WORLD HEALTH ORGANIZATION

ORGANISATION MONDIALE DE LA SANTE

DISTR.: LIMITED
DISTR.: LIMITEE

ESM/CONS. RA/90.5

ENGLISH ONLY

CONSULTATION ON EPIDEMIOLOGICAL
AND STATISTICAL METHODS OF
RAPID HEALTH ASSESSMENT

Geneva, 26-29 November 1990

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

S. Bennett[1], A.J. Woods[2], W.M. Liyanage[3] and D.L. Smith[4]

1. London School of Hygiene and Tropical Medicine, UK
2. LT Consultants, Reading, UK
3. University of Colombo, Sri Lanka
4. World Health Organization, Geneva, Switzerland

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 of Immunization (EPI). Extensions are discussed, including stratification and multiple stages of selection. 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. 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.

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)[1,2,3] to estimate vaccination status among young children. This scheme is a type of cluster sampling, in which a sample of thirty clusters (villages or the like) is selected and seven children of the required age are selected in each cluster. This scheme was designed to allow the estimation of vaccination status to within plus or minus ten percentage points and to a large extent achieves this aim.[2,4] The scheme has been used for its intended purpose of estimating vaccination coverage in many parts of the world.[2]

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 hopelessly impractical. 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[5] but at other times the '30x7' design has been adopted uncritically. A sample size which is adequate to estimate vaccination status to within ten 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 sample 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 30x7 EPI survey' has been expressed[3] and the aim of this paper 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.

The substance of this paper is not, for the most part, original, but is not readily available elsewhere in general terms, although many of the ideas in sections 3 and 5 have been discussed in the context of EPI surveys[3] and have been used in guidelines written for particular surveys by WHO and other bodies[1,6,7,8]. There are many excellent textbooks which describe complex designs and appropriate formulae for their analysis[9], but a certain level of expertise is needed to make the most of these, and this expertise is often not available to workers in the field. Also, the information contained in these books is not usually expressed in the context of household health surveys.

We shall consider mainly 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. We shall also look at some possible extensions to the basic design. The more general aspects of survey methodology may be found elsewhere[10,11,12].

It should be noted that we are considering here surveys whose aim is principally descriptive, the estimation of rates and proportions rather than the modelling of relationships or the testing of hypotheses.

In Section 2 of this paper we outline some of the concepts used in the remainder of the paper. Section 3 describes the selection of the sample and Section 4 discuses 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 principle aim of the study will implicitly define the '<u>basic sampling unit</u>' or 'bsu' (also known as the 'ultimate sampling unit'[7]. 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, say, the use of oral rehydration therapy for children aged 0 - 5, then the sample of such children may be unrepresentative because it will only contain 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 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.

For households there may exist a <u>sampling frame</u>, 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 bsu's 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 <u>rates</u> 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. This variability in the denominator will diminish the precision that we can assign to our estimate of the rate.

Finally, it should be noted that we shall use the term '<u>cluster</u>' in its standard sampling sense to mean a grouping within the population, such as a village or district, from which a subsample may be selected, and not in its EPI usage as that subsample itself. Although we talk in terms of 'villages' the reader may interpret this as urban blocks or enumeration districts or whatever grouping is appropriate. The term 'household' may also be interpreted according to local conditions; a convenient definition may be 'those whose food is prepared by the same person'.

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 villages from each district and a number of households from each village. 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 villages 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 villages and households to be chosen will be discussed in Section 4. Here we only discuss how the selection should be made.

3.1 Selection of clusters

The strategy used for the selection of villages is the same as that used in the EPI method. It will be necessary to have a list of all the villages in the region where the survey is to take place. Some approximate measure of the number of households in each village is also necessary. If one can assume that the mean size of household will not vary greatly from one village to another, then any general measure of village population size will do. The relative size of the villages 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[3].

