Demographic and Health Surveys

Phase III

Sampling Manual

DHS-III Basic Documentation — 6

Demographic and Health Surveys
Phase III

Sampling Manual

DHS-III Basic Documentation Number 6

Macro International Inc.
Calverton, Maryland

November 1996

The Demographic and Health Surveys (DHS) is a 13-year project to assist government and private agencies in developing countries to conduct national sample surveys on population and maternal and child health. Funded primarily by the United States Agency for International Development (USAID), DHS is administered by Macro International Inc. in Calverton, Maryland.

The main objectives of the DHS program are: (1) to promote widespread dissemination and utilization of DHS data among policymakers, (2) to expand the international population health database, (3) to advance survey methodology, and (4) to develop in participating countries the skills and resources necessary to conduct high-quality demographic and health surveys.

For information about the Demographic and Health Surveys program, write to DHS, Macro International Inc., 11785 Beltsville Drive, Suite 300, Calverton, MD 20705, U.S.A. (Telephone 301-572-0200; Telefax 301-572-0999).

Recommended citation:

Macro International Inc. 1996. *Sampling Manual.* DHS-III Basic Documentation No. 6. Calverton, Maryland.

CONTENTS

INTRODUCTION

The Demographic and Health Surveys (DHS) program has conducted more than 70 nationally representative household surveys in more than 50 countries since 1984. The DHS surveys are designed and implemented as single-round operations, and are relatively standardized. Standardization is achieved through the use of model questionnaires, manuals, field procedures, and through the technical support provided by DHS staff. There are two main survey instruments: a household schedule and an individual questionnaire. The household schedule provides a list of household members and basic demographic information about each member. This information is used for the individual questionnaire for selecting respondents who are women of reproductive age. These women are asked to report their reproductive history, knowledge and use of contraception, fertility preferences, and the health status of their young children. A male survey is often implemented in a subsample of the sample households.

Standardization, however, does not apply rigorously to sample design. As long as scientific probability sampling is used and national representativity is ensured, practical and cost problems can often prescribe the final design, especially with respect to the choice of a sampling frame and of area sampling units. However, DHS sampling does have a set of principles from which departure may occur. In this manual, the DHS sampling principles are described, a model sample design is presented together with its variants, and selected issues related to sampling design and implementation are discussed. This is not a general manual on sampling theory and practice. It is only intended as a practical guide to the sampling practices of the DHS project. The reader of this manual is encouraged to refer to the literature on survey sampling for more detailed discussions of sampling theory.

In addition to the general revision of the original text, the sections on sampling errors, and on sample description and documentation have been expanded. Examples of a training manual for mapping and household listing, a full sample description, and computer programs for sampling implementation are included.

The original DHS *Sampling Manual* (IRD, 1987) was prepared by Christopher Scott with the assistance of Graham Kalton and Alfredo Aliaga.[1] This edition of the manual was prepared by Thanh Lê who wishes to acknowledge Alfredo Aliaga, Anne Cross, Christopher Scott, and Ann Way for their valuable comments.

[1]Christopher Scott was also one of the authors of the World Fertility Survey (WFS) *Manual on Sample Design* (WFS, 1975). There may be some overlap between the WFS manual and this manual.

1. DEMOGRAPHIC AND HEALTH SURVEYS SAMPLING POLICY

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.

National coverage

A DHS sample should normally cover 100 percent of the population in the surveyed country; exceptionally, certain geographic areas may be excluded from the sampling frame due to extreme inaccessibility or dispersed population.

Probability sampling

Scientific probability sampling must be used. A probability sample is defined as one in which the units are selected with known and nonzero 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. Nonprobability methods represent a false economy. Although they may yield reasonable estimates in many instances, they cannot provide the confidence that is necessary in the event of unexpected findings. If this occurs, the use of nonprobability methods may lead to controversy and ultimately to criticism of the survey design.

Self-weighting sample

A self-weighting sample is one in which each elementary unit has an equal, nonzero overall probability of selection. DHS samples should be self-weighting unless there is good justification 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. When there are a number of survey objectives, computation of optimal sampling weights would lead to a different set of weights depending on the objective considered, yet only one set of weights can be used. Equal weighting, or self-weighting, is likely to be a good compromising choice. On the other hand, there are circumstances in which these considerations are counterbalanced by advantages in adopting unequal sample weights, particularly between reporting domains (section 2.1) in which case the sample should remain self-weighting within domains.

Preexisting sampling frame

If an adequate preexisting master 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.

Simplicity of design

The sample design should be as simple and straightforward as possible to facilitate accurate implementation.

In the sections which follow, DHS policy is described in relation to a number of specific aspects of sample design.

1.2 Survey Domain

Geographical coverage of each survey should include 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 of reproductive age (15-49). However, in some countries, the coverage may be restricted to ever-married women.

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 and Allocation

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 the capability of the implementing organization and the constraints of funding.

The DHS program is designed for samples of 5,000 to 6,000 women age 15-49.[1] 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, this size provides acceptable levels of sampling error for such key parameters as fertility and infant and child mortality.

Various factors may affect standard DHS sample size. Firstly, most of the planned analyses relate to, or are determined by, women currently in union (formal or informal) in the sample. Further, an important subset of questions relates to children born in the last three or 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 age 15-49 may be sufficient. Secondly, the standard DHS sample size allows up to five or six geographical regions to be distinguished in the tabulation of key variables. Typically, a sample of 1,000 women age 15-49 is needed for each geographical domain for which separate findings are required. However, in some countries there

[1] Sample size in this section refers to the *target* sample. Factors such as undercoverage at the household mapping and listing stage or nonresponse at the fieldwork stage may reduce the target sample by as much as 10 percent. The size of the selected sample should be increased to allow for expected undercoverage and nonresponse in order to maintain the desired target sample.

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 fieldwork 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 is unusual for a country to obtain external funding to carry out a survey utilizing a sample size larger than 10,000.

1.4 Area Sampling Frame

The availability of a suitable sampling frame is a major determinant of the feasibility of conducting a DHS survey. This issue should be addressed in the earliest planning for a survey. Whenever possible, DHS should obtain an area sampling frame from a single source. This is either an existing sampling frame, an existing master sample, the sample of a previously executed survey, or the list of enumeration areas (EAs) from a recently completed census.

The list of areal units should be thoroughly evaluated before it is used. It should cover the whole country, without overlap, and be as up-to-date as possible. Maps should exist for each areal unit or at least groups of units with clearly shown boundaries. Each areal unit should have a unique identification code or a series of codes that, when combined, can serve as a unique identification code. Each unit should have at least one measurement of size estimate (population and/or number of households). Other characteristics of the areal units (e.g., socioeconomic level), if they exist, should be evaluated and retained since they could be used for stratification.

A preexisting master 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. A sample "take" at the final stage or cluster "take" is the number of elementary units (i.e., households) selected in the final sampling stage in each cluster. 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 (section 2.6). A key question with a preexisting sample is whether the listing of dwelling/households is still current or whether it needs to be updated. If updating is required, use of a preexisting sample may not be worthwhile. The potential advantages of using a preexisting sample 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, resulting in respondent fatigue or contamination. DHS encourages the use of an existing sample, provided that it meets technical standards (section 2.6).

1.5 Overview of the Basic Sample Design

The basic procedures involved in the selection of an areal sample are straightforward. Most countries possess convenient area sampling frames, generally in the form of the EAs defined during the most recent population census. These generally come with sketch maps and size estimates, and, in principle, the EAs do not vary greatly in population size. However, in most countries, there are no satisfactory lists of

dwellings, households, or individuals within these EAs, and, in particular, 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 (section 1.4).

Census EAs are sometimes too large (1,000 to 2,000 population) to be economically feasible for a single survey to undertake the listing of all households in the survey's sample. Such EAs, therefore, need to be segmented into smaller areas for a further stage of area sampling before household listing begins. In some cases, the census maps are not accurate enough for the work of segmentation to be done in the office. A field operation may be needed to map and segment these oversized EAs.

A convenient and practical sample design has been developed by DHS based on experience with past surveys. First, a standard segment size is adopted, typically 500 population according to the sampling frame. Every areal unit in the country is then assigned a measure of size equal to the number of standard segments it contains, by dividing the population of the areal unit by 500 and rounding to the nearest whole number. A sample of areal units is then selected with probability proportional to this measure of size. In the selected areal unit with measure of size greater than one, 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 is selected by systematic sampling. In each selected household, a household questionnaire is completed to identify women age 15-49, all of whom are eligible to be interviewed.

Such self-weighting segment samples will usually be employed for DHS surveys. There are a number of variations on this design which are described in section 2.5.

1.6 Stratification

Stratification is the process by which the survey population is divided into subgroups or *strata* that are as homogeneous as possible based on certain criteria. *Explicit* stratification is the actual sorting and separating of the units into the specified strata; within each stratum, the sample is selected independently. Systematic sampling of units from an ordered list can also achieve the effect of stratification. This is called *implicit* stratification.

Strata should not be confused with survey domains. A *survey domain* is a population subgroup for which separate survey estimates are desired (e.g., urban areas, rural areas). A *stratum* is a subgroup in which the sample may be designed differently and is selected separately (e.g., large size areal units, medium size areal units, small size areal units). Survey domains and strata could be the same but they need not be. A survey domain could consist of one or several strata. If only implicit stratification is used, then no explicit strata exist.

Where data are available, explicit stratification should be used and can be based on socioeconomic zones, or more direct relevant characteristics such as level of female literacy, or presence of health and family planning facilities in the areas. Within each explicit stratum, the units can then be ordered according to location, thus providing implicit geographic stratification.

The principal objective of stratification is to reduce sampling error. In a stratified sample, the sampling error depends on the population variance existing *within* the strata but not *between* the strata. For this reason, it pays to create strata with low internal variability. Another major reason for stratification is that, where marked differences exist between subgroups of the population (e.g., urban vs. rural), stratification allows flexible selection of sample allocation and design separately for each subgroup.

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 households/women to be selected per cluster, or the cluster "take," depends on the variable under consideration. For example, in estimating contraceptive prevalence and its determinants, the variables of primary interest tend to be highly clustered, and comparisons are often 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, and 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.[2]

A larger cluster "take" reduces survey field costs. 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 30-40 women for the rural sector. In urban clusters, the cost advantage of a large "take" is generally smaller, and DHS recommends a "take" of about 20-25 women. Where a preexisting recent household list is available, these figures could be further reduced since the main factor favoring a large "take" is saving in listing operations.

1.8 Rationale for Listing

Listing of dwellings and 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 nonzero 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

[2]Data from the World Fertility Survey Assessment project.

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 less than ideal 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 Segmentation, Mapping, Listing, and Main Fieldwork

After selection of the areal units, the next step is segmentation. In most cases, segmentation can only be carried out in the field. Each areal unit, 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 unit. If size measures are required based on a quick count, these can be obtained at the same time (section 2.5).

Selection of the sample segment in each areal unit is the next step. It is important to prevent biased selection. Clear instructions on how to select the segment should be given to the team doing the segmentation in the field, together with necessary parameters (e.g., random number). Control procedures should be introduced to ensure that no conscious biased selection occurs.

