PN-AAZ-429

DEMOGRAPHIC AND HEALTH SURVEYS

SAMPLING MANUAL

Institute for Resource Development
Westinghouse Electric Corporation
P.O. Box 866
Columbia, Maryland 21044, USA

November 1987

Demographic and Health Surveys (DHS) is a five-year Program (1984-1989) to assist governments and private agencies in developing countries with implementing 35 population surveys. The objectives of the Program are: (1) to provide leaders in survey countries with population and health data useful for informed decision making, (2) to develop in participating countries the skills and resources necessary to conduct high-quality demographic and health surveys, (3) to improve survey methods used to study populations in developing countries, and (4) to expand and improve the worldwide body of information on population and health.

DHS is a Program of the Institute for Resource Development (IRD), a subsidiary of Westinghouse Electric Corporation, with the Population Council as a major subcontractor. DHS is funded by the United States Agency for International Development (Contract No. DPE-3023-C-00-4083-00).

CONTENTS

1. DHS SAMPLING POLICY

1.1 General policy on sampling. 1

1.2 The survey domain 2

1.3 Sample size 3

1.4 Basic sample design 4

1.5 Use of a pre-existing sample. 5

1.6 Stratification. 6

1.7 Size of the sample "take" per cluster 7

1.8 Listing and segmentation. 8

1.9 The sampling operations 9

1.10 Sampling errors 10

1.11 Documenting the sample. 11

2. SELECTED SAMPLING TECHNIQUES AND ISSUES

2.1 Disproportionate sampling between domains 12

2.2 Sample weighting. 17

2.3 Systematic sampling 25

2.4 The PPS self-weighting design 27

2.5 A practical model sample, with variants 34

2.6 Subsampling from an existing sample or master sample. 45

2.7 Errors of coverage and nonresponse. 49

2.8 Sampling errors 57

2.9 Sample description and documentation. 59

1. DHS SAMPLING POLICY[1/]

1.1 General Policy on Sampling

Sample design for surveys in the DHS program is guided by a number of general principles, although some modification may be required in the country-specific situation.

First, scientific probability sampling must be used.[2/] Non-probability methods represent a false economy. Although they may yield good estimates, they cannot provide the confidence that is necessary in the event of unexpected findings. If this occurs, the use of non-probability methods may lead to controversy and ultimately to criticism of the survey design.

Second, samples should be self-weighting unless there is good reason to depart from the principle in specific cases. In countries where statistical offices are new or lack resources and/or personnel, the use of weights may present problems. The need to compute weights and carry them as part of the database, the need to assess when and how they should be applied, and to correctly report their use, can be an appreciable burden on staff. Self-weighting designs are likely to work best when there are a number of survey objectives. On the other hand, there are circumstances in which these considerations are counterbalanced by

---

1/ This manual was prepared by Christopher Scott with the assistance of Graham Kalton and Alfredo Aliaga.

2/ A probability sample is defined as one in which the units are selected with known and non-zero probabilities. The term excludes purposive sampling, quota sampling, and such methods as the uncontrolled post-sampling delineation of fixed sized clusters in the field each centered on a sample point.

advantages in adopting unequal sample weights, particularly between reporting domains (see Section 2.1).

Third, if an adequate pre-existing sample or sampling frame is available, it should be used. Similarly, DHS favors appropriate integration with ongoing national survey programs, in the interest of economy and coordination.

Fourth, the sample design should be as simple and straightforward as possible, to facilitate accurate implementation.

In the sections which follow, DHS policy is set out in relation to a number of specific aspects of sample design. In the first part of the manual, technical issues are treated briefly. The second part deals in greater detail with these specific issues.

1.2 The Survey Domain

Geographical coverage of each survey should be the entire national territory unless there are strong reasons for excluding certain areas. If areas must be excluded, they should constitute a coherent domain. A survey from which a number of scattered zones have been excluded is difficult to interpret and to use. The demographic domain for DHS samples is defined as all women aged 15-49.

In many surveys, an important objective is to compare urban and rural populations. Where this is the case, it is necessary to insure that the urban sector is adequately represented in the sample. In a country with an urban population of less than 20 percent, this may require oversampling of the urban sector. In this case, the arguments in favor of a self-weighting sample are overridden. Moreover, once varying weights have been introduced in one domain, there may be reason to introduce them in other domains. For example, several different sampling rates may be needed to permit accurate comparison of regions within a country.

1.3 Sample Size

The issue of sample size is only partly a technical one. The larger the sample the more elaborate the analyses that can be sustained. The choice of sample size involves balancing the demands of analysis with national capability and the constraints of funding.

The DHS program is designed for samples of 5,000 to 6,000 women aged 15-49.[3/] Experience with earlier survey programs such as the World Fertility Survey (WFS) and the Contraceptive Prevalence Surveys (CPS), shows that such a sample size can sustain a variety of analyses. In addition, it provides acceptable levels of sampling error for such key parameters as fertility, infant and child mortality, contraceptive prevalence, and breastfeeding. The survey allows for some adjustment of the size of sampling units for background variables of interest.

Various factors may affect standard DHS sample size. First, most of the analyses planned relate to, or are determined by, ever-married women in the sample. Further, an important subset of questions relates to children born in the last five years. The effective sample size for such analyses depends not only on the original sample size but also on the proportion of currently married women and the current level of fertility. Thus, in countries where entry into union occurs at a relatively early age, or where the fertility rate is relatively high, a smaller initial sample of women aged 15-49 may be sufficient.

Secondly, the standard DHS sample size allows up to five geographical regions to be distinguished in the tabulation of key variables. Typically, a sample of 1,000 women aged 15-49 is needed for each geographical domain

---

3/ Sample size in this section refer to the target sample. Factors such as nonresponse and undercoverage at the mapping/listing stage may reduce the target sample as much as 10 percent.

for which separate findings are required. However, in some countries there may be special reasons for utilizing a larger number of regions; in such cases a somewhat larger sample may be allowed.

In any discussion of the sample size appropriate for a particular survey, it should be noted that a larger sample is more difficult to manage and supervise, especially if the field work period cannot be extended. This fact argues for caution in allowing inflated sample sizes, particularly in countries where survey capability is limited.

Finally, it should be stressed that in determining sample size the availability of funds is a limiting factor. In the framework of the DHS program, it would be unusual for a country to obtain external funding to carry out a survey utilizing a sample larger than 9,000.

1.4 Basic Sample Design

Most developing countries possess convenient area sampling frames in the form of the enumeration districts (EDs) of the most recent population census. These generally come with sketch maps and size estimates, and the EDs do not vary greatly in population size. However, in most countries, there are no satisfactory lists of dwellings, households, or individuals, and no address system outside the more affluent parts of the cities. In general, survey personnel have to make their own lists, although sometimes they can share with other surveys or select a subsample from a master sample (see Section 1.5).

Census EDs are generally too large (typically 1 to 2 thousand population) for it to be economically feasible for a single survey to undertake the listing of all households in the survey's sample of EDs. The EDs, therefore, need segmenting into smaller areas for a further stage of area sampling before household listing begins. In most cases, the census maps are not accurate enough for the work of segmentation to be done in the office. A field operation is needed to select EDs for mapping and segmentation.

A convenient and practical sample design has been developed based on experience with past surveys. First, a standard segment size is adopted, such as the 500 census population. Every ED in the country is then assigned a measure of size equal to the number of standard segments it contains, by dividing its census population by 500 and rounding to the nearest whole number. A sample of EDs is then selected with probability proportional to this measure of size. In the selected EDs, a mapping operation is carried out to create the designated number of segments and one of these is selected at random. This procedure provides a single-stage equal-probability sample of segments which are roughly constant in size. In the selected segments, all dwellings or households are listed, and a fixed fraction of them are selected by systematic sampling. In the selected households, a roster is completed to identify women aged 15-49 and all of these are interviewed.

There are a number of variations on this design (see Section 2.5). Such self-weighting segment samples will usually be employed for DHS surveys, except where an existing sample is utilized.

1.5 Use of a Pre-existing Sample

A pre-existing sample can be accepted only where DHS is confident of the sample design, including its detailed parameters. Often, such samples have unequal weights in different domains; they may also use an average sample "take" at the final stage which differs from that desired for DHS. The task of the sampler for the DHS survey is then to design a subsampling procedure which produces a sample in line with DHS requirements. This will not always be possible. However, the larger the parent sample in relation to the desired DHS subsample, the more flexibility there will be for developing a subsample design (see Section 2.6).

A key question with a pre existing sample is whether the listing of dwelling/households is still usable by DHS or whether it needs to be updated. If updating is required, use of a pre-existing sample may not be worthwhile.

The potential advantages are: (1) economy, and (2) increased analytic power through cross-analysis of two or more surveys.

The disadvantages are: (1) the problem of adapting the sample to DHS requirements, and (2) the problem of repeated interviews with the same household or person (respondent fatigue, contamination).

DHS encourages use of an existing sample, provided that it meets technical standards (see Section 2.6).

1.6 Stratification

Geographical stratification can generally be obtained by systematic sampling of EDs from an ordered census list. In some cases, this can be supplemented by explicit stratification based on socioeconomic zones. If for example, the family planning program is noticeably more active in certain regions, this factor should be represented in the stratification adopted.

The principal objective of stratification is to reduce sampling error. In a stratified sample, the sampling error depends on the population variance existing <u>within</u> the strata but not <u>between</u> the strata. For this reason, it pays to create strata with low internal variablity. Systematic sampling from an ordered list has an effect similar to stratification and is often called "implicit stratification" (see Section 2.3).

Stratification should be introduced only at the first stage of sampling. At the dwelling/household selection stage, systematic sampling is used for convenience; however, no attempt should be made to reorder

the dwelling/household list before selection in the hope of increasing the implicit stratification effect. Such efforts generally have a negligible effect.

1.7 Size of the Sample "Take" per Cluster

The optimum number of women to be selected per cluster depends on the variable under consideration. In estimating contraceptive prevalence and its determinants, the variables of primary interest tend to be highly clustered, and many comparisons are required between geographical areas. For such objectives, calculations based on WFS surveys suggest an optimum average take of 15-20 women per cluster. Other fertility variables are less clustered. Most of the comparisons of interest are non-geographical, e.g., comparisons between age groups or levels of education. These are so-called "cross-classes," whose different categories appear in most clusters. For such variables and analyses, the optimum is substantially higher, normally well over 50.[4/] Information on variances is not available for the data produced by DHS health questions.

A larger cluster take implies a less expensive survey design. However, DHS surveys have a wide range of objectives, including some for which the optimum take is around 20. In view of this multipurpose role, it is suggested that large takes be avoided.

DHS proposes a cluster take of about 40 women for the rural sector. In the urban sector, the cost advantage of large takes is generally smaller, and DHS recommends a take of about 20 women. Where a pre-existing recent household list is available, these figures could be further reduced since the main factor favoring a large take is saving in listing operations.

4/ Data from World Fertility Survey Assessment project.

