

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

Steve Bennett,^a Tony Woods,^b
Winitha M. Liyanage^c & Duane L. Smith^d

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^e) to estimate 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 practically 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 \times 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 \times 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 Frerichs & Tar Tar (10) and Frerichs (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.

^e World Health Organization. *Training for mid-level managers — evaluate vaccination coverage*. Geneva, WHO Expanded Programme on Immunization, 1988. (Document WHO/EPI/MLM/COV/88).

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 so 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:

$$\frac{\text{Number of children aged 5-14 in sample who have visited a health centre in the past month}}{\text{Number of children aged 5-14 in sample}}$$

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 (6 700) by the number of communities to be selected (3) to obtain the sampling interval ($6\,700/3 = 2\,233$). Choose a random number between 1 and 2 233. Suppose this number is 1 814. This should be fitted into position in the list to identify the first community in the sample. Since 1 814 lies between 1 601 and 1 900, community 4 will be chosen. Now add the sampling interval to the initial random number: $1\,814 + 2\,233 = 4\,047$, and so community 6 is chosen. Add the sampling interval again: $4\,047 + 2\,233 = 6\,280$ 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

Box 1. A cumulative list of community sizes

Community	Population size	Cumulative population size
1	1 000	1 000
2	400	1 400
3	200	1 600
4	300	1 900
5	1 200	3 100
6	1 000	4 100
7	1 600	5 700
8	200	5 900
9	350	6 250
10	450	6 700

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 (2^a): 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 200 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

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

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

$$s = \sqrt{[p(1-p)/n]}. \quad (2)$$

The value of \sqrt{D} 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

$$D = 1 + (b-1)roh, \quad (3)$$

where 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 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 before 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 roh must be estimated separately as described in the following paragraphs.

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

The value of 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 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 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 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 roh may be from 0.1 to 0.3 depending on the amount of variation between communities (10, 14). Although in theory 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 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.

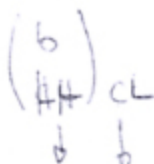
These guidelines for roh , based on the results of the health surveys in developing countries cited above and a review of further studies¹ are necessarily vague, as there will be variability in the value of roh from country to country, from survey to survey and from item to item. One possible contributing factor to the size of roh would be poorly trained interviewers and poor supervision: variability between interviewers could result in a large increase in roh . There is evidence that roh declines slowly with cluster size. In principle it would be best for a particular survey if values of roh can be taken from the results of a previous round of the same survey.

Estimating design effect and precision

Having selected appropriate values of b and 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 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 $roh = 0.10$, a cluster sample size of 7 would imply a design effect of 1.6, 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

¹W.M. Liyanage, unpublished MSc thesis.



to 30, and $n = 30 \times 20 = 600$. If we have some idea of the proportion p in advance, we should use it in the formula, but if not it is best to use $p = 0.5$ as a guess since this maximizes s and hence errs on the safe side. The value of roh is hardest to estimate, but is likely to be high, with more variation in such an item between communities than within each community, so we may take $roh = 0.20$. Using the formula (3) we obtain a design effect of

$$D = 1 + (29 \times 0.20) = 6.8$$

and from (1) the estimate of the standard error is

$$s = \sqrt{0.5 \times 0.5 \times 6.8 / 600} = 0.05$$

or 5%. This indicates that with such a sample size we can be 95% certain that the true proportion of households with latrines will lie within $\pm 10\%$ (2 standard errors) of our estimate. Whether or not this precision is adequate depends on the purpose of our survey. If the design effect had been ignored, we would have predicted a standard error of

$$s = \sqrt{0.5 \times 0.5 / 600} = 0.02,$$

encouraging us to believe that our survey would give much more precise results than would actually be the case.

Suppose that in the same survey we also wished to estimate the proportion of children aged 12-23 months who had been adequately vaccinated by their first birthday. If we could assume that such children are found in about one-quarter of all households, then we would expect to get about 7 responses from each cluster, and we would take this as the value of b . The values of n would be $7 \times 20 = 140$. We might take the value of roh to be 0.10 and following the above calculations would obtain $D = 1.6$ and $s = 5.3\%$, giving a 95% confidence interval of about 11%. Ignoring D would have led us to underestimate the width of the confidence interval as 8%.