The next step is mapping and listing. Mapping refers to drawing a sketch map of the selected areal unit (or segment of an areal unit) that shows, to the extent possible, the location of the dwellings together with landmarks found in the areal unit. The listing should be on a dwelling-cum-household basis (i.e., listing of inhabited dwellings together with all households residing in each dwelling) including dwellings where households are absent at the time of the visit by the listing team. The subsequent interview should cover the current occupants of the listed dwelling whether or not they occupied it at the time of listing. 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.

Once the mapping and household listing operation is completed, the household lists should be sent to the central survey office for the selection of households. Centralization of household selection is necessary so that the completeness of the listing operation can be assessed by experienced survey staff. Discrepancies between the *expected* and the *listed* number of households must be evaluated. Problem areas should be revisited. Sampling fractions could also be readjusted so as to give the expected number of households. In cases where it is not feasible to centralize household selection, especially when regional listers are employed and travel is difficult, supervisors could be trained to do the selection in the field. However, in this situation, evaluation of the results may not be possible.

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 eligible women for the individual interview. Eligible women are previously defined as those who are in the specified age group (15-49) and who slept in the household the night before the interviewer's visit. However, conscious omission of eligible women on the

part of interviewer—by pushing them out of the age limit or by stating that they did not spend "last night" in the household—is a real problem. Measures to eliminate this problem should be undertaken. For example, to remove the incentive for misrecording residency status, DHS has changed the procedure for selecting eligible women: interviews are to be conducted for all women age 15-49 regardless of whether they slept in the household "last night." However, the de facto character of the survey should be maintained at the data analysis stage, by including in the analysis only the women who slept in the household "last night" (section 3.1).

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 variables and subclasses of interest must be produced together with the survey results. This is crucial in evaluating and interpreting the survey results (see section 3.2).

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 several times during the survey operation if this seemingly unimportant chore is to be carried out effectively: (1) at the time of the sample design, (2) at the end of the fieldwork, 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 in section 4. The same section includes the requirements for sample information to be entered on the individual data records.

2. SELECTED SAMPLING TECHNIQUES

2.1 Disproportionate Sampling Between Domains

In a standard DHS survey, each final elementary unit (i.e., 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 larger 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 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.

One strategy for disproportionate sampling involves varying the sampling fractions between domains in order to maximize cost efficiency. It relies on the formula

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

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. 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 often 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 formulae are needed. In what follows, the population standard deviation is assumed to be constant among the domains. Also the assumption is made that any over- or undersampling is implemented by modifying the first stage selection only, leaving the design for subsequent stages unchanged. Let n_h denote the sample size and w_h 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 sampling 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, one needs 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 estimates based on the whole sample. The sampling variance for estimates relating to the whole sample is inflated by a factor:

$$L = \frac{(\Sigma n_h)(\Sigma n_h w_h^2)}{(\Sigma n_h w_h)^2}$$

This factor is never less than 1 so that even though precision is gained in a given domain estimate, there is a loss in the form of increased sampling error at the level of estimates for the whole sample.

Example 2.1.1

A fictional DHS country has an urban population of 1 million and a rural population of 4 million. A sample of 5,000 people (i.e., 0.1 percent of the population) shall be selected. In order to insure more precision in the estimates for the relatively small urban sector, a decision is made to double the urban sampling fraction relative to the rural one, while maintaining the same total sample.

Table 2.1.1 shows the different sampling parameters, the gain in the urban domain variance, and the loss in total variance.

Table 2.1.1

Domain	Population N_h	Proportionate Sample n_h	Disproportionate Sample n'_h	Var'/Var	SE'/SE
	(1)	(2)	(3)	(4)	(5)
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	1.04

Column (1) gives the population estimates.
Column (2) allocates the given total sample of 5,000 between urban and rural domains in proportion to the population in column (1).
Column (3): The sample size is first doubled for the urban domain (yielding 2,000) and the rural domain remains unchanged (4,000). Since this implies a total sample of 6,000, a factor k is introduced. The left-hand column sums to $6{,}000k$. Equating this to the desired sample of 5,000, $k = 5/6$. Hence $2{,}000k = 1{,}667$ and $4{,}000k = 3{,}333$, which are the new disproportionate sample sizes n', for the urban and rural domains, respectively.
Column (4) is simply (n_h/n'_h) for each domain since the sampling variance is inversely proportional to the sample size. The value in column (4) for the total domain is the quantity L discussed above: substituting n'_h for n_h and (n_h/n'_h) for w_h, the formula gives $L = 27/25 = 1.08$.
Column (5) is the square root of column (4).

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

In the above example, the disproportionate sampling plan chosen reduces the sampling standard error in the urban domain to 77 percent of its value under proportionate sampling. This gain is paid for with an increase in the sampling standard error of 10 percent for the rural domain and 4 percent for the total sample.

The equivalent proportionate sample to yield this larger standard error (4 percent higher) for the total sample can be calculated as $n/L = 5{,}000/1.08 = 4{,}630$, a "loss" of 370 cases. On the other hand, the sample size that would be necessary to maintain the same sampling precision in the "total" estimate as with the proportionate sample assuming the disproportionate sampling rates in the above table can be calculated as $nL = 5000 \times 1.08 = 5400$. In other words, an 8 percent increase in sample size would be needed to compensate for the disproportionate sampling.

Example 2.1.2

In a fictional DHS country, there are three important regions, North, Central and South, which differ considerably in population size. What would the implications be if it were decided to select a disproportionate sample designed to yield identical sample sizes in each region? In Table 2.1.2, the total population is assumed to be 5,400,000, and the desired total sample is 5,400.

Table 2.1.2

Region	Population N_h	Proportionate Sample n_h	Disproportionate Sample n'_h	Var'/Var	SE'/SE
	(1)	(2)	(3)	(4)	(5)
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	1.052

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

The equivalent proportionate sample, giving the same sampling error (5.2 percent higher) for the "total" estimate, is $n/L = 5,400/1.107 = 4,878$, so that the number of cases "lost" is 522. The sample size with the same disproportionate sampling rates which would be needed to maintain the same precision for the "total" 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.

Both of the above 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 2:1 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 magnitude 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

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.

2.2.1 Definitions and Calculations of Sample Weights

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

Instead of estimating a total, one may wish to estimate a mean, a rate, a proportion or a ratio. In all 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 one can 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 proportions; these require no raising factors and no weights as long as an equal probability sample design is used. However, if the selection probability varies, the raising factor $N/n = 1/f$, more generally called *weight*, has to be applied separately to each domain for which the sampling fraction is different.

Example 2.2.1

In Table 2.2.1, 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 sample 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

			Sample Data (unweighted)			Population Estimates		
Domain h	Sampling Fraction f_h	Weight $w=1/f_h$	Total Women	Married Women	Percent Married	Total Women	Married Women	Percent Married
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
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), one can 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 the total in column (7) by the total in 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, any common factor can be removed from the weights. For example, all the weights can be divided 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 means, rates, proportions, or ratios are being estimated.

A weighted rate or ratio $R=Y/X$ is estimated with the formula[1]

$$r = \frac{\sum w_h y_h}{\sum w_h x_h}$$

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

$$\bar{y} = \frac{\sum w_h y_h}{\sum w_h n_h}$$

The standard way to introduce weights into the analysis is to include a weight variable on each individual record. In Example 2.2.1, for instance, the data records for each sampled woman in the Eastern

[1]Capital letters indicate the population; small letters indicate the sample.

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 value of 2.5. When a weight variable is included in this way, it is straightforward to employ it to run any analysis required, provided that the computer program is capable of handling weights.

In Example 2.2.1, it is 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 Example 2.2.1, suppose that column (3) represents only the respondents while the numbers initially selected for interview were larger, as shown below:

Table 2.2.2

Domain	Number of Respondents	Number of Women Selected	Response Rate (Percent)
East	900	1,000	90
West	1,000	1,250	80
Coast	1,200	1,412	85

The response rate—number of respondents as a percentage of the number of women selected—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 1,000/0.90 = 1,111 for the East, from 2,000 to 2,000/0.80 = 2,500 for the West, and from 2,500/0.85 = 2,941 for the Coast. Making the weight for the East 1, the weights for the West 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 *within* 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 (for example, from a recent census), 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 is 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.

2.2.2 When is Weighting Necessary?

The overall effect of weighting is small. The sampling fractions in Example 2.2.1 vary by as much as 2.5 to 1, and are closely correlated with the variable of concern which is the percentage of married women. 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 called *design weights*. Obviously, in an equal probability sample no design weights are required, which is why such a sample is termed *self-weighting*.

With respect to nonresponse, 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 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 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.

2.2.3 Standardization of Weights

Caution should be taken in the use of weights in survey analyses. In particular, 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 formulae and programs (section 3.2).

A recommended procedure, which has several advantages, is to *standardize* 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:

$$T = \frac{\sum n_i}{\sum w_i n_i}$$

where n_i is the number of cases bearing the weights 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.

Standardization of the weights in this way has the advantage that if only the weighted n is quoted, the reader will not be seriously misled. In addition, the error 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 (section 3.2).

2.3 Systematic Sampling

Systematic sampling is the selection of sampling units at a fixed interval from a list, starting from a randomly determined point. Compared with random selection, systematic sampling has three advantages: (1) it is easier to perform; (2) it allows easy verification of the selection; and (3) 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, the procedure begins with a random number R that is less than or equal to I. The units to be selected are the ones numbered R, $R+I$, $R+2I$, etc., until the end of the list is reached. If the design specifies the number of units 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 is the number of units to be selected. On the other hand, if the design specifies the sampling fraction or probability f, then the interval is computed 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 in any case.

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

(1) Calculate the interval I rounded to one decimal place.

(2) Find a random number between 1 and $10 \times I$ and place a decimal point before its last digit. This becomes R.

(3) Compute the sequence of sampling numbers: R, $R+I$, $R+2I$, etc.

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

Example 2.3.1

Let $I = 3.4$. Select a random number between 1 and 34, say 23. Then $R = 2.3$. The sampling numbers and selections are as follows:

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

In this example, the method of decimal interval gives an interval which is sometimes 3, sometimes 4, with the desired average of 3.4.

After selection, one must check that the number of units selected is equal to N/I, with an error of not more than 1 (± 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 PPS Self-weighting Sample Design

2.4.1 PPS Sample Design

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

A common sampling plan is to select each PSU with a probability proportional to the estimated population of the PSU. 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 the 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 inversely proportional to the measure of size that was used at the area sampling stage. This means that the sampling fraction for households in area unit A will be 10 times smaller than in unit B, thus canceling 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 in section 2.2.

With this sample design, if the estimates of size used at the first sampling stage were always exactly equal to the number of households in each unit, or even exactly proportional to that number, it would follow that at the second stage, one would be selecting *a fixed number of households* in every selected unit.[2] 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 sampling, together with the self-weighting property itself. The fieldwork is a good deal easier to organize if there are not very large workloads in some areas and very small ones in others. This advantage is particularly significant when sampling ordinary administrative units, which commonly vary widely in population size.

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

---

[2] In a unit 10 times as large one would be selecting a fraction 10 times as small. 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 $m_i = k = constant$.