Selection of a sample of villages 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 village populations and selecting a systematic sample from a random start. For example, suppose it is required to take a sample of three villages from the list of ten villages shown in Table 1. Divide the total population of the villages (6700) by the number of villages to be selected (three) 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 village in the sample. Since 1814 lies between 1601 and 1900, village 4 will be chosen. Now add the sampling interval to the initial random number: 1814 + 2233 = 4047, and so village 6 is chosen. Add the sampling interval again: 4047 + 2233 = 6280 and village 10 is chosen.

This procedure leads to villages being selected with probability proportional to size. It is desirable if, in addition, a constant number of households is selected within each chosen village. 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 village to be selected twice, if that village has a population greater than the sampling interval. This is unlikely to happen if the proportion of villages selected is small (the sampling fraction), unless one village 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 village. It is equally valid (though less informative) to just take one subsample and count each observation twice over. It is not appropriate to select another village instead, or to repeat the whole sampling procedure until no villages are repeated. Either of these approaches invalidates the required probabilities.

If no measure of village population sizes is available at all, it will be impossible to carry out pps sampling, and villages must be selected by simple random sampling. In this case a fixed number of households should still be taken from each selected village, but the responses obtained will have to be weighted in the analysis (see Section 5.3).

3.2 Selection of households

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

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 succeeding households after that.

The EPI recommendation for the first household is suitable:[1,3] this is to choose some central point in the village, 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 village 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[4] (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 under-represented. 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 multi-stage sampling used (see section 6).

In large villages it would be a good idea to spread the sample around by having more than one starting point in different parts of the village. This would also reduce the under-representation of households in the outer parts of the village 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 village, would be acceptable. In every situation a solution should be sought which is appropriate to local conditions.

4. Sample size

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 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 fifty clusters of size six, they will give better estimates than if they were distributed as thirty clusters of size ten. 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 in 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 village 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 village 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 <u>rate of homogeneity</u> (roh) which will be discussed in detail below[13].

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 minus two standard errors) to (estimate plus two 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

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

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

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

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 <u>design effect</u> and is given by

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

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 5.3).

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 questionnairethe 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 5.4). If data from such a survey are not available, b and roh must be estimated separately as described in the following paragraphs.

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 village.

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 that described here, roh is equivalent to the "intra-cluster correlation"[9]; 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 will have a relatively high value of roh, round about 0.10[14]. Although in theory roh can take values up to 1, in practice values above 0.3 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".

Demographic items such as "currently married" and measures of mortality or morbidity such as "ill in past two weeks" will be hardly more likely to produce the same answer from two respondents in the same cluster than from two respondents in different clusters. These questions will have roh very close to zero, about roh = 0.02. For questions of health care practice and of use of health care services such as "use of ORS for last episode of diarrhoea", responses will depend on the level of services locally and on local custom, and the value of roh will probably be around 0.10 - 0.20. The value of roh can also be less than zero, particularly in stratified surveys, but usually a value less than zero may be considered as being due to sampling variation and treated as zero.

These guidelines for values of roh are very rough as little data are available and there will be variability in the value of roh from country to country, from survey to survey and from item to item. The basis for the values of roh given above is experience with a number of health surveys in developing countries[14]. 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. Much more experience is needed before any confidence can be placed in precise values of roh. 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. Further examples of values of roh and D are contained in references 5,8,14 and 15.

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[15] 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 seven would imply a design effect of 1.6, whereas a sample of thirty 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 two often used for the design effect[16] 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 thirty households per cluster, and it is expected that the resources will allow for about twenty clusters to be sampled. Since there will be one response per household, b will be equal 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 maximises 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 villages than within each village, 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 five percent. This indicates that with such a sample size we can be 95% certain that the true proportion of households with latrines will lie within plus or minus ten percent (two 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 seven responses from each cluster, and we would take this as the value of b. The value of n would be 7 x 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 plus or minus about 11%. Ignoring D would have led us to underestimate the width of the confidence interval as 8%.

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

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

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 plus or minus 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 ±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 ±6%, a little less precise than we 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 worse 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 mean 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 6.2) 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 will need to be double that calculated by the usual methods.