Estimating sample size

If the investigator knows that a certain precision is required from the survey, then the necessary sample size may be calculated. Usually it will be a matter of deciding how many cluster samples of a given size b will be necessary. The design effect D should be calculated from (3) as before, and then the number of clusters necessary is given by c where

$$c = \frac{p(1-p)D}{s^2b} \quad (4)$$

For example if p is expected to be around 20% for some measure of disease prevalence, for which we expect roh to be about 0.02, and suppose that we wish to estimate p to within $\pm 5\%$. If we expect to have 20 responses from each cluster, then the value of D will be 1.38 (from (3)). For a confidence interval of $\pm 5\%$ we shall need $s = 0.025$, then from (4) we need $c = 18$ clusters.

If we had failed to take account of the design effect we would have estimated the sample size from equation (4) as 13 clusters. Using equation (1), we see that our result would then have had a predicted standard deviation of 0.029 and a confidence interval of $\pm 6\%$, a little less precise than desired. The small size of the loss of precision in this example is due

only to the small value of D . In many cases, D will be considerably larger, and the precision achieved considerably less than desired. In general, ignoring the design effect in estimating the sample size required will lead to confidence intervals which are wider than desired by a factor of \sqrt{D} .

Such calculations should be made for the most important items in the survey schedule. Ideally c should be chosen to be the largest value given by these calculations in order to satisfy all the requirements. If the sample sizes necessary for different items are grossly different (as may happen in a study which covers both disease prevalence and usage of health-care facilities), it may be advisable to just use a subsample for those questions requiring fewer responses. However, the increase in complexity of the instructions given to interviewers means that this should be used with caution.

One should note that if the prevalence of an item under consideration is expected to be quite low, for example HIV seropositivity which may in some countries be around 2%, then it is not sensible to design a survey to achieve an absolute precision of 5%. In such a case the standard error desired needs to be considered relative to the expected prevalence rate, and would be much smaller, say 0.5% in absolute terms.

If the survey has been stratified (see Section 6) then each stratum should be considered as a separate survey, and sample-size calculations performed for each one to give the precision necessary for that stratum. The precision of the overall national estimate will then be somewhat better than that for any single stratum.

If the survey is one of a series, and the purpose is to estimate the change in some measure since the previous survey, then one needs to estimate the standard error of the change. This will be larger than the standard error of the new estimate of the measure, because of the imprecision of the estimate of the measure from the previous survey. To allow for this, the sample size may need to be double that calculated by the usual methods.

5. Analysis of data

This section describes the methods used to provide estimates of proportions or rates, together with standard errors of those estimates so that confidence intervals can be calculated. A mean value may also be estimated in the same way. We also describe how to calculate D and roh . The methods described below can be carried out on a simple calculator having a square-root key, and the use of a spreadsheet is illustrated in the Annex. The calculations in this and earlier sections may also be programmed easily on a computer using a spreadsheet package, as shown by Frerichs (11).

Estimation of a proportion

Suppose that a number of households have been selected in each of c communities with a view to estimating (by examining their record cards) what proportion of children aged 12-17 months were fully vaccinated on their first birthday. Suppose that in the i^{th} community ($i=1, \dots, c$) these were x_i children whose record cards were examined, and that y_i of these were fully vaccinated as defined by the study.

Then the proportion of children in the p^{th} community who were fully vaccinated will be given by

$$p_i = y_i/x_i.$$

In the survey population as a whole the proportion who are fully vaccinated will be estimated by

$$p = \Sigma y_i / \Sigma x_i \quad (5)$$

i.e. the total number of children vaccinated divided by the total number of children whose cards were examined. This is the straightforward ratio of the sample totals. Note that it is not the same as the average of the p_i 's, which would be incorrect since it does not take account of the variation in the x_i 's.