2.4.2 Method of Selection with PPS

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

a. List the area units with the estimated size M_i shown against each unit.
b. Cumulate the values of M_i and enter this against each unit. Check that the last entry equals the sum of the M_i values, or ΣM_i.
c. If a is the number of area units to be selected, compute the sampling interval $I = \Sigma M_i / a$.
d. Select a random number between 1 and I. Let this be R.
e. Compute the sampling numbers R, $R+I$, $R+2I$, ..., $R+(a-1)I$.
f. For each sampling number, find the first cumulative M_i which equals or exceeds it. The corresponding unit is the one to be selected.

Example 2.4.1

Planned sample size:	$a = 70$
Sum of unit sizes:	$\Sigma M_i = 38{,}500$
Interval:	$I = \Sigma M_i / a = 550$
Random number between 1 and 550:	$R = 90$
Unit selected into the sample:	001, 007, 011, etc.

Area Unit	Size M_i	Cumulative M_i	Sampling Number	Selection
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		
.	.	.		
.	.	.		
.	.	.		
Last 350	90	38,500		
Total	38,500=ΣM_i			

2.4.3 Computation of Sampling Fraction for Households

A two-stage self-weighting sample is assumed with area units at the first stage and households at the second stage; in each selected household all eligible women are to be interviewed. Let a be the number

of area units or primary sampling units (PSUs) to be selected, let M_i be the estimated size and M_i' be the actual size of the i^{th} area unit, measured in terms of the number of households.[3]

The PSUs are selected with systematic PPS sampling using the interval $I_1 = \Sigma M_i / a$. With the sampling points distributed at this interval, it is clear that a unit whose size is M_i has a chance M_i / I_1 of being selected. Thus, the first-stage probability is

$$P_{1i} = \frac{M_i}{I_1} = \frac{a M_i}{\Sigma M_i}$$

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.[4] If P_{2i} is the sampling probability at the second stage (i.e., for household selection) in the i^{th} PSU, then, to get a self-weighting sample with overall probability f so that $P_{1i} P_{2i} = f$, it follows that:

$$P_{2i} = \frac{f}{P_{1i}} = \frac{f \Sigma M_i}{a M_i}$$

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

$$I_{2i} = \frac{1}{P_{2i}} = \frac{a M_i}{f \Sigma M_i}$$

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. Suppose that the units selected in Example 2.4.1 are for the urban sector. Together, these two examples show the complete sampling process for the first three area units selected in the urban sector of a fictional country. For the sake of simplicity, the country is assumed to have a population of less than 1 million.

1.	Desired sample size:	5,000 women age 15-49
2.	Number of women age 15-49 in the country (extrapolated from census to survey date):	168,000

[3]See *Important Note*, page 19.

[4]In estimating the number to be selected, an allowance is made of approximately 10 percent for interview nonresponse and undercoverage in listing (or higher if data from previous surveys, when they exist, show a lower response and coverage rate.) In estimating the number existing, generally data from the latest census are projected to the fieldwork 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 all women in each household are going to be interviewed.

3.	Number of women age 15-49 in the urban sector:	46,600
4.	Sample selected (allowing 10 percent nonresponse):	5,556 women
5.	Desired cluster "take" (urban):	20 women interviewed
6.	Sampling fraction:	$f = 5{,}556/168{,}000 = 1/30.24$
7.	Women selected in urban sector:	46,600/30.24 = 1541
8.	Number of clusters selected (Average of 22 women per cluster):	1,541/22 = 70

For the second stage of sampling households, compute the household sampling interval for each selected area unit:

$$I_{2i} = \frac{1}{P_{2i}} = \frac{aM_i}{f\,\Sigma M_i}$$

where $a = 70$, $f = 1/30.24$, $\Sigma M_i = 38{,}500$, and M_i is the estimated size, taken from Example 2.4.1. This gives $I_{2i} = 0.0550M_i$. For the three units selected in Example 2.4.1, this gives:

Unit 001: $I_{2i} = 0.0550 \times 150 = 8.3$
Unit 007: $I_{2i} = 0.0550 \times 110 = 6.1$
Unit 011: $I_{2i} = 0.0550 \times 140 = 7.7$

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.

Important note: At the start of section 2.4.3, it is stated that the size measures M_i used in the first stage selection are to be understood in terms of the number of *households* existing in the i^{th} PSU. However, they could equally well be the number of *persons* (i.e., the census population). Even though the second stage of sampling operates with households, and the M_i' denote households, nowhere has the assumption been used that the M_i are measured in terms of households. The above procedures apply without any modification 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.

2.4.4 Units Selected with Certainty

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. These large units are said to have been *selected with certainty*, or to be *self-representing units*.

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, the simplest solution might be to split each such unit into two or more approximately equal subunits of size less than I.

The split would be made first 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.

If a substantial number of the units chosen to serve as PSUs are larger than the interval I, then the choice of such a unit to serve as PSU was clearly inappropriate. One solution to this problem is to remove all PSUs that are greater than a certain threshold size (which need not be exactly equal to I) from the list before sampling and to give them special treatment. Assuming the desire 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. For the purpose of sampling error computation, it is important to realize that the term *self-representing PSU* 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[5]

Exact PPS self-weighting sampling, leading to an exactly fixed "take" in each cluster, is not feasible in the context of DHS surveys. In general, the available measure of "size" for selecting the area units will 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 figures may have been inaccurate. In any case, they will usually be out of date by the time of the DHS survey.
- The census areas may not always be correctly identified during the mapping and household operation for the survey.
- Sizes are given in terms of population but the second stage of selection operates with households. Moreover, 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 a constant "take" of women per cluster when using the PPES self-weighting design in the DHS context. WFS and DHS experience suggest that the "take" will typically vary with a coefficient of variation of about 0.4.[6] This is a very substantial degree of variability.

Since the available measures of size provide only an approximation to the true size, 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 1.5 and described in the next section.

---

[5]Throughout this section, it is assumed that the survey design calls for the interview of every eligible woman in each sampled household.

[6]The coefficient of variation (CV) is defined as the ratio of the 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 (from 35 ± 1.64×14).

2.5.1 The Standard Segment Design

In the discussion that follows, it is assumed that census enumeration areas (EAs) are used as first stage units. The design is the same when other areal units than EAs are used. The first step in designing a sample using the standard segment approach is to fix a standard segment size, as small as seems practical, e.g., 500 population. Segments that are too small will be difficult to map due to lack of readily identifiable boundaries. Each EA in the country is allocated a number s_i of segments by dividing its census population by the standard segment size and rounding to the nearest whole number. At this stage no segments are actually identified. If an EA is so small that s_i would be 0, it is combined with the next one on the list with the exception that, if such a small EA is the last listed in a stratum, it should be combined with the preceding one.

The EAs are then sampled with PPS where the "size" is the number of segments, s_i. A field operation is then organized to create the exact number s_i of segments required in each selected EA, and one of these segments is selected at random. Creation of segments involves delineating them on a map of the EA. Within each EA the segments should be approximately equal in population size. However, it is also important to adopt segment boundaries that are easily identifiable. After selection of the segments, a listing operation must be carried out in each selected segment.

The sampling probability at the first stage is

$$P_{1i} = ks_i$$

where k is a constant, and the second stage probability is

$$P_{2i} = \frac{1}{s_i}$$

The overall probability is therefore:

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

Since this is constant, the result is a self-weighting sample of segments. Finally, a third stage is introduced at which households are selected with a probability that is fixed everywhere.

An exactly equivalent procedure for EA selection, somewhat simpler in concept, is to list the EAs with their segments represented by X's. Then a systematic sample is selected among the X's, illustrated by the next example.

Example 2.5.1

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

Table 2.5.1

EA Number	Size (in households)	Segments	Selection
001	140	X	
002	260	X	←
		X	
		X	
		X	
004	40 }	X	←
005	100 }		
006	90	X	
007	390	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 EA in which a segment is selected is considered selected. In Example 2.5.1, $I = 5$, the random start is 2, and the EAs 002, 004/005 (combined), and 007 are selected.

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

The method of selecting EAs with probability proportional to their number of segments, followed by selection of one segment in each EA, is equivalent to a single stage sample of segments. The EAs 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 first stage would. This situation arises typically when only one secondary sampling unit (SSU) is selected in each PSU. In such cases, the first stage is labeled *notional* and there is only one *effective* areal sampling stage.

If PSUs are selected directly from a frame of EAs, and only a few of the largest EAs are subdivided before sampling in order to reduce their size variation, then this is essentially the same as in the standard segment design except that the segmentation is carried out exceptionally, rather than usually. Again, it is a single effective area stage, though the sampling units in this case would best be described as EAs or segments of EAs.

2.5.2 Variant 1: Two Distinct Area Stages

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

At the first stage, EAs are selected with PPS, using the census population data as they were reported in the sampling frame of census EAs to provide the measure of size. In the selected EAs, a number of

segments, s_i, are created and then one segment is selected at random. This is the second stage. At the third stage, households are selected with a probability calculated so as to yield a self-weighting sample.

If M_i is the census size of the i^{th} EA, and if the plan is to select a EAs, then the first stage probability is

$$P_{1i} = \frac{aM_i}{M}$$

where $M = \Sigma M_i$, summed over all EAs existing.

At the second stage, one segment is selected from among s_i, so that

$$P_{2i} = \frac{1}{s_i}$$

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, it follows that

$$P_{1i} P_{2i} P_{3i} = f$$

This gives

$$P_{3i} = \frac{f}{P_{1i} P_{2i}} = \frac{fMs_i}{aM_i}$$

Thus, households are selected in the selected segment of the i^{th} EA by systematic selection with interval

$$I_i = \frac{1}{P_{3i}} = \frac{aM_i}{fMs_i}$$

Note that with this design, one is entirely free to choose the number of segments to be made, s_i, in each EA. However, it is still desirable to have the segments of any given EA as equal in size as possible. Freedom to choose s_i in each EA 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. However, 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. As a result, 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.

Although the computation of s_i for every EA in the country is avoided, one has to cumulate the M_i over all EAs, and there are more computations to be done at the level of the sample EAs. With personal computers available in almost every survey office, a computer program can be easily written to handle the volume of calculations. The choice of software is left to the sampler, but one such as dBase is quite appropriate, especially if the sampling frame can be easily converted into dBase format. dBase allows easy manipulation of a large number of records (e.g., EAs), and hence can be used for a large range of tasks from stratification, to cumulation of size, to sample selection, and even to calculation of sampling probabilities. Examples of dBase programs for different tasks can be found in the appendix.

2.5.3 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 interviewed per ultimate area unit) is decided in advance. It is suggested in section 1.7 that this might reasonably be fixed at an average of about 20 women in the urban sector and 40 in the rural sector. If there are, for example, 1.25 women per household, the "take" would amount to 16 households and 32 households for the urban and rural sector, respectively. Let this target "take" be T.

If segments could be made of average size T, it is arguable that the listing operation could be avoided altogether by using "take-all" or "compact cluster" sampling at the last stage. The objection to this approach 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 minimize the cost of listing.[7] 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 variance due to the correlation between neighbors in the sample. The balance of advantage here is likely to be positive.

As with the standard segment design already described, this design begins by imagining every PSU (normally a census EA) divided into s_i standard segments, but the standard segment is much smaller and equal to T. Note that T, originally worked out in terms of women age 15-49, needs to be converted to a population base (i.e. the number of persons of all ages and both sexes) and then adjusted back to the census date. To convert to population, the latest census can be used to give the ratio *total population/women age 15-49*. To backdate to the census, one can assume that the population growth rate applies also to the growth rate for women age 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.