5. Analysis of data

We describe here 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 a spreadsheet is given in the Appendix.

5.1 Estimation of a proportion

Suppose that a number of households have been selected in each of c villages 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} village ($i=1,...,c$) there 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 i^{th} village 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

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

ie 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.

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

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

A spreadsheet for calculation of s is given in the Appendix, 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 i^{th} village) will have been unknown before the survey began and would probably be different if a different sample of households were taken from the same village. Failure to take account of these factors would lead to underestimation of s, and consequent overconfidence in the precision of the results (see Appendix for an example). In many cases x_i will not vary much between villages, for example when x_i is the number of households selected, and then the simpler formula

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

may be used instead of (6). If p is not too different from the mean of the p_i's, then this may be approximated very closely by

$$s = [\text{standard deviation of the } p_i\text{'s}]/\sqrt{c}. \qquad (8)$$

5.2 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 three children). Estimation of the mean and its standard error are carried out in exactly the same way as for a proportion (Section 5.1) 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 i^{th} village.

5.3 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 3.1, 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 under-represented in the calculation of the proportion p and its standard error. The solution is to weight the responses from this village 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 i^{th} cluster. An unweighted cluster has $w_i = 1$.

The approximate formulae (7) and (8) are unchanged as long as w_i is the same for all units in the cluster.

Weighting may 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.

5.4 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 Appendix.

5.5 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 (socio-economic, health status etc.). Dividing each standard error by the corresponding binomial value (2) gives a new estimate of the design factor (the square root $\sqrt{D}$ of the design effect). 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 villages is selected from the whole region under consideration and a sample of households is visited in each selected village. 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.

6.1 Multistage sampling

In a large region or country where an overall estimate is required it will usually be sensible to select the sample of villages 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 (ie by making a list with cumulative population sizes). Within each selected district, villages would be selected, again by the systematic pps method. The same number of villages 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 village.

It is possible with the systematic pps method described here that the same district is selected twice. This will happen if the population of the district is larger than the sampling interval. In this case two independent samples of villages 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 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 i^{th} district, summed over all villages selected in that district.

The method of sampling described here may be extended to more than two stages if required. It is possible to take a sample of regions, a sample of districts within each selected region and a sample of villages within each selected district. Provided that regions, districts and villages are all chosen with the systematic pps method, that the same number of districts is chosen in each region and the same number of villages in each district, then sample size and analysis are as described in the previous paragraph but with the word 'district' replaced by 'region'.

6.2 Stratification

It may be required 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 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

$$p = W_1p_1 + W_2p_2 + W_3p_3$$

with standard error

$$s = \sqrt{[W_1^2 s_1^2 + W_2^2 s_2^2 + W_3^2 s_3^2]}$$

where W_1 is the proportion of the country's population which belongs to stratum 1, and so on ($W_1 + W_2 + W_3 = 1$). The standard error, s, for the national estimate will be somewhat less than the standard errors for the individual strata.

6.3 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 villages (or districts) by ensuring that the list of villages from which the systematic sample is to be taken is ordered by some measure which is correlated with the the main purpose of the survey. For example, in a survey of the utilisation of mother and child health facilities, there may have been a previous study carried out some years ago into the same subject, or there may be other knowledge available which indicates which villages may be expected to have high levels of utilisation and which villages low levels. If not, one may be able to guess that those villages which are, say, further from the regional capital, or which cover a more widely scattered population, will have lower levels of utilisation than others. Whatever the measure chosen, if the villages can be listed roughly in order from a high to a low level of expected utilisation then the sample selected will contain villages with a spread of utilisation 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[17]. The improvement in precision cannot be quantified adequately to allow its use in sample size calculations.

Note that this procedure is valid only if the villages are ordered before the sample is chosen, and that this same order must be used to analyse all the measures observed in the survey.

7. Discussion

A simplified approach to survey design has been presented, with no attempt to cover all possible types of estimation. There are many books and publications which will give the reader all the necessary formulae for analysis in all sorts of situations[9] . What we have aimed to provide here is 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 will be better than the common practice of assuming that the data came from a simple random sample and using the standard error given by a calculator or standard computer package.