1.8 Listing and Segmentation

Listing of dwellings or households prior to selection of a sample represents an appreciable field cost, but there is no reliable method by which it can be avoided. Indeed, analysis of sample coverage rates in the WFS suggests that more, rather than less, attention to the quality of listing operations is required, if serious biases are to be avoided. In particular, the combination of listing, sampling, and interviewing into a single operation, conducted by the interviewer while moving over the sample area, is an unworkable operation. Even less acceptable is the attempt to avoid listing altogether by having interviewers create clusters as they go along, or select a sample at fixed intervals during a random walk up to a predetermined quota. Such methods are designed to eliminate conscious choice in selection, but they fail to meet the requirement that the sample be selected in such a way as to give a known and non-zero probability to every potential respondent. Essentially, these methods represent a false economy. It is more efficient to reduce the sample size and retain the listing operation.

Listing costs can be reduced by using segmentation to decrease the size of the area which has to be listed; however, segmentation generates its own costs and skill in map making and map interpretation is required. Segmentation becomes progressively more difficult as segments become smaller, because there are not enough natural boundaries to delineate very small segments. Moreover, concentration of the sample into smaller segments increases the sampling error. Since neighbors' characteristics are correlated, a smaller segment captures less of the variety existing in the population; this leads to less efficient sampling. There is a point beyond which it is not useful to attempt further segmentation. As a general rule the average segment size should not be less than 500 population (approximately 100 households) in rural areas and in unplanned urban areas. In planned urban areas, listing is much easier, and an address system may be in effect. In this case there

is no real need for segmentation. However, unless such areas are common and identifiable in advance, it may not be worthwhile to make special arrangements for them.

It is sometimes suggested that listing could be avoided by making segments so small that they are equal to the required sample take per cluster. One could then use a "take-all" rule at the last stage of sampling. Such small segments, however, will generally be difficult to delineate. In planned urban areas, this difficulty may be reduced--one could adopt blocks, or even single buildings, as segments--but urban units of this kind are likely to be homogeneous, containing similar households, and therefore of little use as sampling clusters. A more important objection to the omission of independent listing is that it serves as a check on the completeness of the interviewers' work. Have they covered the whole segment, or have households been omitted, deliberately or by oversight? WFS coverage rates suggest that this kind of error is widespread and careful controls are needed.

1.9 The Sampling Operations

After selection of the area units, usually EDs, the next step is segmentation. In most cases segmentation can only be carried out in the field. Each ED, whether due for segmentation or not, should be visited for verification of maps. When this has been done, the same team can proceed to create the designated number of segments and to delineate them clearly on the map of the ED. If size measures are required based on a quick count, these can be provided at the same time (see Section 2.5).

Selection of the sample segment in each ED is the next step. It is important to apply safeguards here to prevent biased selection. Preferably, the selection should be made in the office, or failing this, by a senior staff member in the field.

The next step is listing. If this is carried out more than one month (urban) or three months (rural) in advance of the interviewing, it should be on a dwelling basis, including empty dwellings. The subsequent interview should cover the current occupants of the listed dwelling whether or not they occupied it at the time of listing. If listing is to be less than a month before interviewing, it would be acceptable to list households as such, with the subsequent interview covering the named household listed. Normally, listing should not be done by the interviewers, and for this reason a gap of at least a month is to be expected for logistical reasons.

Following the listing comes the selection of dwellings/households. This will be carried out at an interval predetermined in the office.

Finally, the interviewing team will visit the area and an interviewer will be assigned to each dwelling/household selected. The interviewer will begin with a brief household interview, listing household members and visitors and identifying among them all persons who spent "last night" in the household. The interviewer will then proceed to interview, among those present last night, all women aged 15-49.

In the event of failure to contact a household or person identified as eligible, the interviewer is required to make two callbacks on different days before the interview is abandoned.

1.10 Sampling Errors

Sampling errors for the main variables and subclasses will be computed in the country or at DHS headquarters (see Section 2.8).

1.11 Documenting the Sample

The task of the sampler does not end with the selection of the sample. The preservation of sampling documentation is an essential requisite for sampling error computation, for linkage with other data sources, and for various kinds of checks and supplementary studies. Experience shows that special efforts are needed during the survey operation if this seemingly unimportant chore is to be carried out effectively: (1) at the time of sample design, (2) at the end of the field work, and (3) at the completion of the data file. If preservation of documentation is delayed, considerable effort will be required to reconstitute the missing information when it is needed.

DHS gives special attention to the issue of documentation. Details of requirements are set out in Section 2.9. The same section includes the requirements for sample information to be entered on the individual data records.

2. SELECTED SAMPLING TECHNIQUES AND ISSUES

2.1 Disproportionate Sampling Between Domains

In a standard DHS survey, each final unit (i.e., each eligible woman) will have an equal probability of selection. This sample design is known as the equal probability of selection method (EPSEM). It may also be called self-weighting because the results can be treated as directly representative of the population concerned, without the need for weights in the analysis.

This section deals with departures from this simple model and, in particular, with the deliberate introduction of different sampling fractions (or probabilities) in different domains of the sample.

There are two main motives for such disproportionate sampling between domains:

(1) Cost efficiency is increased if a higher sampling fraction is used in domains where the population variance is greater and the unit costs are lower. Thus, sampling fractions may be manipulated in order to reach an optimum design.

(2) The survey planner may wish to report findings for a population subgroup which constitutes only a small percentage of the whole population. If a fixed sampling fraction is used everywhere, this small group will be allocated a correspondingly small sample. Sampling error increases as sample size falls, and it may be that, given the overall sample size, the sampling error for this small domain would be unacceptably large. The problem can be resolved by oversampling the small domain, thus reducing its sampling error. When considering the whole population, a weight is introduced to compensate for the unbalanced sampling in the special domain.

The first of these strategies, varying the sampling fractions between domains to maximize cost efficiency, relies on the formula

$$f_h = kS_h / \sqrt{C_h}$$

where f is the sampling fraction, k is a constant, S is the element standard deviation existing in the population (not the standard error of the estimate), C is the cost per unit, and the suffix h designates the domain (or stratum). This allocation of the sample among the strata minimizes the standard error of the overall sample mean for a given budget. This strategy is of limited relevance to DHS, both because the optimum allocation varies for different variables and because the variations between domains in a demographic survey are unlikely to be very great, so that the potential economy is small. In view of the disadvantages of a weighted sample, the "optimum allocation" sample is not recommended for DHS (see Section 2.2).

The second technique, oversampling a small domain of study to give it a more substantial sample and hence reduce the sampling error of its estimate, may sometimes be of value in a DHS survey. In particular, there may sometimes be a call for oversampling the urban sector. An obvious extension is to seek to give an adequate sample to each of several different domains. For example, some national survey organizations favor designs which will yield the same precision in each of a number of regions making up the country. If one assumes that the population standard deviation is the same in all the regions, this strategy would imply an equal sample size in each region, since sampling error is a function of sample size. As regions invariably differ in their population size, this would imply unequal sampling fractions between regions.

In order to evaluate these schemes, some simple formulas are needed. In what follows, we assume that the population standard deviation is constant among the domains. We also assume that any over or under-sampling is implemented by modifying the first stage selection only, leaving the design for subsequent stages unchanged. We write n_h for the sample size and w_h for the weight in domain h.

If the sample in domain h is changed from n_h to n'_h, the sampling variance is multiplied by n_h/n'_h. For example, if the sample is doubled, the error variance is halved. What really matters from the point of view of the analyst is not the sampling variance but its square root, the standard error. This will be multiplied by $\sqrt{n_h/n'_h}$ Thus, to achieve half the standard error, we need to quadruple the sample size--large changes in sample size are needed to achieve modest changes in sampling error.

If the total sample size is kept constant when disproportionate sampling is introduced, the unequal sampling fractions cause an increase in sampling error for the whole sample. It has been shown (Kish, Survey Sampling, 1965) that the sampling variance for estimates relating to the whole sample is multiplied by the factor

$$L = (\Sigma n_h w_h^2)\ (\Sigma n_h)\ /\ (\Sigma n_h w_h)^2$$

This is never less than 1 so that, though we gain precision in a given domain estimate, we pay for it with increased sampling error at the level of whole-sample estimates.

Example 2.1.1

The country of Delicia has an urban population of 1 million and a rural population of 4 million (see Table 2.1.1). We wish to select a sample of 5,000 people (0.1 percent). In order to insure more precision in the estimates for the relatively small urban sector we decide to double the sampling fraction there compared with the rural, while maintaining the same total sample. What do we gain, what do we lose and how do we proceed?

Table 2.1.1

Domain	(1) Pop. = N_h	(2) Proportionate Sample = n_h	(3) Disproportionate Sample = n'_h	(4) var'/var	(5) SE'/SE
Urban	1,000,000	1,000	2,000 k = 1,667	0.60	0.77
Rural	4,000,000	4,000	4,000 k = 3,333	1.20	1.10
TOTAL	5,000,000	5,000	6,000 k = 5,000	1.08=L	1.04

Column (1) gives the population estimates.

Column (2) allocates the given total sample of 5,000 between urban and rural in proportion to the population in column (1).

In column (3) we begin by doubling the sample size for the urban domain (yielding 2,000) and leaving the rural domain unchanged (4,000), but since this implies a total sample of 6,000 we introduce the factor k. Summing the lefthand column we obtain 6,000k. Equating this to the desired sample of 5,000 we obtain k = 5/6. Multiplying out, we obtain the righthand column, which gives the new sample sizes n' for the disproportionate allocation.

Since error variance is inversely proportional to sample size, column (4) is simply n_h/n'_h for each domain. The "total" in column (4) is the quantity L discussed above; putting n'_h for n_h and (n_h/n'_h) for w_h, the formula gives L = 27/25, or 1.08.

Column (5) is the square root of column (4).

Column (4) is the main concern if one is interested in sample size or in costs, since these are approximately proportional to the sampling variance of the estimate. Column (5) is more useful if the concern is with the size of the sampling error.

It will be seen that the disproportionate sampling plan chosen reduces the sampling standard error in the urban domain to 77 percent of its value with proportionate sampling. This gain is paid for with an increase of 10 percent for the rural domain and 4 percent for the whole sample.

With the overall sampling error increased, one may ask what would be the equivalent proportionate sample to yield this larger error for the whole sample. The answer is n/L, or 5,000/1.08 = 4,630. It is as though 370 cases had been removed from the total sample. One might also ask what size of sample would be necessary to maintain the same sampling precision in the whole-country estimate as we had with the proportionate sample assuming the disproportionate sampling rates shown. The answer is nL, or 5,000 x 1.08 = 5,400. So, an 8 percent increase in sample size would be needed to compensate for the disproportionate sampling.

Example 2.1.2

In the country of Bureacracia there are three important regions: North, Central and South (see Table 2.1.2). These differ considerably in population size. What would be the implications if it were decided to select a disproportionate sample designed to yield identically sized samples in each region? We assume the population sizes in column (1) of the table, with a total sample of 5,400.

Table 2.1.2

Region	(1) Pop. = N_h	(2) Proportionate Sample = n_h	(3) Disproportionate Sample: same in all regions = n'_h	(4) $(var'/var)^2$	(5) SE'/SE
North	1,000,000	1,000	1,800	0.56	0.75
Central	2,000,000	2,000	1,800	1.11	1.05
South	2,400,000	2,400	1,800	1.33	1.15
TOTAL	5,400,000	5,400	5,400	1.107=L	1.052

Columns (4) and (5) give the relevant conclusions.

The disproportionate sample reduces the sampling error by 25 percent in the North, increases it by 5 percent in the Central and increases it by 15 percent in the South. For the country as a whole the increase is 5.2 percent.