The first stage of sampling consists of the 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.

---

[7] In section 1.8, it is suggested that listing is 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.

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. This unit is called a *division*. The selected PSU must be mapped into not more than s_i well-defined divisions. Clarity of boundaries is 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 $j = 1, 2, 3 \ldots$ 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 $s_i = \Sigma s_{ij}$; that is, the s_i standard segments in the PSU are allocated among the divisions in proportion to the size of each division, and with at least one in each division.

For example, in one PSU, $s_i = 8$. Suppose that it is found feasible to divide it into only four divisions. After these have been delineated, their sizes are examined and the eight imaginary segments are allocated among the four 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 $s_{ij} = 1$, every household is taken; if $s_{ij} = 2$, every second household is selected systematically; if $s_{ij} = 3$, every third; and so on. The sampling probabilities are as follows:

First stage: Selection of PSUs

$$P_{1i} = \frac{a M_i}{M} = \frac{a s_i}{\Sigma s_i}$$

where a is the number of PSUs selected, M_i is the census size of the i^{th} PSU, and M is the total census size. Note that $s_i = M_i / T'$, rounded to the nearest whole number.

Second stage: Selection of divisions

$$P_{2i} = \frac{s_{ij}}{s_i}$$

Third stage: Selection of households

$$P_{3i} = \frac{1}{s_{ij}}$$

The overall probability is:

$$f = P_{1i} P_{2i} P_{3i} = \frac{a}{\Sigma s_i} = \frac{a T'}{M}$$

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 run of consecutive households from the list instead of using systematic selection. For example, if $s_{ij} = 3$, the list is divided into three equal parts and one is selected 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 section 2.5.4).

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. The difference between a quick count and a full listing, in terms of person-hours of work, is often not very clear. If the latter were used, one 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, all of which are unavoidable.

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.

2.5.4 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, say 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 to the nearest whole number) and select two from the list. This not only reduces the variance arising from intraclass correlation but also reduces the rounding error in s_i, which may be substantial where f_i is large. Even when $f_i = 0.5$, rounding can result in an unacceptably large 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.

2.5.5 A Common Error in PPS Standard Segment Sampling

In the PPS standard segment 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, one segment is selected out of the s_i created, giving a second stage probability $P_{2i} = 1/s_i$. When these two probabilities 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 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 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 are described in section 1.4. In this section, certain aspects of such subsampling are discussed in detail.

2.6.1 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 not a self-weighting sample. In this section, suppose that the master sample is divided into domains $h = 1, 2, 3, \ldots$ for which the overall sampling fractions are f_h. Probabilities or sampling fractions are denoted 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, $f' = f \times f^*$.

Applying this to the domains, it follows that

$$f_h' = f_h f_h^*$$

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

$$f_h^* = \frac{f'}{f_h}$$

The overall sampling fraction f_h is the product of the two sampling probabilities P_{1h} for area selection and P_{2h} 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 the following can be written:

$$P_{1h}^* = \frac{f'}{f_h}$$

where P_{1h}^* is the subsampling rate at the area stage. However, it may sometimes happen that the master sample is designed to yield a cluster "take" which differs from that considered optimal for DHS in one or more domains. In this case the second 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:

$$\frac{P_{2h}'}{P_{2h}} = \frac{b_h'}{b_h}$$

Since the overall sampling fraction f is the product of P_1 and P_2, then:

$$\frac{P_{1h}'}{P_{1h}} = \frac{P_{2h}}{P_{2h}'} \times \frac{f_h'}{f_h} = \frac{b_h}{b_h'} \times \frac{f'}{f_h}$$

In the above formula, f_h' is replaced by f' on the assumption that the DHS sample is to be self-weighting. The ratio on the left is the subsampling rate P_{1h}' required at the area stage. As long as this ratio does not exceed 1 in any domain, the desired subsampling at the area stage can be achieved. If it exceeds 1, one might allow some increase in b_h', the cluster "take" in DHS, but within limits, since one would not wish to deviate excessively from the optimal cluster "take." Another solution is to consider augmenting the sample by a supplementary selection in the domains where the constraints cannot be met.

2.6.2 Updating the Listing

If the lists of dwellings or households provided by the master sample are more than a year old,[8] they will need updating. Since updating requires visiting every dwelling, and in view of the temptation for fieldworkers 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:

- There is a savings in mapping and segmentation. Where segmentation has been done for the master sample, this saving is considerable.
- Use of a common sample allows linkage between the surveys concerned, with enhanced analytic potential.

2.6.3 Disadvantages of Sample Overlap

Where the same households are interviewed in two or more surveys, there are the potential problems of respondent resistance, and if the two surveys cover the same subject matter, contamination or conditioning may occur.

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.

---

[8]The limit is generally lower in unplanned urban areas (six months).

Contamination is the feedback of influence from a previous interview 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 they have been subject to some 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, the mere potential of contamination reduces the value of the second response.

Some 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 an overlap. 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 have to be matched to the old. This will involve serious practical difficulties. 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?

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 (e.g., segments) within the same sample of penultimate area units (e.g., enumeration areas). Where the ultimate area units are segments, this has the advantage of saving work on a new segmentation operation. However, a problem will arise if certain EAs contain only one segment each, leaving no alternative selection for the second survey. A solution might be 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.

2.6.4 Advantages of Sample Overlap

When repeated surveys in the same country are conducted to measure changes over time, then overlapping the samples of the previously conducted survey and the new survey presents certain advantages. The main advantage is that the high correlation between the two surveys results in the reduction of the sampling error for changes.

Different ways of overlapping samples result in different levels of improvement in variance. The three simplest ways of overlapping the two surveys are: same respondents, same dwellings (not necessarily same respondents), and same clusters (not necessarily same dwellings or respondents). The greatest improvement in variance occurs when the same respondents are included in the two surveys. However, biases are also greatest at this level where loss to follow-up and change of household structure are inherent, resulting in the newer sample that is no longer representative of the study population. Some reduction in variance will also be achieved using the same clusters (with additional reduction using the same dwellings). Using the same clusters is also cost-effective since the mapping from the first survey could also be used again. Relisting of the clusters may be necessary if there is a 2-3 year gap between the two surveys.

3. SURVEY ERRORS

The estimates from a sample survey are affected by two types of errors: (1) nonsampling errors, and (2) sampling errors. Nonsampling errors are the results of mistakes made in implementing data collection and data processing, such as failure to locate and interview the correct household, misunderstanding of the questions on the part of the respondent or of the answers on the part of the interviewer, and data entry errors. Noncoverage and nonresponse are also classified as nonsampling errors, and they are the only two types that will be discussed in this section. Even if numerous efforts are made during the implementation of a survey to minimize this type of error, nonsampling errors are impossible to eliminate entirely and difficult to evaluate statistically.

Sampling errors, on the other hand, can be evaluated statistically. The sample of respondents selected in a DHS survey is only one of many samples that could have been selected from the same population, using the same design and expected size. Each of these samples would yield results that differ somewhat from the results of the actual sample selected. Sampling errors are a measure of the variability between all possible samples. Although the degree of variability is not known exactly, it can be estimated from the survey results.

3.1 Errors of Coverage and Nonresponse

Coverage error occurs when there is a 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 attempted, this is called an error of undercoverage; if it is attempted without success, it is an error of nonresponse.[1] This section deals with problems in the definition and estimation of such error rates.

3.1.1 Coverage Error

In DHS surveys, errors of overcoverage, i.e., inclusion of elements that do not belong in the sample, do not occur as often as undercoverage errors, errors due to exclusion of elements that would properly belong in the sample. In the first 63 standard DHS surveys implemented between 1986 and 1995, 25 surveys have coverage of less than 95 percent (of the target sample of women) and 19 surveys have overcoverage of 5 percent or more. Several sources of error may be identified in the problem of coverage. The first type of error arises at the sample implementation stage, notably in the listing stage when listing workers cover less or more than the designated area.[2] Second, where an age limit is used to determine eligibility for interview, distortions in age reporting are another source of undercoverage. Third, where surveys cover de facto women (i.e., those women who slept in the household the night before the survey), there may be deliberate omission of eligible women by the interviewers. Interviewers may consciously misreport a woman's "residency" status as non-de facto, which thereby disqualifies the woman from being eligible for an

[1]Sometimes an interview is deliberately not attempted because it is known that the selected respondent is unavailable or inaccessible. This is classified as nonresponse because failure to attempt the interview was not "erroneous," and the overriding factor was the impossibility of achieving it.

[2]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. These areas were not selected into the sample while figures of projected population from census data were used to calculate sampling fractions for households, resulting in a shortfall of sampled households.

interview. All three types of coverage errors may involve motivated bias by fieldworkers seeking to reduce their workload.

Motivated errors can be controlled by intensive training and close supervision.[3] Error due to an outdated area frame can be reduced by taking special steps to update the frame in areas of known new settlements, 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.

Changing the rule for interviewing women helps in reducing motivated misreporting of residency status. In DHS surveys, the interviewers are now instructed to interview all women age 15-49 regardless of whether they slept in the household the night before the survey. By requiring interviewing of all women, the incentive for misrecording residency status has been eliminated. However, the de facto character of the surveys is maintained at the data analysis stage. Using different fieldworkers to conduct the household schedules and to interview women respondents will also help in eliminating both age distortion and misreporting of residency status.

Coverage errors can be investigated after the survey fieldwork 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.[4] Age distortions can be investigated by studying the discontinuity in trends across the eligibility boundaries of 14-15 and 49-50 years. While it is tempting to introduce comparisons with males as a control, it should be noted that in most societies more males are educated than females, so that heaping at ages 15 and 50 may be less extreme for males.

3.1.2 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:

- Exclusion of clearly identified areas from the sampling frame—In this case, 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.

- Ad hoc exclusions decided during or just prior to fieldwork—In developing countries it is not uncommon for the survey organization to abandon the attempt to conduct fieldwork 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. 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.

---

[3]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.

[4]The growth rate for women age 15-49 can be reasonably assumed equal, or slightly above, that of the population. However, the growth rate for households typically will be substantially smaller.

3.1.3 Nonresponse

At first sight, the concept of nonresponse seems simple and clear: it is the percentage of the persons who should have been interviewed but were not interviewed. Taking into account the distinction between coverage error and nonresponse indicated earlier, this can be modified by saying that the information desired is 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, survey personnel list and select dwellings, interview the household currently in the dwelling, then interview any women age 15-49 in that household. If failure occurs at one of the earlier steps, the information which would enable us to classify the effect at the final level is lacking. For example, if the interviewer cannot find the selected dwelling, it is not known whether it contains a woman eligible for interview; if it does not, then the failure has no effect on the interview response rate.