The standard error, s , of p is obtained from the formula

$$s = [c/\Sigma x_i] \sqrt{[\Sigma y_i^2 - 2p\Sigma x_i y_i + p^2 \Sigma x_i^2] / [c(c-1)]}. \quad (6)$$

A spreadsheet for calculation of s is given in the Annex with an example of its use. This formula is more complex than the formula (2) usually used by standard computer packages in that it takes account of (i) the clustering of the sample and (ii) the variability between clusters of the denominator x_i . This value, (the number of record cards examined in the p^{th} community) will have been unknown before the survey began and would probably be different if a different sample of households were taken from the same community. Failure to take account of these factors would lead to underestimation of s , and consequent overconfidence in the precision of the results (see Annex for an example). In many cases x_i will not vary much between communities, for example when x_i is the number of households selected, and then the simpler formula

$$s = \sqrt{[\Sigma (p_i - p)^2 / (c-1)]} \quad (7)$$

may be used instead of (6).

Estimation of means

At times one will collect data on values which are not simply "yes/no" attributes of the household or person, but counts or other measurable quantities, for example "number of children ever born" or "number of rooms". In this case one may wish to estimate the mean value over the population, for example the mean number of children ever born (although of course one may also estimate a proportion, for example the proportion of women who have given birth to more than 3 children). Estimation of the mean and its standard error are carried out in exactly the same way as for a proportion (Section 5) except that y_i will now be equal to the sum of the numbers of children ever born to all of the x_i mothers interviewed in the p^{th} community.

Weighted analysis

In many situations there will be a need to weight the observations to allow for different probabilities of selection or different levels of non-response. For example suppose clusters were chosen with PPS as in Section 3, and it was intended to visit 25 households in each one, but because of staff illness it was only possible to visit 16 households in one of the clusters. If this fact is ignored, it will lead to that cluster being underrepresented in the calculation of the proportion p and its standard error. The solution is to weight the responses from this community by multiplying them up by 25/16. In more

general terms, this means replacing x_i and y_i each time they occur in formulae (5) and (6) with $w_i x_i$ and $w_i y_i$, giving the more general formulae

$$p = \Sigma w_i y_i / \Sigma w_i x_i$$

and

$$s = [c/\Sigma w_i x_i] \sqrt{[\Sigma w_i^2 y_i^2 - 2p \Sigma w_i^2 x_i y_i + p^2 \Sigma w_i^2 x_i^2] / [c(c-1)]},$$

where w_i is the weight attached to the p^{th} cluster. An unweighted cluster has $w_i = 1$.

If clusters are sampled with probability proportional to size and x_i represents the number of BSUs (households) selected, then the proportion is estimated by \hat{p} , the average of the p_i 's, and we can use formula (7) for its standard error with p replaced by \hat{p} . In other cases the approximate formula (7) ignores the size of the cluster and should not be used if weighting is necessary.

Weighting may also be used to allow for clusters not being selected with probability proportional to size, for example when current size was not known at the time of their selection and they were selected with simple random sampling (or with probability proportional to a poor or very out-of-date measure of size). In this case the weight will be proportional to the actual population of the cluster (or the ratio of this to its old estimate).

Estimation of design effect

The results of any survey may be used to estimate design effects, for use in the same or future surveys. The design effect is estimated by

$$D = \frac{s^2 \text{ from equation (6) or (7)}}{s^2 \text{ from equation (2)}}$$

The rate of homogeneity, roh , may then be estimated as

$$(D - 1) / (b - 1)$$

where b is as defined earlier. An example is given in the Annex.

Imputation of standard errors

In a large survey it may not be feasible to use the correct formulae (6) or (7) to estimate the standard error of every variable. In such a case one may calculate exact standard errors for a few variables of each type (socioeconomic, health status, etc.). Dividing each standard error by the corresponding binomial value (2) gives a new estimate of the design factor (the square root of the design effect \sqrt{D}). For the remaining variables of the survey the simple formula (2) as given by calculator or standard software can be used, and just multiplied by the most appropriate value of \sqrt{D} obtained for variables of similar type.

6. Extensions

The previous sections describe cluster-sampling procedures in a simple context: a sample of communities is selected from the whole region under consideration and a sample of households is visited in each selected community. Such a sampling scheme will be inadequate if the region is very large or if separate estimates are needed for different

geographical areas. In this section we show how the techniques described above can be extended to allow for multistage sampling and stratification.