The equivalent proportionate sample, giving the same sampling error for the whole-country estimate, is $n/L = 5{,}400/1.107 = 4{,}878$, so that the number of cases "lost" is $5{,}400 - 4{,}878 = 522$. The sample size with the same disproportionate sampling rates which would be needed to maintain the same precision for the whole-country estimate as would be yielded by the 5,400 proportionate sample is $nL = 5{,}400 \times 1.107 = 5{,}978$, an increase of 578 cases.

Conclusions

Both of these examples illustrate an important fact--neither the gains nor the losses resulting from disproportionate sampling are very substantial unless the sampling fractions depart a long way from equality. Even the 2-to-1 oversampling in Example 2.1.1 only reduces the urban sampling error by 23 percent; this is paid for by some modest increases in rural and total sampling errors. In Example 2.1.2 the regions vary widely in population size yet the gains and losses from using equal samples in all regions are still minor.

The small size of the gains needs to be set against the disadvantages of a weighted sample which are discussed in the next section. In many cases departure from proportional sampling will not justify the practical inconvenience.

2.2 Sample Weighting

If a sample of size n is selected from a population of size N, using an equal probability design with selection probability $f = n/N$, we can estimate any total in the population by multiplying the corresponding sample total by N/n. The multiplier N/n is often called the "raising factor."

Instead of estimating a total we may wish to estimate a mean, a rate, a proportion, a percentage or a ratio. In all of these cases the raising factor will be applied to both the numerator and the denominator and will cancel out. It follows that means, rates, etc. can be taken straight from the sample; the sample value provides a direct estimate of the population value. For example, if 80 percent of the sample women are married then we estimate that 80 percent of the corresponding women in the whole population are married. Nearly all of the figures presented in a DHS survey are means, rates, or percentages; these require no raising, and no weighting as long as an equal probability sample design is used.

However, if the selection probability varies this simple situation changes. The raising factor $N/n = 1/f$ now has to be applied separately to each domain for which the sampling fraction is different and it does not cancel out in numerator and denominator. Table 2.2.1 provides an example (the raising factors in this case are generally called weights). The sampling fractions adopted are shown in column (1) and the data assumed by the example are in columns (3) and (4). The objective is to estimate the percentage of married women.

The totals for columns (3), (4) and (5) have little meaning and would not normally be calculated, because they are unweighted. They have been entered here for the sake of comparison with the final estimates in the subsequent columns.

Table 2.2.1

Domain	(1) Sampling Fraction	(2) Weight	(3) Sample Data (unweighted)	(4)	(5)	(6) Population Estimates	(7)	(8)
h	f_h	$w_h = 1/f_h$	Total Women	Married Women	Percent Married	Total Women	Married Women	Percent Married
East	1/1,000	1,000	900	810	90	900,000	810,000	90
West	1/2,000	2,000	1,000	800	80	2,000,000	1,600,000	80
Coast	1/2,500	2,500	1,200	900	75	3,000,000	2,250,000	75
TOTAL			3,100	2,510	81	5,900,000	4,660,000	79

The raising factors, or weights, shown in column (2) are used to raise the sample data of columns (3) and (4) to population estimates in columns (6) and (7). Comparing columns (5) and (8) we see that the weights do not affect the estimates within domains and this is as predicted (because the weight is constant within any domain). However, the total in column (8), which is obtained by dividing total column (7) by total column (6), gives 79 percent instead of the 81 percent obtained from the unweighted sample data. The differential weighting has lowered the estimate.

Since the weights appear in the numerator and the denominator of the final estimate we can obviously remove any common factor we wish from the weights. For example, we can divide all the weights by 1,000. Column (2) then becomes 1; 2; 2.5. These are no longer raising factors but they are still weights, and the final estimate of 79 percent married is not changed. Thus it is only the relative values of the weights that matter as long as we are estimating means, rates, proportions, percentages, or ratios.

In symbols, a rate or ratio Y/X is estimated, with weighting, from the formula:

$$(\Sigma w_h y_h)/(\Sigma w_h x_h) \qquad \underline{1}/$$

where x_h and y_h are the sample totals in domain h. If the variable X is the number of cases N, then we obtain the formula for estimating a weighted mean:

$$\bar{y} = (\Sigma w_h y_h)/(\Sigma w_h n_h)$$

1/ Capital letters indicate the population, small letters the sample.

The standard way to introduce weights into the analyses is to include a weight variable on each individual record. In the example above, for instance, the data records for each sampled woman in the Eastern domain would have a value of 1 for the weight variable, those in the Western domain a value of 2 and those in the Coastal domain a weight of 2.5. When a weight variable is included in this way, it is straightforward to employ it to run any analyses required provided that the computer program is capable of handling weights.

Up to this point we have assumed that the weights are used to compensate for unequal selection probabilities. Weights can also be used to compensate for nonresponse, the failure to obtain data for some of the sampled units. In the example described in Table 2.2.1, suppose that column (3) represents only the respondents while the numbers initially selected for interview were larger, as shown below (3'):

	Respondents (3)	Selected (3')	Response Rate (%)
East	900	1000	90
West	1000	1250	80
Coast	1200	1412	85

The response rate--(3) as a percentage of (3')--is indicated for each region and can be incorporated into the weighting system. To compensate for nonresponse, the raising factors could be increased from 1,000 to 1,000/0.9 = 1,111 for the West, from 2,000 to 2,000/0.8 = 2,500 for the East, and from 2,500/0.85 = 2,941 for the Coast. Making the weight for the West 1, the weights for the East and the Coast are then 2.25 and 2.65 respectively. These weights differ, therefore, from the 1:2:2.5 arising from the sampling weights alone because they also compensate for the variability in response rates across the three regions. For instance, the West and Coast have lower response rates than the East, and hence their weights have been increased relative to that

for the East in compensation. Nonresponse adjustments of this type try to compensate for biases introduced by varying response rates in various parts of the sample. They do so by increasing the weights of the respondents to represent the nonrespondents. It should be noted, however, that these adjustments do not attempt to compensate for the bias resulting from any systematic differences between respondents and nonrespondents <u>within</u> the domains.

Another form of weighting adjustment compensates for differences between the achieved distribution for the sample for some characteristic and known population distribution for that characteristic. For example, even with a perfectly implemented equal-probability sample, the age distribution in the sample will differ somewhat from the population age distribution because of sampling fluctuations. If the population age distribution is known (from the census, for example) one can reweight the sample, age-group by age-group, to bring it into line with the population distribution. The same type of weights N_h/n_h are used, where h designates the age-group. This kind of adjustment is known as "post-stratification."

When the population distribution of a characteristic is known, the post-stratification type of adjustment can also be used to compensate for nonresponse and noncoverage.

<u>When is weighting necessary?</u>

The overall effect of weighting is small. The sampling fractions in the example described in Table 2.2.1 vary by as much as 2.5 to 1 and are closely correlated with the variable of concern, percent married. Despite such circumstances favoring a weighted/unweighted differential, it turns out that weighting reduces the final estimate only slightly, from 81 to 79 percent. Weighting adjustments for nonresponse rates would generally have much less effect even than this, because the weights (which are the reciprocals of the response rates) are most unlikely to vary between domains by more than 10 percent.

In a general survey report the main purpose is to provide the best possible estimates of a wide variety of population characteristics based on the sample. If varying sampling fractions have been used in the sample design these should be reflected in the estimates; in other words, in these circumstances weighting should be considered obligatory even if the effects are small. Such weights are termed "design weights". Obviously in an equal probability sample no design weights are required, which is why such a sample is termed "self-weighting".

With nonresponse weights the situation is different. The weighting adjustment corrects only a part of the nonresponse bias and the corrections will nearly always be trivial in developing countries, where response rates are high. It is reasonable to omit such reweighting if the sample is otherwise self-weighting. On the other hand if design weights are to be used for domains there is little added complexity in modifying these weights to take account of variation in nonresponse rates between these domains. The combined weight for domain h will be $1/(f_h R_h)$, where R_h is the response rate.

Finally, post-stratification weighting is better avoided in most developing countries unless one has considerable confidence in the accuracy of the census data used- -or unless there are good reasons for believing that there was severe undercoverage in the survey sample.

In general, DHS recommends self-weighting samples. Where this recommendation is accepted no weights should be computed or used. Where the sample is not self-weighting, design weights, combined with nonresponse weights for the same domains, should be entered in the data record and used in all analyses.

Disadvantages of disproportionate sampling

The disadvantages of disproportionate sampling are essentially the disadvantages of weighting. The weights have to be computed and

standardized (see below) and entered in the data record, then verified. Their presence must be communicated to the user, with advice on their use. One must then remember to use them. For certain analyses and certain programs this requires special precautions (see below). Discussion on this issue in a particular application may waste the time of highly qualified staff. Time may also be wasted trying to determine whether a given table was in fact based on weighted or unweighted data. Finally, in publishing basic tables of results there is a problem with weighted data. Some users may wish to regroup data into new categories. This can only be done if the weighted sample sizes are given for each category. At the same time the unweighted sample sizes are needed as an indicator of the reliability of the estimates. It is usually not possible to provide all this information in basic tables, so in practice, users are not given all the information they need.

Most of these drawbacks are minor but they may create problems for a small statistical office.

Standardization of weights

Caution should be taken in the use of weights in survey analyses. In some computer programs the weight for a record is treated as if it is the frequency of occurrence of the record. This procedure gives the correct weighted estimates, but any sampling errors produced by the program are invalid. Care must be taken to distinguish between the sample size and the sum of the sample weights. As noted above, the sum of the sample weights is generally an arbitrary number, resulting from choosing the weights in any way that gives the correct relative magnitudes. If, for example, the weights are computed simply as the reciprocal of the selection probabilities ($w_h = 1/f_h$), the sum of the weights will exceed the sample size, often to a considerable extent. The incorrect treatment of the sum of the weights as equivalent to the sample size then attributes much greater precision to the survey estimates than

is warranted. Similarly, the application of significance tests (e.g., χ^2 tests) treating the sum of the weights as the sample sizes will grossly overstate the significance levels achieved. The calculation of the sampling errors of weighted estimates must be carried out using the appropriate formulas and programs (see Section 2.8).

A recommended procedure, which has several advantages, is to <u>standardize</u> the weights in such a way that the total weighted sample interviewed is equal to the total unweighted. This means multiplying all weights by the factor $(\Sigma n_i)/(\Sigma w_i n_i)$, where n_i is the number of cases bearing the weight w_i. It is these standardized weights that are entered into the data record for each individual.

Note that within any one category or subset of the sample the equality of weighted n and unweighted n will not hold exactly. It is therefore a waste of effort to compute and apply the standardizing factor to a high degree of accuracy just to insure that the equality holds at the whole sample level to the nearest unit. For DHS surveys it will be sufficient to provide standardized weights correct to the second decimal place.[2/]

Standardization of the weights in this way has the advantage that if weighted n's only are quoted the reader will not be seriously misled. In addition, the error, described above, which results from confusing the sum of the weights with the sample size, will be avoided. Note that the computation of sampling error still needs to take into account the sample structure--stratification and clustering (see Section 2.8).

2/ In entering the weights it may appear convenient to omit the decimal point, leaving it to the user to divide the weights by 100. However, some users may neglect to do so, which can lead to confusion. It is therefore recommended to enter the weights into the data record complete with the decimal point.