To deal with this problem, an example is 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 primarily concerns women, the relevant response rate is *D/C*. But the quantity *C* is not known. It is of interest to know the number of eligible women in all selected households but only the number in the households interviewed, say *C'* is known. Therefore *C* is estimated as follows:

$$Estimated\ C = C' \times \frac{A}{B}$$

This assumes that the number of eligible women per household is the same among nonrespondent as well 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 the first problem. Where two or more steps of sifting are involved, the overall response rate can be estimated by multiplying together the response rates observed at each step. In doing so, the assumption is made 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 there is a need to compute a dwelling response rate as well as the household and women response rates. *Dwelling nonresponse* refers only to the categories *dwellings destroyed* and *dwellings not found* between the listing and the interviewing. If these are represented 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. It is assumed that dwellings in category P have no households while those in Q are typical of other dwellings: thus one drops the first term above but adds in the category Q (dwellings not found) to the household base A'. This gives:

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

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?

This question 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.[5] 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 existing at the time of interviewing in the dwellings selected from the dwelling list.

3.1.4 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 formulae for response rates in terms of such codes. In DHS surveys, the following response categories are used at the household levels:

1*H*	Completed
2*H*	No household member at home or no competent respondent at home
3*H*	Entire household absent for extended period
4*H*	Postponed
5*H*	Refused
6*H*	Dwelling vacant or address not a dwelling
7*H*	Dwelling destroyed
8*H*	Dwelling not found
9*H*	Other

[5] In practice, DHS usually lists not dwellings but structures, or buildings, and all households residing in the structure. The reason is that it is rather time-consuming to identify each and every dwelling; the lister would have to rely on a household member to define his dwelling, especially when there is more than one dwelling in the building.

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:

$$R_H = \frac{1H}{1H + 2H + 4H + 5H + 8H}$$

At the individual level the following response categories are used:

1*I*	Completed
2*I*	Not at home
3*I*	Postponed
4*I*	Refused
5*I*	Partly completed
6*I*	Incapacitated
7*I*	Other

The individual response rate is thus:

$$R_I = \frac{1I}{1I + 2I + 3I + 4I + 5I + 6I + 7I}$$

The category *no eligible woman in the household* is not included in the list since it is irrelevant to the response rate, appearing neither in the numerator nor the denominator. It is assumed that no woman's questionnaire will be provided for such cases.

Whenever the *other* code is used, the interviewers should specify the reason for nonresponse. At the household level, the analyst should review a printout of the other codes and recode as many as possible into the existing categories. Similarly, all other codes for the individual interview should be examined and recoded. Any questionnaire in which the household or the woman were deemed ineligible should be destroyed. An ineligible household may be one in a dwelling unit that does not lie within the sample area. An ineligible woman may be one who was reported 16 years old in the household questionnaire, but later turned out to be 14 (in which case her age in the household questionnaire should be corrected appropriately).

The overall response rate is obtained by multiplying the household and the 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 previous section on coverage error). In such cases, it is recommended that the analyst estimate the proportion p_o of the population so excluded, and multiply the overall response rate by the factor $1 - p_o$.

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

$$R = (1 - p_o) \times R_H \times R_I$$

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.

3.2 Sampling Errors

A sampling error is usually measured in terms of the *standard error* for a particular statistic (mean, percentage, etc.), which is the square root of the variance. The standard error can be used to calculate confidence intervals within which the true value for the population can reasonably be assumed to fall. For example, for any given statistic calculated from a sample survey, the value of that statistic will fall within a range of plus or minus two times the standard error of that statistic in 95 percent of all possible samples of identical size and design.

If the sample of respondents had been selected as a simple random sample, it would have been possible to use straightforward formulae for calculating sampling errors, such as $\sqrt{(pq/n)}$ and $S/\sqrt{n}$ for the estimated standard errors of a proportion and a mean, respectively. However, a DHS sample would most often be the result of a multistage stratified clustered design, and, consequently, it will be necessary to use more complex formulae.

The computer software used to calculate sampling errors for DHS is the Sampling Error Module of ISSA.[6] This module uses the Taylor linearization method of variance estimation for survey estimates that are ratio estimates (means or proportions). It uses the Jackknife repeated replication method or the balanced repeated replication for variance estimation of more complex statistics such as fertility and mortality rates.

The Taylor linearization method treats any percentage or average as a ratio estimate, $r = y/x$, where y represents the total sample value for variable Y, and x represents the total number of cases in the group or subgroup under consideration. The variance of r is computed using the following formula, with the standard error being the square root of the variance:

$$Var(r) = SE^2(r) = \frac{1-f}{x^2} \sum_{h=1}^{H} \left[\frac{m_h}{m_h - 1} \left(\sum_{i=1}^{m_h} z_{hi}^2 - \frac{z_h^2}{m_h} \right) \right]$$

in which

$$z_{hi} = y_{hi} - r x_{hi} \quad , \quad and \quad z_h = y_h - r x_h$$

where h represents the stratum which varies from 1 to H,
m_h is the total number of clusters selected in the h^{th} stratum,
y_{hi} is the sum of the values of variable y in cluster i in the h^{th} stratum,
x_{hi} is the sum of the number of cases in cluster i in the h^{th} stratum, and
f is the overall sampling fraction, which is usually so small that it can be ignored.

The Jackknife repeated replication method derives estimates of complex rates from each of several replications of the parent sample, and calculates standard errors for these estimates using simple formulae. Each replication considers *all but one* clusters in the calculation of the estimates. Pseudo-independent

[6]The Integrated System for Survey Analysis (ISSA) was developed specifically for the DHS program, and has been used for all aspects of data processing, from data entry, to editing, to tabulation. The sampling error module has been added recently to allow the calculations of sampling errors for complex demographic rates such as fertility and mortality rates using the Jackknife method. Before this module was introduced, the CLUSTERS program, developed for the WFS, was used to compute sampling errors for the DHS surveys. Only the Taylor linearization method is used in CLUSTERS.

replications are thus created. If there are k nonempty clusters, then k replications will be created. The variance of a rate r is calculated as follows:

$$Var(r) = SE^2(r) = \frac{1}{k(k-1)} \sum_{i=1}^{k} (r_i - r)^2$$

in which

$$r_i = kr - (k-1)r_{(i)}$$

is the value estimate of the i^{th} pseudo-independent replication and where

- r is the estimate computed from the full sample of k clusters,
- $r_{(i)}$ is the estimate computed from the reduced sample of k-1 clusters (i^{th} cluster excluded), and
- k is the total number of clusters.

In addition to the standard error SE of an estimate r, other parameters that should be of great interest are, for each estimate, the design effect (*DEFT*), the relative standard error (SE/r), and the 95 percent confidence intervals ($r \pm 2SE$). The *DEFT* is defined as the ratio between the standard error using the given sample design and the standard error that would result if a simple random sample had been used. A *DEFT* value of 1.0 indicates that the sample design is as efficient as a simple random sample, while a value greater than 1.0 indicates the increase in the sampling error due to the use of a more complex and less statistically efficient design.

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 as ISSA and CLUSTERS cannot be used.

A DHS survey produces an extremely large number of estimates, many of which are presented as percent 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, appear in the *Guidelines for the DHS Main Survey Report*.

4. SAMPLE DESCRIPTION AND DOCUMENTATION

4.1 Minimal Sample Description

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

1. Statement that the sample is a probability sample.
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. If the number of stages differs between strata, it should be mentioned. Count "effective" stages only. Thus, the standard segment design would be described as two-stage, the first stage being segments of enumeration areas, the second stage being dwellings/households.
4. If feasible, identify the nature of the units used as primary (first 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 sector of the national population 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, together with overall response rate (see section 3.1).
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.

Example 4.1.1

The sample for the Ghana DHS was a two-stage stratified self-weighting probability sample, representative of the entire country. A total of 4,405 households and 4,488 women in 144 census enumeration areas were successfully interviewed, with an overall response rate of 96 percent. The areal sample was a subsample of the 1988 Ghana Living Standards Survey.

Example 4.1.2

The sample for the Mali DHS was a stratified probability sample covering the national population, with the exception of the rural areas of Tombouctou and Gao (5 percent of the national total). Two sampling stages were used in the urban sector and three in the rural. In the urban sector, 60 census enumeration areas were selected at the first stage, and households at the second stage. In the rural areas, 34 districts were selected at the first stage; at the second stage, 2 villages were selected in each sampled district; and, at the third stage, households were selected. The sample was self-weighting within each sector but the urban selection probability was twice the rural. A total of 3,047 households and 3,200 women were successfully interviewed, the overall response rate being 98 percent.

4.2 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 fieldwork to take into account

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 fieldwork and data entry.

The description should include all the items just mentioned, but in full detail. The next section contains 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 format would be, first, a short summary description similar to that previously outlined, and second, a complete account of each sampling stage, working down the hierarchy of stages from PSUs to individual women.

4.2.1 Verbal Description

Insofar as sample design and parameters differ among domains (regions, urban, rural), the sample should be described and documented for each domain separately. The following list serves as a guideline for the points to be discussed in the description of the sample design:

1. Describe the population covered and the population excluded, including their size.

2. Describe the sampling frame used at each stage and any procedure used for frame updating. The frame may be described in three separate parts:
   a. Primary area sampling frame: this is the frame of PSUs which in principle covers the entire population.
   b. Secondary area sampling frames: there can be more than one, and these cover area units at various levels within selected PSUs, up to and including *ultimate area units* (UAUs). This part will not be necessary if the sample goes directly from PSUs to dwelling/household units.
   c. The list frame within sample UAUs.

   For each type of frame, the following items are to be specified:
   - The source of the frame, the physical form in which the frame is kept, and the identification system.
   - The type of units involved, and type of information available on the units, especially measures of size and characteristics used for stratification.
   - The date when the frame was created.
   - Any information on subsequent updating, whether it was done prior to, independently of DHS, or as part of DHS operations.
   - The updating procedure, for example, reclassification (urban/rural) of units, changes in boundaries, improved mapping, updating of information on characteristics especially measures of size.
   - If listing was done, provide the date of listing, describe whether or not it was part of DHS operations, describe the unit of listing (dwellings or households), and whether or not it involved the preparation of entirely new lists or merely updating of existing lists.
   - When an existing sample or master sample was used, provide the above information for the master sample up to and including the stage at which the operation moves from the master sample to the DHS sample. From and including that stage, provide the information for the DHS sample.

3. Describe the sample selection procedures.

a. Specify whether the domain formed an explicit stratum for independent sample selection, whether the domain was divided into more than one explicit stratum. Explain the basis of stratification and identify the strata.
b. At each sampling stage, describe the basic selection method used. Note if systematic selection is used, at least in the first sampling stage. Since systematic selection is equivalent to implicit stratification; describe the basis on which the first stage list or frame is ordered (e.g., geographical, alphabetical, by size of unit, etc.).[1] If selection with probability proportional to size is used, identify the "size" variable.

   Description of selection methods may be:
   i. Fixed probabilities at each stage.
   ii. First stage probability fixed within each region but varying between regions in a specified ratio (e.g., North:Center:South = 3:2:1). In the remaining stages, probabilities used are constant throughout.
   iii. First stage: probability proportional to size (e.g., the size being the census population). Second stage: probability inversely proportional to first stage probability.

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

4.2.2 Numerical Parameters

The two useful numerical parameters are the selection probabilities and the weights. These should be discussed in the full sample description as follows:

1. Give selection probabilities that are fixed, distinguishing between conditional and overall probabilities. Where selection probabilities 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. Specify any weights recommended for use.

---

[1]When systematic sampling is used at the first stage, the standard formulae 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 two 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 two selections. If the number of units selected is odd, the last of the implicit strata is made with three 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.