Multistage sampling

In a large region or country where an overall estimate is required, it will usually be sensible to select the sample of communities in at least two stages. For example, if the country is split into a number of administrative districts one would take a sample of districts by the systematic PPS method described in Section 3 (i.e. by making a list with cumulative population sizes). Within each selected district, communities would be selected, again by the systematic PPS method. The same number of communities must be selected in each district. If some districts are very small it may be sensible to combine them. Households would be selected in the usual way, with again the same number selected in each community.

With the systematic PPS method described here it is possible that the same district may be selected twice. This will happen if the population of the district is larger than the sampling interval. In this case two independent samples of communities should be selected from this district.

Decisions on the sample size will be made exactly as in Section 4, except that b will now be the expected number of responses per district and c will be the number of districts in the sample. The value of roh is now an indicator of the ratio of between-district variances to within-district variances. In theory, this requires an estimate of roh from a survey of similar multistage design. In practice, such estimates are not available, and the best one can do is probably to use the values given in Section 4 as guidelines, and bear in mind that they will be overestimates, as the value of roh is likely to decline slowly with the size of the primary cluster used.

The analysis will follow exactly the same pattern as in Section 5 except that x_i and y_i now refer to the number of responses and the number of positive responses respectively in the p^{th} district, summed over all communities selected in that district.

The method of sampling described here may be extended in exactly the same way to more stages if required.

Stratification

It may be necessary to obtain separate estimates for, say, the urban and rural sectors of the population, or for different provinces or ecological zones. Each province (etc.) will be a stratum, and a sample should be selected independently from each stratum. The sample size and structure for each stratum should be chosen with the conditions and needs of that stratum in mind, as if a separate survey were being carried out in that stratum alone. The samples may be of a different type and/or size for each stratum.

An estimate for each stratum may be calculated together with its standard error by treating each stratum as a separate survey. A stratified estimate for the whole country may then be calculated by weighting the stratum estimates by the stratum populations. For example, suppose there are three strata and the estimates from them are p_1 , p_2 and p_3 , with

standard errors s_1 , s_2 and s_3 respectively. Then the estimate for the whole country would be

$$p = V_1 p_1 + V_2 p_2 + V_3 p_3$$

with standard error

$$s = \sqrt{[V_1^2 s_1^2 + V_2^2 s_2^2 + V_3^2 s_3^2]}$$

where V_1 is the proportion of the country's population which belongs to stratum 1, and so on ($V_1 + V_2 + V_3 = 1$). The standard error s for the national estimate will be somewhat less than the standard errors for the individual strata.

Implicit stratification

Stratification usually leads to a small reduction in the standard error of the overall estimate p , compared to the error that would have been obtained if the survey had not been stratified. Another way of obtaining such a reduction is by *implicit stratification*. This is simply carried out at the time of selection of communities (or districts) by ensuring that the list of communities from which the systematic sample is to be taken is ordered by some measure which is correlated with the main purpose of the survey. For example, in a survey of the utilization of mother-and-child health facilities, there may have been a previous study carried out some years ago on the same subject, or there may be other knowledge available which indicates which communities may be expected to have high levels of utilization and which communities low levels. If not, one may be able to guess that those communities which are further from the regional capital, or which cover a more widely-scattered population, will have lower levels of utilization than others. Whatever the measure chosen, if the communities can be listed in approximate order from a high to a low level of expected utilization, then the sample selected will contain communities with a spread of utilization levels, and the estimated proportion p will be more precise. The standard error will be reduced, and its estimate s given by (6) will be somewhat of an overestimate (15). The improvement in precision cannot be quantified adequately to allow its use in sample-size calculations.

7. Conclusion

A simplified approach to survey design has been presented, with no attempt to cover all possible types of estimation. We have rather aimed to provide a set of guidelines which will enable the practitioner to plan a survey in a way which will give a reasonably representative sample, without any great bias, and of a suitable size to give adequate precision without wasting resources. The values given for the rate of homogeneity have of necessity been approximate, but variability between surveys and between variables is such that precise advice is impossible. The methods of analysis presented here offer an improvement on the common practice of assuming that the data came from a simple random sample and using the standard errors given by a calculator or standard computer package.