2.3 Systematic Sampling

Systematic sampling means the selection of units at a fixed interval from a list, starting from a randomly determined point.

Compared with random selection, systematic sampling has three advantages:

- It is easier to perform.
- It allows easy checking of the selection.
- If the list is in some order, the method provides a degree of stratification in respect to the variable on which the list is based.

Because of these advantages systematic selection is much more often used than random selection. In real life most lists do contain some degree of ordering.

Systematic sampling is normally carried out as follows, assuming a whole number interval I. We begin with a random number less than or equal to I. Let this be R. We then select the units numbered until we arrive at the end of the list:

$$R;\ R + I;\ R + 2I;\ \text{etc.}$$

If the design specifies the <u>number of units</u> to be selected, the interval I is computed as N/n, rounded to the nearest whole number, where N is the number of units in the list and n the number to be selected. On the other hand, if the design specifies the <u>sampling fraction</u> f (or the probability) we compute the interval as $I = 1/f$. In this case, if I is not a whole number there may be an appreciable error in rounding it to the nearest whole number. It is suggested that where I is a non-whole number less than 5, the decimal interval method be used. Moreover, if the same non-whole number interval is to be used repeatedly in the sample selection, then the decimal interval method should be used <u>in any case</u>.

Selection with a decimal interval may be carried out as follows:

(1) Express the interval I rounded to one decimal place (the necessary modifications, if two or more decimal places are carried, will be obvious).

(2) Find a random number between 1 and 101 and place a decimal point before its last digit. This becomes R.

(3) Compute the sequence of sampling numbers:

$$R;\ R + 1;\ R + 2I;\ \text{etc.}$$

(4) The whole number part of each sampling number indicates the unit selected.

Example 2.3.1

Suppose the interval I is 3.4.

Select a random number between 1 and 34. Suppose 23 is selected. Then R = 2.3. The sampling numbers and selections are as below:

Sampling Number	Unit Selected
2.3	2
5.7	5
9.1	9
12.5	12
etc.	etc.

It will be seen that the method gives an interval which is sometimes 3, sometimes 4, with the desired average of 3.4.

Note: If the random start R is less than 1 the first member of the sequence yields no selection. The method is still valid, however, with R + I giving the first selection.

It is advisable to check, after sampling, that the number of units selected is equal to N/I, with an error of not more than 1, where I is the interval actually used.

Often the sample design calls for numerous systematic samples, for example, a systematic sample of households may be needed within each selected area unit. In this situation a separate random start R should be determined independently for each sample.

2.4 The PPS Self-weighting Design

Different sampling probabilities may be used in different domains, or strata, of the sample (see Section 2.1). This principle may be pushed to the extreme by selecting every primary sampling unit (PSU) with a different probability.

A common sampling plan is to select the PSUs with a probability proportional to the estimated population of each unit. Thus, if unit A is estimated to be 10 times as large as unit B, it is given 10 times as many chances of being selected. This gives a sample biased in favor of large units but the bias is corrected later. This method is called "sampling with probability proportional to size," or "PPS sampling."

One way of correcting the bias is to use the opposite system at the household sampling stage, i.e., sampling with probability <u>inversely</u> proportional to the measure of size that was used at the area sampling stage. This means that the sampling fraction <u>for households</u> in area unit A will be 10 times smaller than in unit B, thus cancelling the bias introduced at the first stage. A given household in unit A now has exactly the same probability of selection as a household in unit B. Since this is an equal-probability sample no weighting is needed in the analysis, i.e., the sample is self-weighting. The advantages of self-weighting samples have been discussed (see Section 2.2).

With this sample design, if the estimate of size used at the first stage were always exactly equal, or proportional, to the number of households in the unit, it follows that at the second stage one would be

selecting a fixed number of households in every selected unit.[3/] In practice, the estimate of size is inaccurate, to varying degrees in different situations. This method is sometimes called "sampling with probability proportional to estimated size," or "PPES sampling." If the above sampling plan is followed using PPES sampling, one gets a household sample in each PSU which is only approximately constant.

It is the approximate constancy of the workload in each area unit that constitutes the main attraction of the PPES self-weighting sample, together with the self-weighting property itself. The field work is a good deal easier to organize if we do not have very large workloads in some areas and very small ones in others. This advantage is particularly significant where we are sampling ordinary administrative units, which commonly vary widely in population size.

A further advantage of PPS (or PPES) sampling is that, in general, it greatly reduces the sampling error for estimates of totals. It will also, in general, improve sampling efficiency for means, rates, percentages, and ratios.

Method of selection with PPS

Selection with PPS (or with PPES) is most easily performed as follows:

(1) List the area units, with the estimated size M'_i shown against each.[4/]

3/ In a unit 10 times as large one would be selecting a fraction 10 times as small, i.e., the same actual number. More generally: if in the i-th area m_i households are selected from a total of M_i and if the sampling fraction f_i is made to vary inversely with M_i, then $f_i = k/M_i$ where k is constant. But $f_i = m_i/M_i$. Therefore $m_i = k$ = constant.

4/ M', with a prime, is used here instead of M to signify that the size is estimated. In the next section M (without a prime) is introduced to represent the actual size.

(2) Cumulate the values of M'_i and enter this cum M'_i against each unit. Check that the last entry equals the sum of the M'_i values, namely M'.

(3) If a is the number of area units which it has been decided to select, compute the sampling interval $I = M'/a$.

(4) Select a random number between 1 and I. Let this be R.

(5) Compute the sampling numbers

$$R;\ R + I;\ R + 2I;\ R + 3I;\ \text{etc.}$$

(6) For each sampling number, find the first cum M'_i which equals or exceeds it. The corresponding unit is the one selected.

It will be seen that this procedure describes systematic sampling with probability proportional to the values M'_i. Below is an example:

Example 2.4.1

	Area Unit	Size M'_i	Cum M'_i	Sampling Numbers
	001	150	150	90
	002	60	210	
	003	80	290	
	004	70	360	
	005	130	490	
	006	90	580	
	007	110	690	640
	008	140	830	
	009	150	980	
	010	70	1,050	
	011	140	1,190	1,190
	012	110	1,300	
	•	•	•	
	•	•	•	
	•	•	•	
End	350	90	38,500	
Total		38,500		

Planned sample size: a = 70
Sum of unit sizes: M' = 38,500
Interval: I = M'/a = 550

Random number selected between 001 and 550: R = 090

Units selected into the sample: 001, 007, 011, etc.

Computation of sampling fraction for households

We assume a two-stage sample with area units at the first stage and households at the second. In each selected household all eligible women are to be interviewed.

Let a be the number of area units (PSUs) to be selected, let M'_i be the estimated size and M_i the actual size of the i-th area unit, measured in terms of the number of households.[5/] We write M' and M for the totals $\Sigma M'_i$ and ΣM_i.

As we have seen, the PSUs are selected with systematic sampling using the interval $I_1 = M'/a$. With the sampling points distributed at this interval, it is clear that a unit of size M'_i has a chance M'_i/I_1 of being selected. Thus, the first-stage probability is

$$p_{1i} = M'_i/I_1 = aM'_i/M'$$

In working out the sample design the first step is to fix the overall sampling fraction f. This must be the ratio of the number of women to be selected to the number of eligible women existing.[6/] If p_{2i} is the sampling probability at the second stage (that is, for household selection) in the i-th PSU, then since we want a self-weighting sample with overall probability f we have

$$p_{1i}\, p_{2i} = f$$

so that

$$p_{2i} = f/p_{1i} = f\, M'/(aM'_i)$$

5/ See the Important note, Page 33.

6/ In estimating the number to be selected we make an allowance of approximately 10 percent for interview nonresponse and undercoverage in listing. In estimating the number existing we generally use data from the latest census projected to the field work date. Note that the number of women per household need not be considered. The sampling probability for women is the same as that for households because we are going to interview all women in each household.

The sampling interval I_{2i} for household selection in the i-th PSU is

$$I_{2i} = 1/p_{2i} = aM'_i/(f\ M')$$

This must be computed for each PSU selected, then used for systematic sampling of households, according to the method described in Section 2.3.

Example 2.4.2

This example expands on Example 2.4.1. Together they show the complete sampling process for the first three area units selected in the urban sector of an imaginary country. For the sake of simplicity, the country has a population of less than 1 million.

GIVEN	Desired sample size	5,000 women aged 15-49
	Number of women aged 15-49 in country (extrapolated from census to survey date)	168,000
	Number of women aged 15-49 in urban sector	47,100
ASSUMED	Sample selected (allowing approximately 10% nonresponse)	5,500 women aged 15-49
	Desired cluster take (urban)	20 women interviewed
COMPUTED	Sampling fraction f =	5500/168000 = 1/30.55
	Women selected in urban sector	47100/30.55 = 1542
	Number of clusters selected (average of 22 women per cluster)	1542/22 = 70

Example 2.4.1 serves as a representation of the urban sample selection for the first stage (area units).

For the second stage (households) we proceed as follows:
Compute for each selected area unit the household sampling interval:

$$I_{2i} = 1/p_{2i} = a\ M'_i\ /\ f\ M'$$

where

$a = 70$

$f = 1/30.55$

$M' = 38500$

and M'_i is the estimated size, taken from Example 2.4.1.

This gives $I_{2i} = 0.0555\ M'_i$

For the three units selected in Example 2.4.1 this gives:

<u>Unit 001</u> $I_{2i} = 0.0555 \times 150 = 8.3$

<u>Unit 007</u> $I_{2i} = 0.0555 \times 110 = 6.1$

<u>Unit 011</u> $I_{2i} = 0.0555 \times 140 = 7.8$

The final step is to list the households in the selected areas and to select a systematic sample of households, using the above intervals and following the method of Example 2.3.1.

<u>Important note</u>: In introducing Examples 2.4.1 and 2.4.2 we have stated that the size measures M'_i used in the first stage selection are to be understood in terms of the number of <u>households</u> existing in the i-th PSU. However, they could equally well be the number of <u>persons</u> (i.e. the census population). Even though the second stage of sampling operates with households, and the M_i (without a prime) denote households, we have nowhere used the assumption that the M'_i are measured in terms of households. The above procedures apply <u>without any modification</u> if the M'_i represent population sizes; in particular there is no need to introduce any household size estimate at any point in the operations. In practice census data are more often available in terms of population than in terms of households and their use is at least as appropriate for the present purpose.

<u>Units selected with certainty</u>

In systematic PPS sampling with interval I, any unit whose size equals or exceeds I is certain to be selected. If the method is maintained in such cases, units larger than the interval may be selected two or more times.

Examining the list of PSUs before sampling begins, but after computation of the interval, will show whether there are many units of size greater than I. If there are very few, not more than five, perhaps the simplest solution is to split each such unit into two or more approximately equal subunits of size less than I. Initially, the split is made on paper only. The size measure for the original unit is divided equally among the subunits and sampling proceeds. Later the split is "materialized," either by drawing a line on the map of the area, or by identifying a suitable dividing line during the first field visit to the area.

It often happens that a substantial number of the units chosen to serve as PSUs are larger, than the interval I. In this case the choice of such a unit to serve as PSU was clearly inappropriate. It would be better to start over again. All of the units designated as PSUs, greater than a certain threshold size (which need not be exactly equal to I), should be pulled out of the list before sampling, to receive special treatment. Assuming we wish to maintain the same sampling interval then a part of the sample will fall into each one of these units. They are not, therefore, sampling units but strata, by definition. A new, smaller type of sampling unit has to be designated to serve as PSU within these areas.