Thus distinction is made between explicit strata, defined in the normal way, before sampling, and implicit strata defined after systematic sampling by the procedure of pairing PSUs selected consecutively.

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 men/husbands are sampled (for additional men/husbands' interviews), descriptions of these sampling operations should be added in similar terms.

4.3 Sample Documentation in Database Format

Even a 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.

The following example of a database, computerized using any spreadsheet software, contains a complete list of sample UAUs also referred to as *clusters*, providing several items of information for each cluster. Such a database should be created as soon as the sample areas have been selected. Subsequent information can be completed as listing and fieldwork proceed. Where the sample is selected systematically, all clusters and other area units should be listed in the order of selection. In all cases, the list of clusters should be ordered hierarchically by sampling domain, and within that, by explicit stratum, by primary sampling unit (PSU), then by each subsequent stage unit if relevant. When there is only one area sampling stage, then the UAU or cluster is the same as the PSU.

The following columns should be included, except where not applicable for a particular design:

1. *Identification*
   A unique identification number of the cluster must be provided. Normally, this will consist of three parts: (i) an identification number defining the administrative structure (e.g., region, urban/rural sector, district, etc.) of the cluster; this includes, where relevant, the identification number used for the area in the population census or master sample frame from which the DHS sample has been drawn; (ii) an identification number providing full information on sample structure (e.g., identification of the domain, explicit stratum, PSU, other higher stage unit, etc.) of the cluster; (iii) a simpler processing number which may have been assigned to the cluster (e.g., a sequential number appearing on the questionnaires and data files), usually called the *cluster number*.

2. *PSU (Primary Sampling Unit)*
   For each domain, and within it, for each explicit stratum as used in DHS, the following must be provided:
   a. the total measure of size (e.g., population or households as recorded in the frame);
   b. the total number of PSUs before selection; and
   c. the number of PSUs selected.

   For each PSU selected provide the following:
   d. the measure of size;
   e. its selection probability;
      (If the DHS sample has been selected from a master sample or from an existing sample at the stage concerned, this selection probability will be shown in two parts:
      e' :selection probability of the PSU in the existing frame or master sample; and
      e'' :selection probability as applicable to the DHS after taking a subsample of the PSUs from the existing frame or master sample.)

f. the total number of *secondary sampling units* (SSUs) in this PSU (if the sample consists of more than one area stage); and
g. the number of SSUs selected.

3. *SSU (Secondary Sampling Unit)*

If the sample consists of more than one area stage, for each PSU, and for each unit at each stage preceding the UAU or cluster, if present, provide the following:

a. the measure of size of the unit;
b. its selection probability;
   (If the DHS sample has been selected from a master sample or from an existing sample at the stage concerned, this selection probability will be shown in two parts:
   b': selection probability of the unit in the existing frame or master sample; and
   b'': selection probability as applicable to the DHS after taking a subsample of the units from the existing frame or master sample.)
c. the total number of next (lower) stage units in this unit; and
d. the number of units selected for the next stage.

4. *UAU (Ultimate Area Unit)*

For each UAU, provide:

a. its measure of size;
b. its selection probability;
   (This is to be specified as b' and b'' above if the DHS UAUs have been selected from the UAUs in an existing frame or master sample;) and
c. the required sampling rate within the UAU as originally designed.

5. *Dwellings/households*

If the listing and ultimate sampling units are *dwellings* rather than households, provide for each UAU:

a. the number A of dwellings listed;
b. the number B of dwellings selected;
c. of the number of dwellings selected, the number C of dwellings found to exist at the time of fieldwork, i.e., excluding the demolished, vacant and nonresidential units;
d. of the number of dwellings found to exist, the number D of dwellings successfully contacted, i.e., excluding those not found, inaccessible, temporarily vacant, with no competent respondent, and whose residents refused;
e. of the number of dwellings successfully contacted, the number E of households found;
f. the number F of households selected if this number is different than E;
g. of the number of households selected, the number G of households successfully interviewed;
h. the subsampling rate H within UAU for the households:

$$H = \frac{B}{A} \times \frac{F}{E}$$

(Note that if dwellings are selected in the DHS, all households found in the dwellings are interviewed so that $F/E = 1$.)

i. the overall sampling rate for households in the UAU:

$$I = H \times (4b)$$

j. the response rate for the household interviewed:

$$J = \frac{D}{C} \times \frac{G}{F}$$

If the listing and sampling units are households, only a subset of the items in part 5 will be present:

e. the number E of households listed;
f. the number F of households selected;
g. the number G of households successfully interviewed;
h. the subsampling rate within UAU:

$$H = \frac{F}{E}$$

i. the overall sampling rate for households in the UAU; and
j. the response rate for households interviewed:

$$J = \frac{G}{F}$$

6. *Eligible women*

Provide for each UAU:

a. the number a of eligible women found in the households successfully interviewed in 5g;
b. the number b of women selected, if this number is different than a;
c. the number c of women successfully interviewed;
d. the response rate d for the individual interviews:

$$d = \frac{c}{b}$$

e. the overall response rate e for the individual interviews:

$$e = d \times J$$

f. the overall sampling rate f for women in the UAU:

$$f = \frac{b}{a} \times (5i)$$

7. *Weights*
   If weighting was done by UAU or by units at some higher stage within the domain, the sample weight may be specified as a combination of several factors, for example:
   a. the design weights, inversely proportional to *f* (in 6f);
   b. the adjustment for nonresponse, inversely proportional to *e* (in 6e);
   c. the adjustment on the basis of external population control totals; this may be introduced to correct the distribution of the sample among domains to agree with the population distribution. In DHS, this adjustment is not often necessary; and
   d. the overall sampling weight, which is the product of 7a, 7b and 7c.

Information of the kind specified in this section 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 clusters. 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 using survey data. Some examples of such uses are given in the next section.

4.3.1 Need for Specific Items of Sampling Information

- 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. Subsequent surveys may be linked at the different levels of sampling. For these reasons the probabilities are needed at all sampling stages.

- 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 identification system, as well as the presence of any breaks in the ordering which may occur at the boundaries of explicit strata. It is not uncommon for the area frame to be 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 the sample units selected consecutively are paired together to represent implicit strata.

- 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.

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

- The numbers of interviews attempted and achieved are needed for computation of response rates, and possibly for corrective weighting.

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

- 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.

The frequent occurrence of unforeseen problems of field implementation must be dealt with urgently and 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.

When 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 sample may in turn be used in future surveys, this emphasizes the obligation to provide detailed sample documentation for the benefit of future users.

4.4 Sample Documentation in the Data File

Some of the sampling information mentioned earlier should appear on the survey data file for each individual; this includes: (a) stratum identifier (at least for explicit strata); (b) PSU identifier (in the order of selection); and, (c) weight (if any). Documentation accompanying the data file should, of course, identify these items.

References

Institute for Resource Development Inc. (IRD). 1987. *Sampling Manual.* DHS Basic Documentation No. 8. Columbia, Maryland: IRD.

International Statistical Institute. 1975. *Manual on Sample Design.* World Fertility Survey Basic Documentation. Voorburg, Netherlands: International Statistical Institute.

Kish, Leslie. 1965. *Survey Sampling*. New York: Wiley.

Lê, Thanh. 1993. *Sampling practice in the Demographic and Health Surveys.* Paper presented at the 49th Session of the International Statistical Institute, August, Firenze, Italy.

Macro International Inc. 1995. *Guidelines for the DHS-III main survey report.* Working draft. In-house publication.

Mitra, S.N., Charles Lerman, and Shahidul Islam. 1993. *Bangladesh Contraceptive Prevalance Survey - 1991: Final report.* Dhaka: Mitra and Associates.

Rutstein, Shea and Mohamed Ayad. 1996. *The use of panels in studying demographic trends: Experience from Morocco.* Paper presented at the annual meeting of the Population Association of America, May, New Orleans, Louisiana.

Scott, Chris and Trudy Harpham. 1987. Sample Design. *In The World Fertility Survey: An assessment,* ed. J. Cleland, C. Scott, and D. Whitelegge. London: Oxford University Press. 309-346.

Verma, Vijay. 1991. *Sampling Methods*. Training Handbook. Tokyo, Japan: Statistical Institute for Asia and the Pacific.

APPENDIX A: EXAMPLE OF MANUAL FOR MAPPING AND HOUSEHOLD LISTING

The following is merely a model manual that needs to be adapted to each country's situation and sample design. The cluster is the smallest area unit selected for the DHS. Depending on the type of sampling frame used, the cluster could be many things. If the sampling frame is the latest population census conducted in the country, then a cluster could be an enumeration area as defined by the census. If the sampling frame is the roster of villages, in the case of the rural areas, then the cluster could be an entire village, a segment of the village or a group of villages. If the sampling frame is a list of urban blocks, then the cluster could be an urban block. The definitions of terms, in section A.4, should match the definitions adopted by the local statistical office that maintains the sampling frame. The administrative units in which the clusters are located are country-specific.

A.1 Introduction

The DHS is a national sample survey designed to provide information on fertility, family planning, and health. The survey will interview women between the ages of 15 and 49. The women will be from households randomly selected from a set of sample points which are clusters of households. Prior to interviewing, all households located in the selected clusters will be listed. The listing of households for each cluster will be used in selecting the final sample of households to be included in the DHS.

The listing operation consists of visiting each cluster, recording on listing forms a description of every structure together with the names of the heads of the households found in the structure, and drawing a location map of the cluster as well as a sketch map of the structures in the cluster.

A.2 Responsibilities of the Listing Staff

Persons recruited to participate in the listing operation will work in teams consisting of two enumerators. A coordinator will monitor the entire operation.

The responsibilities of the *coordinator* are to:

1. obtain base maps for all the clusters included in the survey;
2. arrange for the reproduction of all listing materials (listing manuals, mapping and listing forms);
3. assign teams to clusters;
4. obtain travel allowances for the teams;
5. arrange for transportation of the teams to the field;
6. monitor the receipt of the completed listing forms at the central office; and
7. verify that the quality of work is acceptable.

The responsibilities of the *enumerators* are to:

1. contact local officials in each cluster to inform them about the listing operation and to obtain their cooperation;
2. identify the boundaries of the cluster;
3. draw a map showing the location of the cluster;
4. draw a detailed sketch map of the cluster;
5. list all the households in the cluster in a systematic manner; and
6. communicate to the coordinator problems encountered in the field and follow his instructions.

The two enumerators in each team must work at the same time in the same area. First, they identify the cluster boundaries together. Then one enumerator prepares the location map and sketch map while the other does the household listing. The sketch map and the household listing form must be prepared in tandem.

A.3 Listing Materials

The materials needed for the household listing operation are:

1. Manual for Mapping and Household Listing
2. Base map of the area containing the cluster
3. Map Information Form (Form DHS/1)
4. Household Listing Form (Form DHS/2)
5. Segmentation Form (Form DHS/3)[1]

A.4 Definition of Terms

A *base map* is a reference map that contains one or more clusters. It shows the boundaries of the clusters, and the principal physical features such as mountains, rivers, and roads.

A *cluster* is the smallest area unit in the DHS.

A *dwelling unit* is a room or group of rooms normally intended as a residence for one household (for example: a single house, an apartment, a group of rooms in a house).