Acknowledgements

This article had its origins in work carried out for WHO/SHS on sampling in primary health care reviews. We are grateful to S. Lwanga, S.A. Sapirie, J. L. Tulloch and J. E. Dowd of the World Health Organization for their contributions, and to many colleagues for helpful comments.

SUMMARY

General guidelines are presented for the use of cluster-sample surveys for health surveys in developing countries. The emphasis is on methods which can be used by practitioners with little statistical expertise and no background in sampling. A simple self-weighting design is used, based on that used by the World Health Organization's Expanded Programme on Immunization (EPI). Topics covered include sample design, methods of random selection of areas and households, sample-size calculation

and the estimation of proportions, ratios and means with standard errors appropriate to the design. Extensions are discussed, including stratification and multiple stages of selection. Particular attention is paid to allowing for the structure of the survey in estimating sample size, using the design effect and the rate of homogeneity. Guidance is given on possible values for these parameters. A spreadsheet is included for the calculation of standard errors.

RÉSUMÉ

Méthodes générales simplifiées pour les enquêtes sanitaires utilisant le sondage par grappes dans les pays en développement

Cet article présente des directives générales concernant l'exécution d'enquêtes sanitaires utilisant le sondage par grappes dans les pays en développement. L'accent y est mis sur des méthodes utilisables par des praticiens peu spécialisés en statistique et sans formation de base en matière de sondages. Ces méthodes font appel à un plan comportant un dispositif d'auto-évaluation simple, inspiré de celui qui est utilisé par le Programme élargi de vaccination (PEV) de l'Organisation mondiale de la Santé. Les sujets traités couvrent le plan de sondage, les méthodes de sélection aléatoires des zones et des

ménages, le calcul de la taille des échantillons et l'estimation des proportions, des taux et des moyennes, avec les erreurs types appropriées au plan. L'article traite aussi de questions telles que la stratification et les différentes étapes de la sélection. On insiste sur l'importance de tenir compte de la structure des enquêtes pour estimer la taille des échantillons, en utilisant l'«effet de plan» et le taux d'homogénéité. Des conseils sont donnés sur les valeurs qu'il serait possible d'attribuer à ces paramètres. Une feuille de calcul pour les erreurs types est jointe.

ANNEX

Estimating the standard error of a ratio and its design effect

The use of a simple spreadsheet for the calculation of an estimate and its standard error using the precise formula (6) is demonstrated using the following example. The use of the approximate formula (7) for the standard error is also shown, and the design effect is calculated. The sample size is much smaller than those encountered in practice but all the important steps in the calculation are demonstrated.

Six communities are selected using the systematic PPS procedure. Twenty households are chosen in each community in order to estimate, for the population, the proportion of recently-pregnant mothers who have received postnatal care.

The data are:

Community	Number of recently-pregnant women	Number receiving postnatal care
1	2	2
2	7	5
3	4	3
4	6	3
5	4	1
6	3	0

The spreadsheet is constructed as follows:

	y_i	x_i	y_i^2	x_i^2	$x_i y_i$	p_i
	2	2	4	4	4	1.00
	5	7	25	49	35	0.70
	3	4	9	16	12	0.75
	3	6	9	36	18	0.50
	1	4	1	16	4	0.25
	0	3	0	9	0	0.00
Total	A=14	B=26	C=48	E=130	F=73	

Here $c=6$ is the number of communities; y_i is the number of recently-pregnant mothers in the i^{th} community who have received postnatal care; x_i is the number of recently-pregnant mothers in the sample from the i^{th} community.

The estimated proportion is

$$p = A/B = 0.5385.$$

The standard error s , as given by (6), is calculated as follows:

New quantity	Calculated as	Value
p^2	$p \times p$	0.2900
G	$2 \times p \times F$	78.621
H	$p^2 \times E$	37.70
J	$C - G + H$	7.079
K	$J / [c \times (c-1)]$	0.2360
L	\sqrt{K}	0.4858
s	$c \times L / B$	0.1121

The 95% confidence interval for the true proportion is $0.5385 \pm (2 \times 0.1121)$, i.e. 0.314 to 0.763.