The solutions described above will eliminate all cases of "sampling units selected with certainty," a phrase which is logically contradictory, and will allow systematic PPS selection to proceed in the manner described earlier. Nevertheless, certain statisticians prefer to retain the fiction of "sampling with certainty" by referring to such units as "self-representing PSUs." For the purpose of sampling error computation it is important to realize that this terminology is misleading. The self-representing units are in fact strata, while the new, smaller units within them are the true PSUs.

2.5 A Practical Model Sample with Variants[7/]

Exact PPS self-weighting sampling, leading to an exactly fixed take in each cluster, is not feasible in the context of DHS surveys. The available measure of "size" for selecting the area units will, in general, be the population figures of the last census. Several factors intervene to make this no more than a rough approximation to the size measure that would result in a fixed take of women per cluster:

- The census may have been inaccurate. In any case it is always to some degree out of date by the time of the DHS survey.
- The census areas may not always be correctly identified when used for the survey (mapping error).
- Sizes are given in terms of population but the second stage of selection operates with households and the number of women per household varies.
- Nonresponse and undercoverage, whether at the household or individual level, introduce a further source of variation.

In practice, therefore, it is not possible to achieve anything like a constant take of women per cluster when using the PPES self-weighting design in the DHS context. WFS experience suggests that the take will typically vary with a coefficient of variation of about 0.4.[8/] This is a very substantial degree of variability.

7/ Throughout this section we assume that the survey design calls for the interview of every eligible woman in each sample household.

8/ The Coefficient of variation (CV) is defined as the ratio of standard deviation to the mean. For example, with CV = 0.4, in a sample with a mean take of 35 women successfully interviewed, the standard deviation is 0.4 x 35 = 14 and one could expect 10 percent of the clusters to show a take less than 12 or greater than 58.

In practice, then, the available measures of size provide only an approximation to the ideal, and PPES sampling serves only to control the most extreme variations in area unit sizes. It is reasonable therefore to treat these size measures in a very approximate way, using them in broad size groups instead of exact figures. This suggests the possibility of the "standard segment" strategy outlined in Section 1.4 and described below:

The standard segment design

The first step is to fix a standard segment size, as small as seems practical, for example, 500 population. Segments that are too small will be difficult to map.

Census enumeration districts (EDs) are used as first stage units. On the basis of the census, each ED in the country is allocated a number s_i of segments by dividing its population by the standard segment size and rounding to the nearest whole number. At this stage no segments are actually identified. If an ED is so small that s_i would be zero, it is combined with the next one on the list. Exception: if such a small ED is the last listed in a region or stratum, it should be combined with the preceding one.

The EDs are then sampled with PPS where the "size" is the number of segments, s_i.

A field operation is then mounted to create the exact number s_i of segments required in each selected ED, and one of these segments is selected at random.

The sampling probability at the first stage is

$$p_{1i} = ks_i \text{ where } k \text{ is a constant}$$

and the second stage probability is

$$p_{2i} = 1/s_i$$

The overall probability is therefore

$$p_{1i}p_{2i} = k$$

Since this is constant we have a self-weighting sample of segments. Finally we introduce a third stage at which households are selected with a probability that is fixed everywhere.

An exactly equivalent procedure for ED selection, somewhat simpler in concept, is to list the EDs with their segments represented by X's. We then select a systematic sample among the X's. An example is shown below.

Example 2.5.1

Census sizes of EDs are assumed available in terms of households. Assuming 5 persons per household the standard segment size can be taken as 100 households.

ED Number	Size (Households)	Segments	Selection
001	140	X	
002	260	X	◄---
		X	
		X	
003	210	X	
		X	
004	40)	X	◄----
005	100)		
006	90	X	
007	390	X	
		X	
		X	
		X	◄---
008	etc.	etc.	

The segments are entered as X's and a systematic sample is selected at a fixed interval I. Any ED in which a segment is selected is considered selected. In the example we have I = 5, the random start is 2, and the EDs 002, 004 + 005 (combined), and 007 are selected.

The method of selecting EDs with probability proportional to their number of segments, followed by selection of one segment in each ED, is equivalent to a single stage sample of segments.

Creation of segments involves delineating them on a map of the ED. Within each ED the segments should be approximately equal in population size, but it is more important to ensure that their boundaries are clearly defined. A warning on segmentation with this design is included in Note 2 at the end of this section.

With a standard segment of 500 population, any ED with a population less than 750 will have $s_i = 1$. It follows that if the average ED size is not too large, e.g. 900 or less, there is likely to be a substantial proportion of EDs which do not require segmentation. This should be borne in mind in planning the segmentation operations.

After selection of the segments a listing operation must be carried out in each segment. The choice between listing <u>dwellings</u> and listing <u>households</u> is discussed in Section 1.9.

<u>Variant 1: Two distinct area stages</u>

In the design described above the two area stages collapse into one. But this effect is achieved only by working out in advance, for every PSU in the whole country, the number of segments it will contain. This chore can be avoided by adopting a truly two-stage area sample.

At the first stage we elect EDs with PPS, using the census data, as they stand, to provide the measure of size.

In the selected EDs we create a number of segments (s_i), then select 1 at random. This is the second stage.

At the third stage we select households with a probability calculated so as to yield a self-weighting sample.

If M_i is the census size of the i-th ED and if we plan to select a EDs, then the first stage probability is

$$p_{1i} = aM_i/M, \text{ where } M = \Sigma M_i \text{ summed over all EDs existing.}$$

At the second stage we select 1 segment from among s_i, so that

$$p_{2i} = 1/s_i$$

The overall sampling fraction f, for all three stages together, is calculated from the ratio of the number of women to be selected to the number existing. Since the sample is to be self-weighting we have

$$p_{1i}\ p_{2i}\ p_{3i} = f$$

This gives

$$p_{3i} = f/\ p_{1i}p_{2i} = fMs_i/aM_i$$

Thus, we select households in the selected segment of the i-th ED by systematic selection with interval

$$I_i = 1/p_{3i} = aM_i/fMs_i$$

Note that with this design we are entirely free to choose the number of segments to be made, s_i, in each ED. However it is still desirable to have the segments of any given ED as equal in size as possible.

How does this design compare with the standard segment plan? There are three main differences:

(1) Although the measure of size M_i is used more exactly in the first stage it is doubtful whether this will yield any significant advantage in practice, for the reasons already mentioned.

(2) The computation of s_i for every ED in the country is avoided. However, we do have to cumulate the M_i over all EDs and there are more computations to be done at the level of the sample EDs. The gain here is at best marginal.

(3) Freedom to choose s_i in each ED should be an advantage if the mapping team is skilled. The key issue here is the clarity with which segments can be mapped. Where natural features for segment delineation are scarce it would be profitable to reduce s_i to ease the mapping problem. But to leave this decision in the hands of the field team responsible for segmentation may be unwise, first because the judgment involved is a difficult one and second because the team can save work for itself by reducing s_i, which is motivation for a wrong decision. This being so, there is a danger of ending up with a consistently smaller set of s_i than planned. This does not cause any bias, but it leads to inflated listing costs because the segments are bigger than planned. In some places, maps are so good that segmentation can be done in the office; here the advantage of freedom in the choice of s_i may be significant.

Variant 2: Standard segment with compact cluster

In the sample design strategies considered in this manual the desired size of the "take" (i.e., the number of women to be selected per ultimate area unit) is decided in advance. We suggested in Section 1.7 that this might reasonably be fixed at an average of about 40 women in the rural sector and 20 in the urban. Counting 1.25 women per household this would amount to 32 households and 16 households respectively. Let this target "take" be T.

If we could make segments of average size T it is arguable that we could avoid the listing operation altogether by using "take-all," or "compact cluster" sampling at the last stage. The objection to this is that it is very difficult to map such small segments with clarity because there are not enough natural features to provide the boundaries.

The following design provides the option of creating such small segments aimed to be about the size of the target take, T. This option is taken up where convenient, but where there are no suitable boundaries the segment maker is allowed to create segments of a size consistent with good mapping and hence to minimize the cost of listing.[9/] It also minimizes the interviewer's walking time between households. On the other hand it maximizes the intraclass correlation effect, that is, the increase in error variance due to the correlation between neighbors in the sample. The balance of advantage here is likely to be positive.

Like the standard segment design already described, this design begins by imagining every PSU (normally a census ED) divided into s_i standard segments, but in the present case the standard segment is much smaller and equal to T. Note that T, originally worked out in terms of women aged 15-49, needs to be converted to a population basis (i.e. the number of persons of all ages and both sexes) and then adjusted back to the census date. For the first operation the latest census can be used to give the ratio (total population)/(women aged 15-49); typically this will be about 4.5. For back-dating to the census we can assume that the population growth rate applies also to the growth rate for women aged 15-49. Let the new value of T based on these adjustments be T'. Then s_i is obtained by dividing the census population of the PSU by T' and rounding to the nearest whole number.

9/ In Section 1.8 we suggested that listing was desirable even with compact cluster sampling, to provide adequate control of the interviewer's coverage of the segment. If this is accepted, compact cluster sampling still has the advantage of minimizing the amount of listing.

The first stage of sampling consists of selection of PSUs with probability proportional to s_i. This may be achieved by working out s_i for every PSU in the census list. However, an adequate approximation, which should involve less work, would be to select with probability proportional to census population. Since s_i is proportional to the population the only error is one of rounding. In the present case, segment size is so small that this error can be tolerated. In any case, whether or not s_i is computed explicitly for every PSU in the sampling frame, it must be computed for every PSU selected.

Up to this point the design does not differ essentially from the standard segment design already described. But from now on the two diverge.

Instead of creating the s_i standard segments in each selected PSU, an intermediate area unit is introduced which is essentially a group of standard segments. We call this unit a <u>division</u>. The selected PSU must be mapped into not more than s_i well defined divisions. Clarity of boundaries is here given absolute priority and there is no longer any requirement that these area units be approximately equal in population. Within the i-th PSU the divisions are numbered j = 1, 2, 3 ... so that any one division is identified by the pair of numbers i, j.

Once the divisions have been delineated each one must be allocated a whole-number measure of size s_{ij} such that $s_{ij} = \Sigma s_i$; that is, the s_i standard segments in the PSU are <u>allocated among the divisions</u> in proportion to the size of each division, and with at least one in each division.

For example, suppose a PSU has $s_i = 8$. Suppose it is found feasible to divide it into only 4 divisions. After these have been delineated their sizes are examined and the 8 imaginary segments are allocated among the 4 divisions in proportion to each division's estimated population size. This might lead, for example, to the allocation: 1, 2, 2, 3. These figures then become the size measures s_{ij} of the divisions.

The next step is to select one division, with probability proportional to s_{ij}, in each PSU.

Finally, the households are listed in the selected division and a fraction $1/s_{ij}$ of them are selected. Thus, if a division has s_{ij} = 1, we take every household; if s_{ij} = 2, we select every 2nd household systematically; if s_{ij} = 3, every 3rd; and so on.

The sampling probabilities are as follows.

<u>1st stage</u>: PSUs

$$p_{1i} = as_i / \Sigma s_i \text{ (or } aM_i/M)$$

where a = number of PSUs selected, M_i = census size of the i-th PSU, M = census total, $s_i = M_i/T'$ to the nearest whole number.