A *structure* is a free-standing building that can have one or more rooms, for residential or commercial use. Residential structures can have one or more dwelling units (for example: single house, apartment building). In the case where one household inhabits several small dwellings, as can sometimes be found in the rural areas, all the dwellings together, whether they are fenced in or not, constitute a structure.

The *head of household* is the person who is acknowledged as such by members of the household and who is usually responsible for the upkeep and maintenance of the household.

A *household* consists of a person or group of related or unrelated persons, who live together in the same dwelling unit, who acknowledge one adult male or female as the head of the household, who share the same housekeeping arrangements, and are considered as one unit. In some cases one may find a group of people living together in the same house, but each person has separate eating arrangements; they should be counted as separate one-person households. Collective living arrangements such as hostels, army camps, boarding schools, or prisons will not be considered as households.

A.5 Locating the Cluster

The coordinator will provide the listing team with a base map containing the cluster assigned to his team. The cluster is identified by a code (for example, cluster code 002). Upon arrival in an area, the team will use the map to identify all the boundaries of the cluster. In most clusters, the boundaries follow easily recognizable natural features such as streams or lakes, and constructed features such as roads or railroads.

---

[1] This form is needed only if segmentation of some area units is necessary.

However, lines may be invisible (especially in rural areas), in which case, the team should obtain assistance from local authorities to identify the boundaries.

Before doing the listing, the team should tour the cluster to determine an efficient route of travel for listing all the structures. Divide the cluster into sections if possible. A section can be a block of structures. It is useful to make a rough sketch map of the cluster indicating the boundaries of the sections, as well as the relative location of landmarks, public buildings (e.g,. such as schools, churches, and markets), and main roads. This rough sketch will serve as guide for the team when they begin the main work.

A.6 Preparing Location and Sketch Maps

The coordinator will designate one enumerator as the *mapper*. The second enumerator will be the *lister*. Although the two have separate tasks to perform, they must move around the cluster together and work in tandem; the mapper prepares the maps, and the lister collects information on the structures (and corresponding households) indicated on the sketch map.

The mapping of the cluster and the listing of the households should be done in a systematic manner so that there are no omissions or duplications. If the cluster consists of a number of blocks, then the team should finish each block before going to the adjacent one. Within each block, start at one corner of the block and move *clockwise* around the block. In the rural area where the structures are frequently found in small groups, the team should work in one group of structures at a time and in each group they can start at the center (choosing any landmark, such as a school, to be the center) and move around it *clockwise*.

On the first page of the Map Information Form (Form DHS/1), the mapper will prepare a location map of the cluster. First, fill in the identification box for the cluster. All the information needed for filling in the identification box is provided by the coordinator. In the space provided, draw a map showing the location of the cluster and include instructions on how to get to the cluster. Include all useful information to find the cluster and its boundaries directly on the map and in the space reserved for observations if necessary.

On the second page of Form DHS/1, draw a sketch map of all structures found in the cluster. It is important that the mapper and lister work together and coordinate their activities, since the structure numbers that the mapper indicates on the sketch map must correspond to the serial numbers assigned by the lister to the same structures.

On the sketch map, mark the starting point with a large X. Place a small square at the spot where each structure in the cluster is located. For any nonresidential structure, identify its use (for example, a store or factory). Number all structures in sequential order beginning with "1." Whenever there is a break in the numbering of structures (for example, when moving from one block to another), use an arrow to indicate how the numbers proceed from one set of structures to another. Although it may be difficult to pinpoint the exact location of the structure on the map, even an approximate location is useful for finding the structure in the future. Add to the sketch map all landmarks (such as a park), public buildings (such as a school or church), and streets or roads. Sometimes it is useful to add to the sketch map landmarks that are found outside the cluster boundaries if they are helpful in identifying other structures inside the cluster.

Use the marker or chalk provided to write on the entrance to the structure the number that has been assigned to the structure. Remember that this is the serial number of the structure as assigned on the household listing form, which is the same as the number indicated on the sketch map. In order to distinguish the DHS number from other numbers that may exist already on the door of the structure, write DHS in front

of the number; for example, on the door of structure number 3, write DHS/3, and on the door of structure number 54 write DHS/54.

A.7 Listing of Households

The lister will use the Household Listing Form (Form DHS/2) to record all households found in the cluster. Begin by entering the identification codes of the cluster. The first two columns are reserved for office use, leave them blank.

Complete the rest of the form as follows:

Column (1) [Serial number of structure]: For each structure, record the same serial number that the mapper enters on the sketch map.

Column (2) [Address/description of structure]: Record the street address of the structure. Where structures do not have visible street addresses (especially in the rural area), give a description of the structure and any details that help in locating it (for example, in front of the school, next to the store, etc.).

Column (3) [Residence Y/N]: Indicate whether the structure is used for residential purposes (eating and sleeping) by writing Y for "Yes," or in cases where a structure is used only for commercial or other purposes, write N for "No." Structures used both for residential and commercial purposes (for example, a combination of store and home) should be classified as residential (ie. mark Y in column 3). Make sure to list any dwelling unit found in a nonresidential structure (for example, a guard living inside a factory).

Do not forget to list vacant structures and structures under construction, as well as structures where the household members refuse to cooperate, or are not at home at the time of the listing. In such cases, the columns that follow (4 and 5) should be left blank, and in Column (6) [Observations], give some explanation (for example: under construction, refusal, not at home, etc.).

Column (4) [Serial number of household in structure]: This is the serial number assigned to each household found in the structure; there can be more than one household in a structure. The first household in the structure will always have number "1." If there is a second household in the structure, then this household should be recorded on the next line, a "2" is recorded in Column (4), and Columns (1) to (3) are left blank.

Column (5) [Name of head of household]: Write the name of the head of the household. There can only be one head per household. If no one is home, ask neighbors for the name of the head of the household. If a name cannot be determined, leave this column blank. Note that it is not the name of the landlord or owner of the structure that is needed, but the name of the head of the household that lives there.

Column (6) [Observations]: This space is provided for any special remarks that might help the interviewing team locate the structure or identify the household during the main survey fieldwork.

If the structure is an apartment building (or a block of flats), assign one serial number to the entire structure (only one square with one number appears on the sketch map), but complete Columns (2) through (6) for each apartment in the building individually. Each apartment should have its own address, which is the apartment number.

The listing team should be careful to locate hidden structures. In some areas, structures have been built so haphazardly that they can easily be missed. If there is a pathway leading from the listed structure, check to see if the pathway goes to another structure. People living in the area may help in identifying the hidden structures.

A.8 Segmentation of Large Area Units[2]

Area units that are very large in population size must be subdivided into several small segments, only one of which will be retained for the survey. In this case, the cluster corresponds to a segment of the area unit. The exact area units that need segmenting will be communicated to the listing teams together with the number of segments to be made in each large area unit. When the teams arrive in the area unit, they should first identify the boundaries of the area unit and draw a location map of the area unit, then proceed to segment it.

The ideal would be to have segments of approximately equal size, but it is also important to adopt segment boundaries that are easily identifiable. First draw a sketch map of the entire area unit. Using identifiable boundaries such as roads, streams, and electric power lines, divide the area unit into the designated number of roughly equal-sized segments. On the map of the cluster, show clearly the boundaries of the segments created. Number the segments sequentially. Estimate the size of each segment in the following manner: quickly count the number of dwellings in each segment, add them up and calculate the proportion of dwellings for each segment.

Example: A cluster of 620 dwellings has been divided into 3 segments and the results are as follows:

Segment 1 :	220 dwellings,	or	220/620	= 35 percent
Segment 2 :	190 dwellings,	or	190/620	= 30 percent
Segment 3 :	210 dwellings,	or	210/620	= 35 percent
Total :	620 dwellings,	or	620/620	= 100 percent

On Form DHS/3 (Segmentation Form), write the size of the segments in the appropriate columns (number and percent) and calculate the cumulative size (percent). The last cumulative size must be equal to 100.

Example:

Segment Number	Number of dwellings	Percent	Cumulative percent
1	220	35	35
2	190	30	65
3	210	35	100

For each large area unit to be segmented, a random number will be selected in the central office and included in the file. Compare this random number with the cumulative size. Select the first segment whose cumulative size is greater than or equal to the random number.

[2]This section is to be removed from the manual if no segmentation of large area units is necessary.

Example: Random number: 67
Segment selected: Segment number 3

Draw a detailed sketch map of the selected segment and list all households found in the selected segment.

A.9 Quality Control

To ensure that the work done by each listing team is acceptable, a quality check will be performed. The coordinator will do an independent listing of 10 percent of each cluster. If errors are found in 2 percent or more of the relisted sample, the whole cluster will be relisted. If less than 2 percent of the relisted sample are wrong, corrections will be made on the household listing form, and no relisting is necessary.

