen. Add the sampling interval again: 4047 + 2233 = 6280 and community 10 is chosen.

This procedure leads to communities being selected with probability proportional to size. It is desirable if, in addition, a constant number of households is selected within each chosen community. Then, overall, each household in the population will have an equal probability of being in the sample. Such a sampling procedure is said to be self-weighting and leads to the simplified formulae for analysis given in Section 5. If some other scheme is used it is unlikely that the sample will be self-weighting, and a weighted analysis will be necessary. Even the straightforward unweighted value of a proportion taken from such a sample would be a biased estimator of the true population value.

It should be noted that in selecting a PPS sample as described above it is possible for the same community to be selected twice, if that community has a population greater than the sampling interval. This is unlikely to happen if the proportion of communities

<table>
<thead>
<tr>
<th>Community</th>
<th>Population size</th>
<th>Cumulative population size</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1000</td>
<td>1000</td>
</tr>
<tr>
<td>2</td>
<td>400</td>
<td>1400</td>
</tr>
<tr>
<td>3</td>
<td>200</td>
<td>1600</td>
</tr>
<tr>
<td>4</td>
<td>300</td>
<td>1900</td>
</tr>
<tr>
<td>5</td>
<td>200</td>
<td>1930</td>
</tr>
<tr>
<td>6</td>
<td>1000</td>
<td>4000</td>
</tr>
<tr>
<td>7</td>
<td>1600</td>
<td>5600</td>
</tr>
<tr>
<td>8</td>
<td>200</td>
<td>5800</td>
</tr>
<tr>
<td>9</td>
<td>350</td>
<td>6150</td>
</tr>
<tr>
<td>10</td>
<td>450</td>
<td>6600</td>
</tr>
</tbody>
</table>
selected is small (the sampling fraction), unless one community is very much bigger than all the others. If it should happen, the correct procedure to follow would be to select two subsamples of households from within this community. It is equally valid (though less informative) to take only one subsample and count each observation twice over. It is not appropriate to select another community instead, or to repeat the whole sampling procedure until no communities are repeated, since either of these approaches invalidates the required probabilities.

If no measure of community population sizes is available at all, it will be impossible to carry out PPS sampling, and communities must be selected by simple random sampling. In this case a fixed number of households should still be taken from each selected community, but the responses obtained will have to be weighted in the analysis (see Section 5). This will necessitate a count of the total number of households in each selected community.

Selection of households

The ideal procedure for the selection of households would be to have a list of all households in the community and to choose a selection from the list at random. If such a list does not exist, and if the community is small, then a list can be created by carrying out a quick census, or perhaps by consulting community leaders.

If this is not practicable then some means has to be used which ensures that the sample is as representative as possible. This will usually involve two stages: a method of selecting one household to be the starting point and a procedure for selecting successive households after that.

The EPI recommendation for the first household is suitable ($Z^2$): that is, to choose some central point in the community, such as the market; choose a random direction from that point, count the number of households between the central point and the edge of town in that direction, and select one of these houses at random to be the starting point of the survey.

The remaining households in the sample should be selected to give as widespread a coverage as possible of the community consistent with practicality. It is possible to follow the EPI strategy of simply going to the household whose door is nearest to the current household, but whereas this procedure is adequate for the purposes of EPI sampling (3) (where children of the right age are found only in a small proportion of households visited) it is unlikely to be adequate in general. It would be better to choose, say, the fifth nearest household, and better still to select all the households completely at random.

Some procedure needs to be adopted for dealing with dwellings which contain several households. If these are infrequent, it is best to select all the households within the selected dwelling, as this prevents households in multi-household dwellings from being underrepresented. If most dwellings contain more than one household, as for example in the compounds common in some parts of Africa, then the compound may be treated as a cluster and multistage sampling used (see Section 6).

In large communities it would be a good idea to spread the sample around by having more than one starting point in different parts of the community. This would also reduce the underrepresentation of households in the outer parts of the community inherent in having just one central starting point.

The above ideas should be seen only as suggestions. Any method which achieves a random or near-random selection of households, preferably spread widely over the community, would be acceptable as long as it is clear and unambiguous, and does not give the field worker the opportunity to make personal choices which may introduce bias. In every situation a solution should be sought which is appropriate to local conditions.

4. Sample size and precision

Precision, clustering and variability

In deciding on an appropriate sample size for a survey, one is faced with the need to strike a balance between precision and cost. Ideally, one would decide on the precision needed and calculate the sample size accordingly. In practice, however, resources are always limited and often the best one can do is to calculate what sort of precision can be achieved with the resources available. This is valuable; in particular if the achievable precision is poor then perhaps the decision should be made not to carry out the survey at all.

The precision of the estimates made from the survey will depend on the size of the sample and the amount of clustering, and the item whose value is being measured. The size of the population from which the sample is selected has little effect in practice, and may be ignored. The larger the sample, other things being equal, the more precise any estimates will be. For the same overall total sample size, however, a survey in which a large number of clusters is selected, and a few households visited in each, will give more precise results than a survey in which a larger number of households is visited in each of a smaller number of clusters. For example, a survey in which 300 mothers are interviewed will usually give more precise results than one in which 300 mothers are interviewed, but if the 300 are distributed as 50 clusters of size 6, they will give better estimates than if they were distributed as 30 clusters of size 10. In opposition to this, a larger sample size and more clusters (even if somewhat smaller) will lead to an increased workload, which in turn means increases in costs and time.

The precision of an estimate also depends on the item itself and how even is its distribution across the population. For example, suppose the overall (unknown) proportion of households with a pit latrine in the region were 40%; if the proportions in each community in the region varied very little (say from 35% to 45%) then a small number of clusters selected would give a reasonably precise estimate; if, on the other hand, the proportions in each community varied more widely (say from 0% to 80%) then one would need a considerably larger sample to be sure of obtaining the same precision. This variability is measured by the rate of homogeneity (roh) which will be discussed in detail below (6).