<u>2nd stage</u>: Divisions

$$p_{2i} = s_{ij}/s_i$$

<u>3rd stage</u>: Households

$$p_{3i} = 1/s_{ij}$$

<u>Overall</u>: $$f = p_{1i}\, p_{2i}\, p_{3i} = \frac{as_i}{\Sigma s_i} \frac{s_{ij}}{s_i} \frac{1}{s_{ij}} = \frac{a}{\Sigma s_i} = \frac{aT'}{M}$$

It follows that f is constant throughout, so that the sample is self-weighting.

A minor variant would be, at the household selection stage, to select a <u>run</u> of consecutive households from the list instead of using systematic selection. For example, if s_{ij} = 3 we divide the list into three equal parts and select one at random. Since compact cluster sampling is the ideal for which this design is aiming, this modification appears more consistent with that aim than would systematic selection (see Note 1 on run sampling at end of this section).

Apart from its relative complexity, the main problem with this design is how to estimate the sizes of the divisions. The usual solution is a "quick count," carried out in a rough manner, for example by counting houses (rural) or dwellings (urban). But this must be done for the whole PSU. How much difference is there, in person-hours of work, between a quick count and a full listing? If the latter were used we could go straight to the compact cluster by dividing the list into s_i runs and selecting one; there would be no "divisions" and indeed no mapping to be done within the PSU. It seems doubtful whether the difference between a quick count and a full listing can more than pay for the mapping work, the various calculations, and the full listing within the selected division which is in any case unavoidable.

Summary

The standard segment design is simple in concept and easily carried out in the field. The two variants considered, two stage segment design and standard segment with compact cluster, are both acceptable methods, although they are more complex and offer only marginal advantages.

If compact cluster sampling is desired, the simplest method is still the standard segment design with run sampling from the household list.

NOTES

1. Run sampling from a list

Systematic sampling at the household stage maximizes the amount of walking to be done by interviewers between interviews. If the settlement pattern is very dispersed, or the household sampling fraction very low, this may be unacceptably costly. In such cases the alternative of run sampling for household selection may be preferred. The following procedure is recommended.

Let f_i be the required household sampling fraction computed for the area i.

Compute $1/f_i$ and round to the nearest whole number, $= s_i$. Divide the list into s_i equal (or nearly equal) runs by drawing $(s_i - 1)$ horizontal lines across the list at equal intervals. Number the runs and select one with a random number.

A frequently used variant is to make twice as many runs (compute $2/f_i$ and round) and select two from the list. This not only reduces the error variance arising from intraclass correlation but also reduces the rounding error in s_i, which may be substantial where f_i is large. If f_i 0.5 even this can give an unacceptably large rounding error. Perhaps the simplest rule is to revert to systematic sampling for household selection in any area i for which the computed f_i exceeds 0.5; otherwise, use the method of selecting two half-runs just described.

2. A common error in PPS standard segment sampling

In this design the size measure s_i for PPES sampling is computed by dividing the census population M_i by the standard segment size. The first stage probability is $p_{1i} = ks_i$.

At the second stage we select 1 segment out of the s_i created, giving a second stage probability $p_{2i} = 1/s_i$.

When these two are multiplied together the s_i cancels so that the overall probability is constant.

A common error is to use the wrong rule for segment creation. Instead of working to create the designated number of segments s_i, the segmentation workers believe that they are required to create segments of the standard size. As the population may have changed since the census, this may yield a number of segments different from s_i. In this situation the s_i will not cancel when p_{1i} and p_{2i} are multiplied together and the sample will not be self-weighting.

The error seems to occur because the concept of the standard segment is overstressed when instructing the segmentation personnel or their supervisor. It is important to emphasize to them that in all cases they must create exactly the designated number s_i of segments.

2.6 Subsampling from an Existing Sample or Master Sample

The problems involved in subsampling from an existing sample or master sample were described in Section 1.5. Here certain aspects are presented in detail.

Selecting a self-weighting subsample from a non-self-weighting master sample

It often happens that the master sample from which one wishes to subsample is non-self-weighting. In this section we suppose that the master sample is divided into domains h = 1, 2, 3 ... for which the overall sampling fractions are f_h.

We denote probabilities or sampling fractions for the master sample by p or f, and the corresponding target values required for the new DHS sample by p' or f'.

The problem is simply to find subsampling rates f* which will yield an equal probability sample. In general, the probability for the DHS sample will be the product of the probability for the master sample and the subsampling probability, that is, $f' = ff^*$.

Applying this to the domains, we have

$$f'_h = f_h f^*_h$$

Since the DHS sample is to be self-weighting, f'_h is to be constant f'. Thus:

$$f^*_h = f'/f_h$$

The overall sampling fraction f_h is the product of the two sampling probabilities $p_{1,h}$ for area selection and $p_{2,h}$ for household selection.

Normally, the subsampling with rate f^*_h will be carried out at the area sampling stage. In this case, the last equation applies directly to the area stage, so that we can write:

$$p^*_{1,h} = f' / f_h$$

where $p^*_{1,h}$ is the subsampling rate at the area stage. However, it may sometimes happen that the master sample is designed to yield a cluster take, in one or more domains, which differs from that considered optimal for DHS. In this case the 2nd stage sampling probabilities for the master sample and for the DHS sample will differ in the same ratio. If b_h is the target cluster take for the master sample and b'_h that for the DHS sample then we have:

$$p'_{2,h} / p_{2,h} = b'_h / b_h$$

Using the fact that the overall sampling fraction f is the product of p_1 and p_2, we obtain:

$$p'_{1,h} / p_{1,h} = (f'_h/p'_{2,h}) / (f_h/p_{2,h})$$

$$= (p_{2,h}/p'_{2,h}) (f'_h/f_h)$$

$$= (b_h/b'_h) (f'/f_h)$$

In the last line we have once again replaced f'_h by f', on the assumption that the DHS sample is to be self-weighting.

The ratio on the left is the subsampling rate $p^*_{1,h}$ required at the area stage. As long as this does not exceed 1 in any domain, the desired subsampling at the area stage can be achieved. If it does, one might allow some increase in b'_h, the cluster take in DHS, but there is a limit to the flexibility that can be accepted here. If this fails to meet requirements, the subsampling cannot be achieved within the constraints set. Even then, there is still the possibility to be considered of augmenting the sample by a supplementary selection in the domains where the constraints cannot be met.

Updating the listing

If the lists of dwellings provided by the master sample are more than two years old, they will need updating.[10/] Since updating requires visiting every dwelling, and in view of the temptation for field workers to report "no change" for their own convenience, it is preferable to organize an independent relisting.

The need for a relisting removes most of the advantage offered by a preexisting sample. However, two possible benefits remain:

(1) There is a savings in mapping and segmentation. Where segmentation has been done for the master sample this savings is considerable.

(2) Use of a common sample allows linkage between the surveys concerned, with enhanced analytic potential. If only the areas and not the households are in common, this advantage is, however, modest and studies profiting from such a linkage are rare.

Disadvantages of sample overlap

Where the same households are interviewed in two or more surveys there are the potential problems of respondent resistance and contamination.

10/ The limit is generally lower in unplanned urban areas, where even one year may be too long.

Respondent resistance is rarely a problem in developing countries, where response rates are typically higher than in industrialized countries. Only if the earlier survey has imposed a heavy burden on the respondent (as in the case of certain household economic surveys and nutrition surveys) should one expect to encounter resistance when a new survey is implemented.

Contamination is the feedback of influence from the interviewer to the respondent. Such effects are often weaker than expected. Experiments show that people remember little of the detail of an interview and do not change their behavior because someone asks them a few questions. However, some types of question are particularly subject to contamination. Questions testing knowledge which are expressed in the form "Have you heard of X?" cannot reasonably be asked a second time since X was mentioned by name on the first occasion. Even if the effect of contamination is small in reality, one cannot be sure of this; the mere potential of contamination reduces the value of the second response.

If contamination problems appear serious in a particular case, what measures can be taken to avoid interviewing the same household twice in the event that a common master sample is used by two surveys?

If the same household list is used by both surveys and if the sum of the cluster takes for the two surveys never exceeds the total units existing in the ultimate area units, there is no difficulty in preventing overlapping. The simplest solution is to select the household sample for the second survey only after removing the households selected for the first (the household sampling fraction will need adjustment to yield the same number of households as would be obtained were the full list used).

If the household list is updated or renewed after the first survey the new list would somehow have to be matched to the old. The practical difficulties here are serious. Not only will the volume of work in matching be considerable, but there will be uncertainties and ambiguities

which may further lead to biases. For example, is this really "the same household" when some of its members have gone away? It is not advisable to attempt such an operation in a developing country.

In some master samples there are two or more stages of area sampling. In this case it may be possible to select for the second survey a systematically different sample of ultimate area units (UAUs) within the same sample of penultimate area units. Where the UAUs are segments this still has the advantage of saving work on a new segmentation operation.

This solution was adopted in one of the earliest DHS surveys, however, a problem arose. A small percentage of EDs contained only one segment, leaving no alternative selection for the second survey. It was decided to use the same segment in these few cases, to examine the results for contamination effects, comparing such repeated segments with new segments, and if a significant difference was found to remove this small subgroup from the analyses for variables subject to contamination.

2.7 Errors of Coverage and Nonresponse

By coverage error we mean lack of correspondence between the sample as designed and its implementation in the form of attempted interviews. Nonresponse, on the other hand, relates to interviews attempted but not achieved. Thus, if an interview is, erroneously, not even attempted we call this an error of undercoverage; if it is attempted without success we term it nonresponse.[11/] This section deals with problems in the definition and estimation of such error rates.

11/ Sometimes an interview (or group of interviews) is deliberately not attempted because it is known that the selected interviewee is unavailable or inaccessible. We class this with nonresponse: failure to attempt the interview was not "erroneous," and the overriding factor was the impossibility of achieving it.

Coverage error

Errors of overcoverage, though possible, are rare. A mass of evidence indicates that a net undercoverage is widespread in developing country surveys. Coverage rates of 90 percent appear to be quite common.

The explanation of such systematic failure has not been clearly identified but two main sources of error appear to be involved. The first arises at the listing stage: in many surveys listing workers appear to have covered less than the designated area.[12/] Secondly, where an age limit is used to determine eligibility for interview, age distortions are often seen which have the effect of excluding persons from the eligible age range. Both of these errors appear to involve motivated bias by field workers seeking to reduce their workload.

Motivated errors can only be controlled by intensive training and close supervision.[13/] Note, however, that age distortions around the eligibility limits, though motivated, are not necessarily conscious or deliberate. In many developing country surveys the age of certain older respondents is estimated by a process not far from guessing. Training interviewers to guess "objectively," without bias, is difficult. In such cases bias cannot be eliminated entirely.

12/ It is possible that in some cases the shortfall in the number of households listed arises because the area sampling frame, based on the census, omits areas in which construction has taken place since the census, for example areas on the edges of cities.

13/ In principle, motivated age distortion in a survey like DHS could be controlled by using different field workers to list household members and to interview women respondents. This was done in a few WFS surveys but the additional cost would be prohibitive in DHS, where there is no call for an independent household survey.

Error due to an outdated area frame can be reduced by taking special steps to update the frame in areas of known new settlement, in particular, new housing estates or squatter areas on the outskirts of cities, and camps housing refugees (if these are to be included in the survey).