The approximate formula (7) gives $s = 0.1482$. The difference between this figure and that given above arises because the x_i 's are very variable.

The standard error assuming a simple random sample is given by (2) as

$$s_{SRS} = \sqrt{\{(0.5385) \times (1-0.5385)/26\}} = 0.0978,$$

thus ignoring the design of the study would have led us to assign our estimate a confidence interval from 0.343 to 0.734, which is 13% narrower than the correct value.

The design effect is estimated as

$$D = s^2/s_{SRS}^2 = (0.1121)^2/(0.0978)^2 = 1.31.$$

Since $b = \sum x_i/6 = 4.333$, roh may be estimated in this case by $(D - 1)/(b - 1) = 0.093$.

REFERENCES — RÉFÉRENCES

- HENDERSON, R. H. & SUNDARESAN, T. Cluster sampling to assess immunization coverage: a review of experience with a simplified sampling method. *Bulletin of the World Health Organization*, **60** (2): 253-260 (1982).
HENDERSON, R. H. & SUNDARESAN, T. Sondage par grappe pour évaluer la couverture vaccinale: examen d'une expérience sur une méthode simplifiée de sondage [résumé]. *Bulletin de l'Organisation mondiale de la Santé*, **60** (2): 259-260 (1982).
- LEMESHOW, S. & ROBINSON D. Surveys to measure programme coverage and impact: a review of the methodology used by the Expanded Programme on Immunization. *World health statistics quarterly*, **38** (1): 65-75 (1985).
LEMESHOW, S. & ROBINSON, D. Enquêtes sur la couverture et l'impact des programmes: méthodes quantitatives utilisées par le Programme élargi de vaccination. *Rapport trimestriel de statistiques sanitaires mondiales*, **38** (1): 65-75 (1985).
- LEMESHOW, S. ET AL. A computer simulation of the EPI survey strategy. *International journal of epidemiology*, **14**: 473-481 (1985).
- ROTHENBERG, R. B. ET AL. Observations on the application of EPI cluster survey methods for estimating disease incidence. *Bulletin of the World Health Organization*, **63** (1): 93-99 (1985).
ROTHENBERG, R. B. ET AL. Utilisation des méthodes de sondage par grappes du PEV pour estimer l'incidence d'une maladie: observations [résumé]. *Bulletin de l'Organisation mondiale de la Santé*, **63** (1): 97-98 (1985).
- COCHRAN, W. G. *Sampling techniques*. Chichester, John Wiley & Sons, 1977.
- KISH, L. *Survey sampling*. London, John Wiley & Sons, 1965.
- EL BINDARI-HAMMAD, A. & SMITH, D. L. *Primary health care reviews, guidelines and methods*. Geneva, World Health Organization, 1989.
- VERMA, V. *The estimation and presentation of sampling errors*. London, World Fertility Survey, 1982. (Technical Bulletin No. 11).
- LEMESHOW, S. ET AL. *Adequacy of sample size in health studies*. Chichester, John Wiley & Sons, 1990.
- FRERICHS, R. R. & TAR TAR, K. Computer-assisted rapid surveys in developing countries. *Public health reports*, **104**: 14-23 (1989).
- FRERICHS, R. R. Simple analytic procedures for rapid microcomputer-assisted cluster surveys in developing countries. *Public health reports*, **104**: 24-35 (1989).
- CASLEY, D. J. & LURY, D. A. *Data collection in developing countries*. Oxford, Oxford University Press, 1981.
- ABRAMSON, J. H. *Survey methods in community medicine*. Edinburgh, Churchill Livingstone, 1984. (Third edition).
- KISH, L. ET AL. *Sampling errors for fertility surveys*. London, World Fertility Survey, 1976. (Occasional paper No. 17).
- KISH, L. & HESS, I. On variances of ratios and their differences in multistage samples. *Journal of the American Statistical Association*, **54**: 416-446 (1959).