The usual way to measure the precision of an estimate is by its standard error. We can then construct a 95% confidence interval for the true value.
from (estimate \(-2\) standard errors) to (estimate \(+2\) standard errors). If we denote the average number of responses achieved to an item per cluster by \(b\) and the total number of responses to the item in the survey by \(n\), then the standard error of an estimated proportion \(p\) may be written in the form

\[
s = \sqrt{p(1-p)/n}.
\]

Note that this is an extension of the simpler formula used when the data are assumed to come from a simple random sample, the binomial formula

\[
s = \sqrt{p(1-p)/n}.
\]

The value of \(\sqrt{n}\) measures the increase in the standard error of the estimate due to the sampling procedure used.

\(D\) is known as the design effect and is given by

\[
D = 1 + (b-1)\text{roh},
\]

where \(\text{roh}\) is the rate of homogeneity mentioned above and \(b\) is the average number of responses to the item per cluster (see below). The value of \(D\) (or equivalently of \(\text{roh}\)) will be estimated in the light of experience of previous surveys of similar design and subject matter. Such a value may be used for guidance on sample size decisions because the current survey is carried out, but once the analysis is under way, standard errors should be calculated using the methods of Section 5. The simple formula (1) should not be used for this unless \(D\) has been evaluated anew (see Section 5).

If a survey of similar design (using the same size of sample per cluster) has been carried out previously, then for any particular item in the questionnaire the design effect may be estimated from the data of that survey by the ratio of the appropriate cluster-sample variance to the variance as if it were a simple random sample (shown in Section 5). If data from such a survey are not available, \(b\) and \(\text{roh}\) must be estimated separately as described in the following paragraphs.

\textbf{Estimating \(b\) and \(\text{roh}\)}

It makes sense to choose the number of households to be visited in each cluster on practical grounds, for example, the number that can be completed in one full day's work by a team of interviewers. It would be inconvenient to choose a cluster-sample size that would involve the interviewing team in spending parts of a day in different places.

For any given item in the survey schedule, the value of \(b\) can then be obtained. If there is one response per household then \(b\) will be equal to the number of household visits achieved in each cluster. If there is one response for, say, each child aged 12-23 months, then \(b\) will be the expected number of such children to be seen in each community.

The value of \(\text{roh}\) may be thought of as a measure of the variability between clusters as compared to the variability within clusters. In a single-stage cluster sample such as the one described here, \(\text{roh}\) is equivalent to the "intra-cluster correlation" (5); in a more complex design such as a stratified multistage survey, \(\text{roh}\) is composed of the components of variability from all stages of the design.

The value of \(\text{roh}\) will be higher for those items whose value varies more between clusters. For example, because families in the same area tend to have broadly similar socioeconomic status, variables such as "husband's occupation: clerical" will be more likely to produce the same response for two individuals in the same cluster than for individuals in separate clusters. Such socioeconomic variables may have a relatively high value of \(\text{roh}\), around \(0.20\) (14).

Demographic items such as "currently married" and measures of mortality will be hardly more likely to produce the same answer from two respondents in the same cluster than from two respondents in different clusters, and will have \(\text{roh}\) very close to 0, around \(0.02\) (14). Questions of general morbidity, such as "ill in past two weeks" may have similarly low values, but morbidity from specific infectious diseases may have much higher values, up to 0.3 (4). For questions of health-care practice and of use of health-care services such as "use of ORS for last episode of diarrhoea" or immunization coverage, responses will depend on the level of services locally and on local custom, and the value of \(\text{roh}\) may be from 0.1 to 0.3 depending on the amount of variation between communities (10, 14). Although in theory \(\text{roh}\) can take values up to 1, in practice values above 0.4 are uncommon, except for variables which are specific to the locality rather than the household, and hence clustered by definition, such as for example "health centre within 30 minutes walk". The values of \(\text{roh}\) can also be <0, particularly in stratified surveys, but usually a value <0 may be considered as being due to sampling variation and treated as 0.

These guidelines for \(\text{roh}\), based on the results of the health surveys in developing countries cited above and a review of further studies\(^1\) are necessarily vague, as there will be variability in the value of \(\text{roh}\) from country to country, from survey to survey and from item to item. One possible contributing factor to the size of \(\text{roh}\) would be poorly trained interviewers and poor supervision: variability between interviewers could result in a large increase in \(\text{roh}\). There is evidence that \(\text{roh}\) declines slowly with cluster size. In principle it would be best for a particular survey if values of \(\text{roh}\) can be taken from the results of a previous round of the same survey.

\textbf{Estimating design effect and precision}

Having selected appropriate values of \(b\) and \(\text{roh}\) for the most important items in the survey, one can then calculate the design effect \(D\) using the formula (3). Although experience is limited, it is known (14) that \(\text{roh}\) is more likely to be constant from one survey to another than is \(D\). The value of \(D\) increases with cluster sample size, for example with \(\text{roh} = 0.10\), a cluster sample size of 7 would imply a design effect of 1.8, whereas a sample of 30 from each cluster would lead to a design effect of 3.9. Use of the formula (3), however approximate, is more likely to be appropriate than the value of 2 often used for the design effect regardless of cluster size or type of item.

For example, consider a household survey in which an item of major interest is the proportion of households with a pit latrine. Suppose a reasonable workload for a team of interviewers is 30 households per cluster, and it is expected that resources will allow for about 20 clusters to be sampled. Since there will be one response per household, \(b\) will be equal

\(^1\)W.M. Liyanage, unpublished MSc thesis.