Coverage errors can be investigated after the survey by a variety of methods. The sample can be extrapolated to the total population and the last census can be extrapolated to the survey date for comparison. This check should be done separately for the number of households and the number of eligible women.[14/] Age distortions can be investigated by studying the discontinuity in trends across the eligibility boundaries 14-15 and 49-50. While it is tempting to introduce males as a control here, it should be noted that in most societies more males are educated than females, so that heaping at ages 15 and 50 will be less extreme for males.

Deliberate restriction of coverage

In many surveys, whether in developed or developing countries, certain parts of the national territory are deliberately excluded from the survey for reasons of difficult access. Two distinct cases arise:

(1) Exclusion of clearly identified areas from the sampling frame - Here it is usual to state the coverage limitation in the survey report, which then becomes a report on the remainder of the country. Such exclusions are not regarded as coverage or response error but simply as part of the definition of the survey domain.

14/ The growth rate for women aged 15-49 can be reasonably assumed equal, or slightly above, that of the population. But the growth rate for households typically will be substantially smaller.

(2) Ad hoc exclusions decided during or just prior to field work - In developing countries it is not uncommon for the survey organization to abandon the attempt to conduct field work in certain sample clusters, whether due to floods, civil disturbance or other practical constraints. Here the exclusions usually occur after sample selection. If such excluded areas form a meaningful domain it may be acceptable to deal with the problem by redefining the survey domain, in the manner of case (1) above. More commonly, however, the excluded areas will not "make sense" and will have to be accepted as constituting error. This should be classified as nonresponse rather than coverage error, in line with the definition at the beginning of this section.

Nonresponse

At first sight, the concept of nonresponse seems simple and clear--what percentage of the persons who should have been interviewed were not interviewed. Taking account of the distinction between coverage error and nonresponse indicated earlier, we can modify this by saying that we want to know what percentage of attempted interviews failed.

In practice, there are two features, found in some sample designs, which complicate this simple issue.

First, in many surveys the final units for interviews are identified through a progressive sifting process. For example, in the typical DHS we list and select dwellings, interview the household currently in the dwelling, then interview any women aged 15-49 in that household. If failure occurs at one of the earlier steps, we do not have the information which would enable us to classify the effect at the final level. For example, if the interviewer cannot find the selected dwelling we do not know whether it contains a woman eligible for interview; if it does not, then the failure has no effect on the interview response rate.

Second, in some surveys there is a policy of replacing nonrespondents. Although this is not allowed in DHS, there is a similar arrangement that is allowed, indeed recommended. Where the household in the selected dwelling moves away between the listing and the interview, DHS recommends interviewing the household (if any) that moves in to replace it. Should such interviews be counted as successful?

To deal with the first problem, we begin with the case in which there are only two steps in the sifting process, namely households and women. There are four quantities of potential interest in computing response rates:

A Households selected

B Households interviewed

C Women selected

D Women interviewed

Since the survey concerns women, the relevant response rate is D/C. But the quantity C is not known. We would like to know the number of eligible women in all selected households but we only know the number in the households interviewed, say C'. We therefore estimate C as follows:

$$\text{Est. } C = C' \times \frac{A}{B}$$

This assumes that the number of eligible women per household is the same among nonrespondent as respondent households. This assumption is not very convincing, but the effect of any departure from it on the estimate of C is likely to be very small. On this basis the response rate, D/C, becomes:

$$\frac{B}{A} \times \frac{D}{C'}$$

It will be seen that this is the product of the response rates observed at the two respective stages, households and women. This basic principle gives the solution to our first problem. Where two or more steps of sifting are involved we can estimate the overall response rate by multiplying together the response rates observed at each step. In doing so we assume that the response/nonresponse outcomes at the different steps occur independently.

Turning to the DHS, and assuming the dwellings households women progression already mentioned, it follows that we need to compute a dwelling response rate as well as the household and women response rates. "Dwelling nonreponse" can only refer to the categories dwellings destroyed and dwellings not found between the listing and the interviewing. If we represent these by P and Q respectively, with E = dwellings selected, then the dwelling response rate is (E - P - Q)/E, so that the overall response rate is:

$$R = \frac{E - P - Q}{E} \times \frac{B}{A'} \times \frac{D}{C'}$$

where A' relates to the households in the dwellings found and still existing. In practice, P and Q are always very small categories and it is usual to collapse the dwelling and household steps into one. We assume that dwellings in category P have no households while those in Q are typical of other dwellings: thus we drop the first term above but add in the category Q (dwellings not found) to the household base A'. This gives:

$$R = \frac{B}{A' + Q} \times \frac{D}{C'}$$

Turning to the problem of replacement, this is easily settled by looking more closely at the logic of the design. The design calls for the listing and selection of dwellings, and then for the interview of the household found in the dwelling at the time of the survey. Since in many areas there is no address system, the initial listing operation has to identify the dwellings in terms of the names of the occupying households, but these merely serve as addresses. The fact that, in some cases, a new household moves in, between the time of listing and interview, does not mean that replacement of a sampling unit has occurred. Thus, such cases do not require any special treatment. Moreover, just as a new household moving in does not constitute a replacement, so the case of a household moving out after the listing without another moving in does not constitute nonresponse. The target household sample is defined as the set of households found at the time of interviewing, in the dwellings selected from the dwelling list.

Operational procedures

It remains to operationalize these conclusions in the form of response codes to be entered on the questionnaires and field records, and to express the formulas for response rates in terms of such codes.

We assume that the listing operation includes unoccupied dwellings and that the household survey is on a de facto basis, so that a household temporarily absent (specifically, absent the night before the survey) is not in the target population.

At the household level, the following response categories can be imagined:

1H Completed

2H Household present but no competent respondent at home

3H Household absent night before interview

4H Postponed

5H Refusal

6H Dwelling vacant or address not a dwelling

7H Dwelling destroyed

8H Dwelling not found

9H Other

Note that "household" above refers to any household found in the dwelling, not necessarily the household named at the time of the listing operation.
The household response rate is then:

$$\frac{(1H)}{(1H) + (2H) + (4H) + (5H) + (8H)}$$

At the individual level the following response categories are possible:

1I Completed

2I Not at home

3I Postponed

4I Refused

5I Partly completed[15/]

6I Other

15/ Partly completed. These cases will ultimately be classified (1I) or (4I) at the data entry stage, depending on whether or not they are retained in the data file.

The individual response rate is thus:

$$\frac{(1I)}{(1I) + (2I) + (3I) + (4I)}$$

It will be noted that we have not included the category, "no eligible woman in the household." This category is irrelevant to the response rate, appearing neither in the numerator nor the denominator (analogously to category 3 at the household level). We omit it for this reason and assume that no woman's questionnaire will be provided for such cases.

Finally, it will be seen that we have omitted the category "other" (9H and 6I) from the response rate formulas. This is an approximation introduced for the sake of simplicity, on the assumption that the category is hardly ever used.

It would be desirable to have the interviewers specify in writing the reason for nonresponse, whenever the "other" code is used. The number of "other" cases should then be counted and if, at the household level and interview level together, they exceed one percent, an attempt should be made to allocate them according to the following principles:

(1) If the household or woman was correctly selected and ought to have been interviewed, the case should be counted in the denominator.
Examples:

Omitted due to time pressure
Omitted in error
Sick or incapacitated

(2) If the household or woman was wrongly included in the list or sample, the case should be excluded altogether from the formula.
Examples:

Dwelling does not lie within the sample area
Woman over or under age

(3) Any others which still resist classification can be dropped from the formula.

The overall response rate is obtained by multiplying the household and individual level rates. However, one further adjustment will be required if there has been a deliberate exclusion of certain areas (assuming that this has not been absorbed through a redefinition of the survey domain (see subsection on coverage error above). In such cases it is recommended to estimate the proportion p_0 of the population so excluded and to multiply by the further factor $(1 - p_0)$.

In summary, the final overall estimated response rate is obtained from the formula

$$R = (1 - p_0) \times \frac{(1H)}{(1H)+(2H)+(4H)+(5H)+(8H)} \times \frac{(1I)}{(1I)+(2I)+(3I)+(4I)}$$

Such response rates should be computed and published separately for the main geographical domains of the sample as well as the whole survey domain. If the sample is self-weighting within domains but has different weights in different domains, the response rates should be computed and published for each differently weighted domain, and the whole survey rate should be completed and published both with and without weighting.

2.8 Sampling Errors

The sample designs for surveys in the DHS program involve features such as stratification, clustering and sometimes weighting. All these features affect the sampling errors of the survey estimates. The standard formulas for sampling errors found in statistical texts and used in most computer programs do not take account of these features, and hence their use with DHS surveys is inappropriate. These formulas (such as $\sqrt{(pq/n)}$ and $S/\sqrt{n}$ for the estimated standard errors of a proportion and a mean respectively) apply only with samples selected by simple random sampling with replacement. Their use with DHS surveys would generally tend to understate the sampling errors of the survey estimates, often to an appreciable extent.

The computations required to estimate sampling errors of survey estimates based on complex sample designs like those used for DHS surveys are more burdensome than those based on simple random samples. In recent years a number of computer programs have been developed to perform these computations. One of these programs, such as the WFS CLUSTERS program, will be used to compute sampling errors with DHS surveys. The correct computation of sampling errors with complex sample designs requires knowledge of the stratum and primary sampling unit to which each sampled individual belongs. It is therefore essential that this information be recorded on each individual's computer data record; otherwise programs such CLUSTERS cannot be used.

A DHS survey produces an extremely large number of estimates, many of which are presented as percentage distributions in tables. Analysts require standard errors for these estimates, and also for differences between estimates. They want to know, for example, not only the standard errors of the percentages of educated and less educated women with a particular characteristic, but also the standard error of the difference in the percentage of educated and less educated women with this characteristic. Even with the availability of computer programs to compute the standard errors, it is not possible to compute standard errors for all the survey estimates of interest; and even if it were possible, the inclusion of all of them in the survey report would make the report unwieldy. For these reasons, only a selection of sampling errors will be computed with DHS surveys. Generalized sampling error models will then be developed from the computed standard errors to enable readers of the survey report to infer the standard error of any estimate in which they are interested.

Suggestions regarding the selection of variables and categories for publication of sampling errors, and on their mode of presentation, will appear when DHS publishes its recommendations on the tabulation and analysis of survey data.

2.9 Sample Description and Documentation

This section specifies the requirements for describing and documenting DHS samples.

Minimal sample description

In any report on survey results it is usual to include a brief description of the sample. Such a description should include the following:

(1) Statement that the sample is a probability sample (see definition on p. 1).

(2) Statement that the sample is stratified. Details of stratification are not required. In practice almost all samples are stratified, but the fact is worth mentioning to confirm that technical standards are being maintained.

(3) Number of sampling stages. Count "effective" stages only (see Note 1 at the end of this section). If the number of stages differs between strata, mention this.

(4) If feasible, identify the nature of the units used as primary (1st stage) sampling units (PSUs). In any case, give the number of PSUs in which the survey was conducted.

(5) Statement regarding self-weighting. If the sample is not self-weighting there should be a brief statement of how the weights vary.

(6) Statement regarding coverage. If any substantial sector of the national population (e.g., over 3 percent) is excluded from the sampling frame this should be stated with the percentage of population involved.