Acknowledgements This paper grew out of 'Sampling for Primary Health Care Reviews' written for and funded by WHO/SHS. We are grateful to Drs. Steve Lwanga, Steve Sapirie, Jim Tulloch and Ed Dowd of the World Health Organization for their contributions, and to colleagues at the London School of Hygiene and Tropical Medicine for discussions.

References

1. World Health Organization. Training for mid-level managers - Evaluate vaccination coverage. Geneva, WHO Expanded Programme on Immunization, n/d.

2. Henderson, R. H. and Sundaresan, T. Cluster sampling to assess immunization coverage: a review of experience with a simplified sampling method. Bulletin of the World Health Organization, 60: 253 - 260 (1982).

3. Lemeshow, S. and 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: 65 - 75 (1985).

4. Lemeshow, S. et al. A computer simulation of the EPI survey strategy. International Journal of Epidemiology, 14: 473 - 481 (1985).

5. 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: 93 - 99 (1985).

6. World Health Organization. Diarrhoea morbidity, mortality, and treatment practices: household survey manual. Geneva, WHO Diarrhoeal Diseases Control Programme, 1986.

7. Verma, V. The estimation and presentation of sampling errors. Technical Bulletin No. 11. London, World Fertility Survey, 1982.

8. Verma, V., Scott, C. and O'Muircheartaigh, C. Sampling designs and sampling errors for the World Fertility Survey (with discussion). Journal of the Royal Statistical Society, Series A, 143: 431 - 473 (1980).

9. Cochran, W. G. Sampling techniques. London, John Wiley and Sons, 1977.

10. Casley, D. J. and Lury D. A. Data collection in developing countries. Oxford, Oxford University Press, 1981.

11. Abramson, J. H. Survey methods in community medicine. Edinburgh, Churchill Livingstone, 1984.

12. Ross, D. A. and Vaughan, J. P. Health interview surveys in developing countries: a methodological review. Studies in family planning, 17: 78 - 94 (1986).

13. Kish, L. Survey sampling. London, John Wiley and Sons, 1965.

14. Liyanage, W. Design effects in health and immunisation surveys. University of Reading: unpublished M.Sc. thesis (1986).

15. Kish, L., Groves, R. M. and Krotki, K. P. Sampling errors for fertility surveys. Occasional paper No. 17. London, World Fertility Survey, 1976.

16. World Health Organisation. Sample size determination (a user's manual). Geneva, WHO Epidemiological and statistical methodology unit, 1986.

17. Kish, L. and Hess, I. On variances of ratios and their differences in multi-stage samples. Journal of the American Statistical Association, 54: 416 - 446 (1959).

Appendix: 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 villages are selected using the systematic pps procedure. Twenty households are chosen in each village in order to estimate, for the population, the proportion of recently pregnant mothers who have received postnatal care.

The data are:

Village	No. of recently pregnant women	No. 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_iy_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	D=130	E=73	

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

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 x p	0.2900
F	2 x p x E	78.621
G	p^2 x D	37.7
H	C − F + G	7.079
L	H / [c x (c−1)]	0.2360
M	square root of L	0.4858
s	c x M / B	0.1121

The 95% confidence interval for the true proportion is 0.5385 ±2 x 0.1121, i.e. 0.3143 to 0.7627.

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 mean of the p_i's is 0.5333 and their standard deviation is 0.3629, and so (8) gives s = 0.1481, very close indeed to the value given by (7).

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

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

thus ignoring the design of the study would have led us to assign our estimate a confidence interval from 0.3429 to 0.7341, which is 13% narrower than the correct value.

The design effect is estimated as

$$D = s^2/s_{srs}^2 = (0.1121)^2/(0.0978)^2 = 1.315.$$

Since $b = \Sigma x_i/6 = 4.333$, roh may be estimated in this case by $(D - 1)/(b - 1) = 0.073$.