(7) Total final sample. Number of households and women successfully interviewed, plus overall response rate (see Section 2.8 for definitions).

(8) If the sample is based on a master sample, or on the sample of another survey, this should be stated.

In practice, this information can be covered in two or three sentences. Two imaginary examples are given below.

> The sample for the Zimpopo DHS was a 2-stage stratified self-weighting probability sample representative of the entire country. A total of 4,421 households and 5,309 women in 123 census enumeration areas were successfully interviewed, with an overall response rate of 93.4 percent. The a real sample was a subsample of the Zimpopo National Master Sample.
>
> The sample for the Ambrosia DHS was a stratified probability sample covering the national population, with the exception of the Island of Polyp (9 percent of the national total). Two sampling stages were used in the urban sector and three in the rural, the primary sampling units being respectively blocks (134 selected) and census enumeration districts (106 selected). The sample was self-weighting within each sector but the urban selection probability was twice the rural. A total of 5,124 households and 6,106 women were successfully interviewed, the overall response rate being 91.7 percent.

Note that the field work dates should always be included, whether in the sample description itself or in an accompanying account of the survey operations.

Full sample description

A complete description of the sample design should be written as soon as the design is finalized and the areas have been selected. It should be updated immediately after field work in order to take account of departures from the plan as well as nonresponse. A final update should be made after the first marginals have been run, since some questionnaires may have been rejected or lost between field work and data entry.

The description should include all the items mentioned above, but in full detail. Below is a checklist of points to be covered, divided for convenience between qualitative items requiring verbal description and numerical items. Note that the checklist order is not a particularly suitable one for presentation of the sample description. If the description is to be published, it should be presented in a logical order showing a natural development. A recommended schema would be, first, a short summary

description similar to that outlined above, and secondly, a complete account of each sampling stage, working down the hierarchy of stages from PSUs to individual women.

Verbal description

(1) Identify nature of sampling unit for each stage. If different for different domains, give details.

(2) Describe sampling frame used at each stage. Two cases are possible:

(a) Sampling frame already exists (e.g., census EDs). In this case give date when frame was made or when last updated.

(b) Sampling frame must be made. Describe the frame creation operation (e.g. segmentation, in office or in field?; listing, dwellings or households?). Give the date of the sampling frame. NOTE: The time interval between creation of the frame and use of the frame is an important parameter. If small area units are to be combined to make larger ones, describe this.

(3) Describe any explicit stratification used (see Note 2 at end of this section). Explain the basis of stratification and identify the strata.

(4) Note if systematic selection is used, at least in the first sampling stage. Systematic selection is equivalent to implicit stratification (see Note 2); therefore, describe the basis on which the 1st stage list or frame is ordered (e.g., geographical, or alphabetical, by size of unit, etc.).

(5) Describe how the selection probabilities vary or stay constant. If selection with probability proportional to size is used, identify the "size" variable. Here are three examples:

(a) Fixed probabilities at each stage

(b) 1st stage probability fixed within each region but varying between regions in the ratio:

North : Center : South = 3 : 2 : 1

Remaining stages: probabilities constant throughout

(c) 1st stage: probability proportional to size ("size" = 1980 census population).

2nd stage: probability inversely proportional to 1st stage probability.

The verbal description should normally include some discussion of the reasons for choosing the different design features.

Numerical parameters

(1) Where selection probabilities are fixed, give these (distinguish between conditional and overall probabilities). Where they vary within one sampling stage and where these variations are compensated by opposite variations at a later stage, it is sufficient to give the overall probability for the stages taken together, with the formula and data enabling the individual stage probabilities to be calculated.

(2) For each of the explicit geographical strata give the sample numbers as in the following table:

	Number of area units selected at each area stage			Number of households		Among Households Successfully Interviewed...	
Stratum	1st	2nd	..	Selected	Successfully Interviewed	Number of Women Aged 15-49 Listed	Number of Women Aged 15-49 Interviewed

(3) Specify any weights recommended for use.

The full sample description should be available in the country and in DHS files for every survey. It will be too detailed for publication in a short report of survey results, but it should be included in any full methodological report of the survey.

Finally, where children are sampled (e.g., for anthropometric measures) or husbands (for additional husbands' interviews) descriptions of these sampling operations should be added in similar terms to the above.

Sample documentation

Even the above full description of the sample -the maximum that can reasonably be published- will not normally include all the information that certain users may require. In particular, for computation of sampling errors and for linkage of survey data with other data sources, additional details are needed. This subsection specifies all the information that should be retained in the survey files to meet such potential needs.

The essential requirement is for the complete specification of the sample in terms of the number of units selected, with their selection probabilities (or the formulas and data from which these can be computed), in every explicit stratum and at every sampling stage. The records should also identify the position of each sampling unit in the sampling hierarchy, i.e., in which higher stage unit and in which stratum it falls. The numbers of units existing in the sampling frame are also useful for the first area stage, number of PSUs in each explicit stratum, and essential for the last stage, number of households listed in each ultimate area unit (UAU). Also needed are the number of interviews attempted and achieved in each UAU. Where systematic sampling is used at the first sampling stage, information is needed on the order in which the units were selected, that is, the order in which they appear in the frame. The ID numbering system used in the final record needs to be related, on the one hand, to the order of selection, and on the other, to any independent ID system, such as census ED numbers, which may be in current use and of potential relevance as a supplementary data source. These relationships may be reported either by including the alternative IDs in the data file or by providing a separate file showing the correspondence between two or more systems. Finally, for any domain within which the sample is self-weighting, the overall sampling fraction for the final units, households or individuals, should be reported.

Documentation should also identify any weights that appear on the data record and explain both their derivation and their use.

Provision of all this information may appear an arduous chore, but in fact, each one of these items serves a specific analytic need and each one has been used in one or more studies carried out by CPS or WFS. Some examples of such uses are given below. Meanwhile, note that most of the data required can be supplied by preserving the worksheets used for sample selection together with the instructions and computations used in completing them (sampling intervals and their derivation, strata definitions, etc.). Some further items will appear in the individual computerized data file (see below).

Need for specific items of sampling information

(1) Selection probabilities are needed for computation of weights. Even if the ultimate sample is self-weighting, there may be supplementary data available at the level of a sampling stage which is not self-weighting, or checks with independent data sources may be made at such levels, or subsequent surveys may be linked at such levels. For these reasons the probabilities are needed at all sampling stages.

(2) Detailed information on stratification and on systematic selection is needed for the computation of sampling errors. When systematic sampling is used, such computation often proceeds by the retrospective creation of implicit strata by grouping the selected PSUs in pairs. For this purpose one needs to know the order in which they were originally selected and how this order is represented in the survey ID system, as well as the presence of any breaks in the ordering such as occur at the boundaries of explicit strata. It is not uncommon that the area frame is reordered before selection to improve implicit stratification; it is also common to renumber the survey clusters before starting the data processing. Unless the reordering is fully documented, any such rearrangement threatens the validity of the process by which sample units selected consecutively are paired together to represent implicit strata (see Note 2 below).

(3) The number of PSUs existing in the frame, even if only approximate, enables the analyst to obtain the mean PSU size, a key feature in any understanding of the sample.

(4) The number of units listed is of value in estimation of cost parameters and for monitoring sample implementation.

(5) Numbers of interviews attempted and achieved are needed for computation of response rates (and possible corrective weighting).

(6) Recording of alternative IDs which relate the sample units to other data sources will be of value when such sources are used for checking or for improved estimation.

(7) Overall sampling fractions are needed for raising to national totals, whether for direct reporting purposes or for checking against census or other national level data.

Information of the kind specified above is obtained at three stages: when selecting the sample, at the field stage, and after production of the final data record. At all three stages the information must be given in terms of the same list of selected area units. Experience shows that this requirement for coordinated record keeping over different survey stages is not easily fulfilled by overburdened survey offices in developing countries.

The frequent occurrence of unforeseen problems of field implementation further aggravates the problem; these have to be dealt with urgently and are rarely documented. Special efforts are needed to ensure that all such problems are described, together with the solutions adopted, and to note any departures from the initial sample design.

Where a survey sample is based on a master sample, or on a sample selected for a prior survey, the information required may not be available. Since the DHS survey sample may in turn be used in future surveys, this emphasizes the long-term obligation to provide detailed sample documentation for the benefit of future users.

Sample documentation in the data file

Some of the sampling information mentioned above should appear on the survey data file for each individual; this includes:

Stratum identifier (at least for explicit strata)

PSU identifier (should appear in the record in the order of selection)

Weight (if any)

Documentation accompanying the data file should, of course, identify these items.

NOTES

1. Sampling stages

The concept of a "sampling stage" is not as clear-cut as it may seem. In the "standard segment" design, described in Section 2.4, we pointed out that the final sample of segments can be viewed either as a two-stage sample, with EDs at the 1st stage and segments at the 2nd, or as a single stage sample of segments. In the latter case, the EDs are introduced only to reduce the amount of work in making segments. They do not actually affect the final sample obtained and, in particular, they do not tend to cluster the selection of the subsequent stage, as a true 1st stage would. This situation arises typically when only one secondary sampling unit (SSU) is selected in each PSU. In such cases we say that the 1st stage is notional and there is only one effective a real sampling stage. In the minimal sample description mentioned above we recommended that only effective stages be mentioned. Thus, the standard segment design would be described as 2-stage, the 1st stage being segments of EDs, the 2nd dwellings/households.

Suppose that PSUs are selected straight from a frame of EDs, but that a few of the largest EDs are subdivided before sampling in order to reduce their size variation. This is essentially the same as the segment design except that the segmentation is carried out exceptionally, rather than usually. Again, it is a single effective area stage, though the sampling unit in this case would best be described as EDs or segments of EDs.

A further ambiguity arises over units selected with certainty, or "self-representing" units (see Section 2.4). Strictly, these are not sampling units but strata. If they occur during 1st stage selection, it is the level below that constitutes the 1st stage. Obviously, this complicates the sample description. We suggest that, where the number of such units is small (e.g. less than 5 percent), no special mention of them is necessary in the minimal sample description. Where the number is larger, or where the sample designer treated such units separately, the description should refer to this explicitly. Below is an imaginary example:

> In the smaller municipios (less than 50,000 population) the sample is 3-stage: municipios, sectores, dwellings. Municipalities larger than 50,000 appear with certainty and here the sample is 2-stage: sectores, dwellings.

2. Implicit stratification

When systematic sampling is used at the 1st stage, the standard formulas for sampling error computation are likely to overestimate the sampling error. This is because systematic selection spreads the sample evenly throughout the list; if there is any systematic trend on some variable as one goes down the list, the method is equivalent to stratification in terms of that variable. It is as though the list were divided into strata, with one PSU selected in each stratum.

Sampling error computations can take account of stratification provided at least 2 PSUs are selected in every stratum. When systematic sampling is used, the sampling error is usually estimated by assuming implicit strata, delineated after sample selection in such a way that each contains just 2 selections. If the number of units selected is odd, the last of the implicit strata is made with 3 selected units. The computation proceeds by taking the sample PSUs in pairs in the order in which they were selected. It is for this reason that the ordering in the original list needs to be preserved in the data file.

We thus distinguish between explicit strata, defined in the normal way, before sampling, and implicit strata defined after systematic sampling by the above procedure of pairing PSUs selected consecutively.