A.10 Examples of Symbols for Mapping, and Mapping and Listing Forms

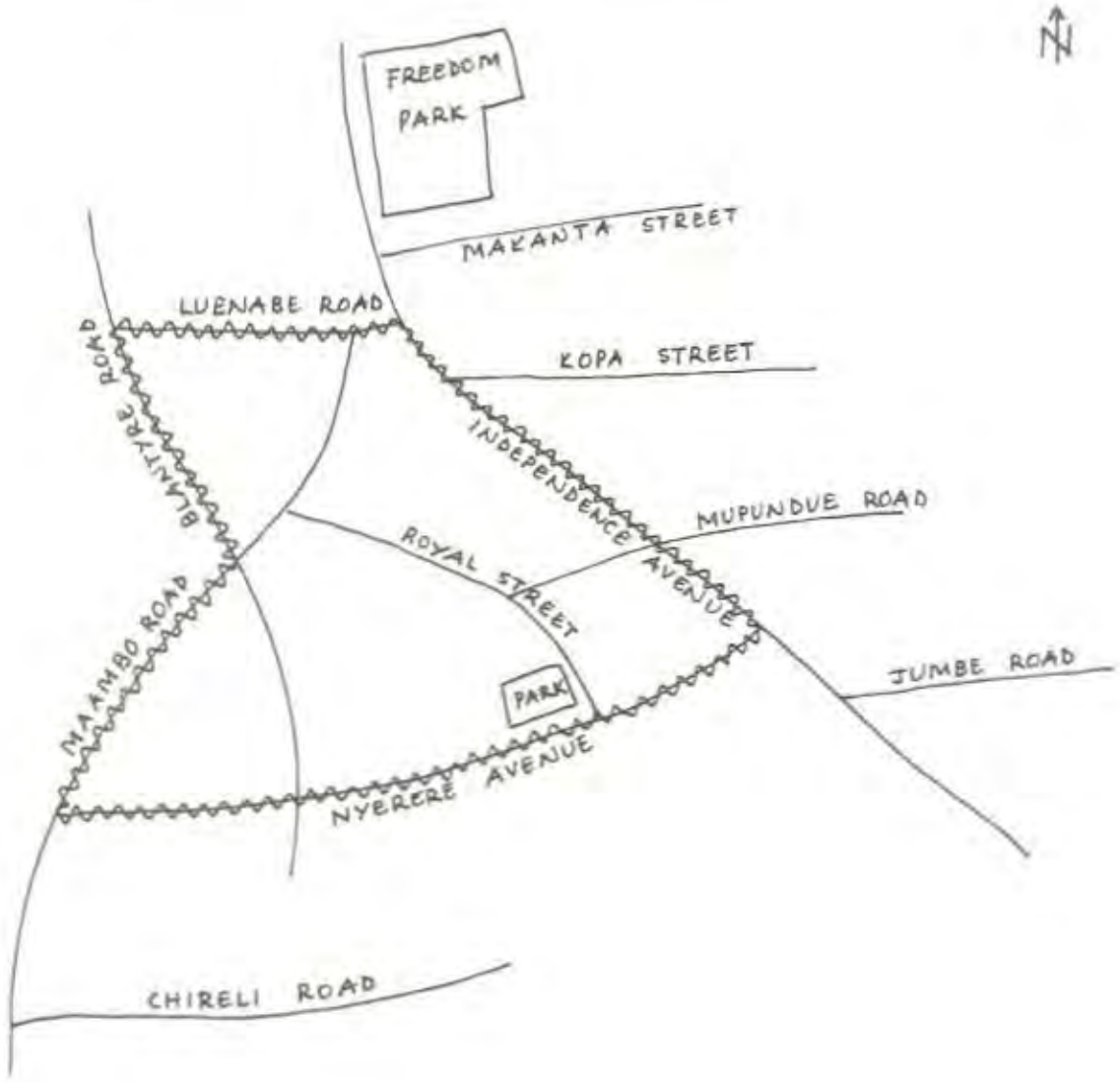
STANDARD SYMBOLS SUGGESTED FOR MAPPING

Orientation to the North	
Boundaries of the cluster	
Paved road	
Unpaved (dirt) road	
Footpath	
River, creek, etc.	
Bridge	
Lake, pond, etc.	
Mountains, hills	
Water point (wells, fountain, etc.)	
Market	
School	
Administrative building	
Church, temple	
Mosque	
Cemetery	
Residential structure	
Non-residential structure	
Vacant structure	
Hospital, clinic, etc.	
Electric pole	
Tree, bush	

IDENTIFICATION			
PROVINCE	KAYES	PROVINCE CODE	1
DISTRICT	DIEMA	DISTRICT CODE	04
TOWN~~/VILLAGE~~	DIEMA	TOWN/VILLAGE CODE	02
NAME OF MAPPER	Harrison Sidibe	CLUSTER CODE	017
NAME OF LISTER	John Melaku	DHS CLUSTER N°	001

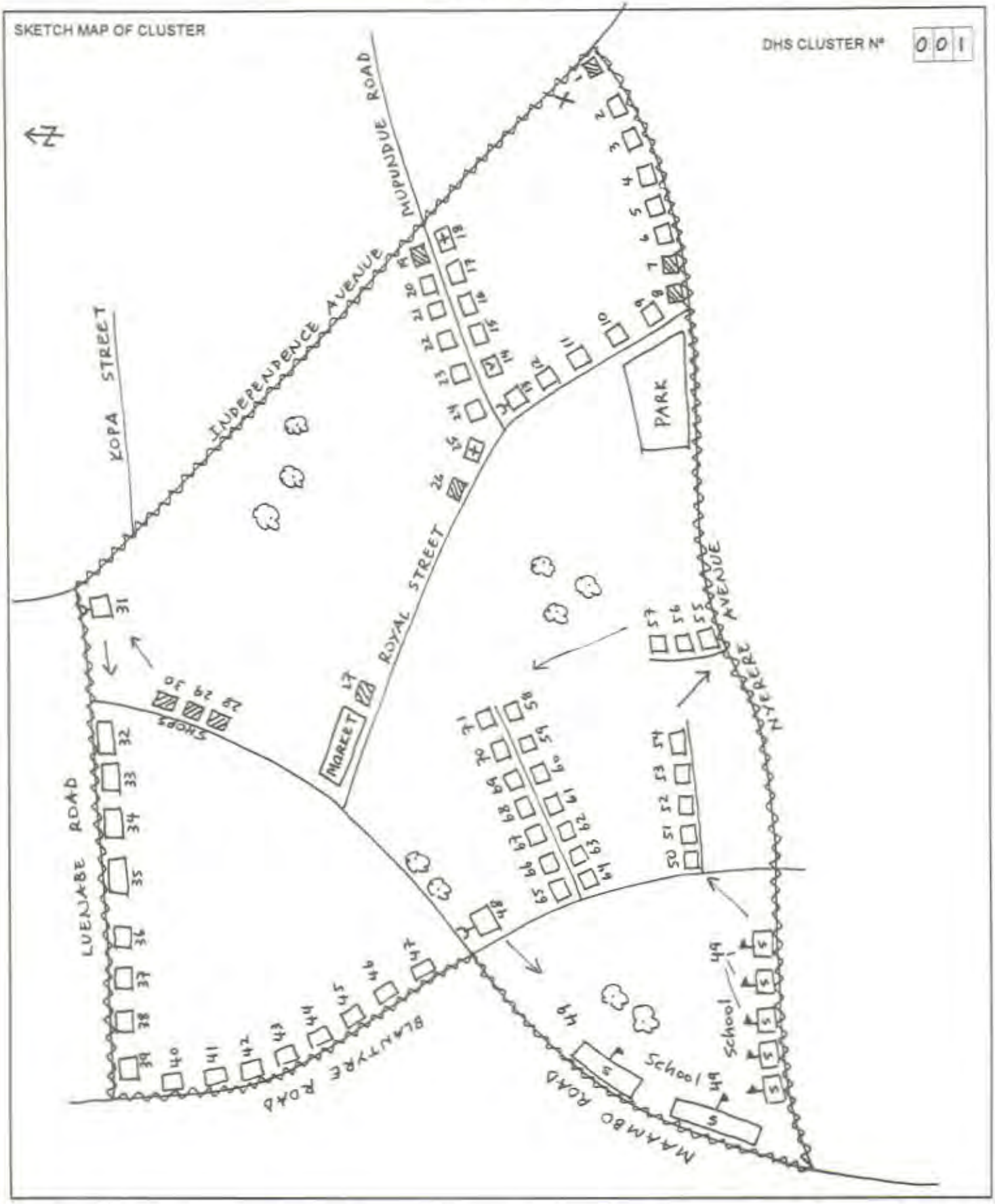
OBSERVATIONS:

LOCATION MAP OF CLUSTER



SKETCH MAP OF CLUSTER

DHS CLUSTER N° 0 0 1



Form DHS/2

DEMOGRAPHIC AND HEALTH SURVEY
HOUSEHOLD LISTING FORM

DHS CLUSTER N° 001

LEAVE BLANK		SERIAL N° OF STRUC-TURE	ADDRESS/DESCRIPTION OF STRUCTURE	RESIDENCE Y/N	SERIAL N° OF HOUSEHOLD IN STRUCTURE	NAME OF HEAD OF HOUSEHOLD	OBSERVATIONS
HH TO INTERVIEW	HH NUMBER	(1)	(2)	(3)	(4)	(5)	(6)
		1	Nyerere Avenue	N			Pharmacy Star
		2	6 Nyerere Avenue	Y	1	Diane Obote	
		3	8 Nyerere Avenue	Y	1	Eugene Kariba	
					2	Dorothy Uchi	
		4	10 Nyerere Avenue	Y	1		No one at home.
		5	12 Nyerere Avenue	Y	1	Sam Lowa	
		6	14 Nyerere Avenue	Y	1	Harrison Coulibali	
					2	Paul Liande	
					3	Harry Fiwale	
		7	Avenue Nyerere	N			In construction
		8	Nyerere Avenue	N			In construction
		9	22 Royal Street	Y	1	George Sidibi	
		10	20 Royal Street	Y	1		Refused
		11	18 Royal Street	Y	1	Chief Seidou	
		12	16 Royal Street	Y	1	Ann Tonde	
		13	Mupundue Road	N			Mosque
		14	4 Mupundue Road	N			Vacant
		15	6 Mupundue Road	Y	1	Suzanne Ibenga	
		16	8 Mupundue Road	Y	1	David Chouta	
					2	Joseph Lupiya	
		17	10 Mupundue Road	Y	1	Eleni Fahmi	
		18	10A Mupundue Road	Y	1	Doctor Tadesse	Home upstairs, clinic downstairs

Form DHS/3

DEMOGRAPHIC AND HEALTH SURVEY
SEGMENTATION FORM

IDENTIFICATION	
PROVINCE KOULIKORO	PROVINCE CODE 4
DISTRICT DIOLA	DISTRICT CODE 02
TOWN~~/VILLAGE~~ DIONGAGA	TOWN/VILLAGE CODE 06
NAME OF MAPPER WOLDE CONATE	CLUSTER CODE 023
NAME OF LISTER ANDRE LUENA	DHS CLUSTER Nº 015

NUMBER OF SEGMENTS TO BE CREATED 03

Segment Number	Number of dwellings	Percent	Cumulative percent
1	220	35%	35%
2	190	30%	65%
3	210	35%	100%
4			
5			
6			
7			
8			
9			
10			
11			
12			

RANDOM NUMBER BETWEEN 1 AND 100: 067

SEGMENT SELECTED: 03

APPENDIX B: EXAMPLE OF A FULL SAMPLE DESCRIPTION

B.1 INTRODUCTION

The Bangladesh Demographic and Health Survey (BDHS) covered the population residing in private households throughout the country. The design for the BDHS called for a representative probability sample of 10,000 completed individual interviews with ever-married women under the age of 50. It was designed to produce reliable estimates of all variables for each of the five administrative divisions in the country, in addition to urban and rural estimates.

The area sampling frame used for the BDHS was the new Integrated Multipurpose Master Sample (IMPS), developed by the Bangladesh Bureau of Statistics (BBS) following the 1991 Population Census, to be used for demographic and socioeconomic surveys.

B.2 THE INTEGRATED MULTIPURPOSE MASTER SAMPLE DESIGN

The sample for the IMPS was designed to be nationally representative, stratified and selected in two stages. Each of the five administrative divisions of the country was stratified into three groups: 1) statistical metropolitan areas (SMAs),[1] 2) municipalities, and 3) other rural areas. As Barisal division did not include any SMA, this stratification scheme resulted in 14 strata.

In the rural areas, the primary sampling unit (PSU) was the *mauza*. In total, 52 *mauzas* were selected with probability proportional to size (PPS), the size being the number of households counted in the 1991 census. The selection of the *mauzas* was a systematic, one-stage operation, carried out independently for each of the five rural strata. In each rural stratum, implicit geographic stratification was introduced by ordering the *zillas* in a serpentine manner; similarly, *thanas* within *zillas* were also ordered geographically. *Mauzas* themselves were ordered by census codes within the *union parishads* that constituted the *thanas*; *union parishads* were also ordered sequentially by census codes.

In the urban areas (municipalities and greater parts of SMAs), the PSU was the *mahallah*. There were 70 *mahallahs* in the SMAs and 50 in the municipalities, selected with PPS. The selection procedure was the same as for the rural areas, again independent for each of the nine urban strata. Implicit stratification was achieved as follows: *mahallahs* within *wards* within *thanas* within *zillas*. *Mahallahs* and *wards* were ordered sequentially by census codes while *thanas* and *zillas* were ordered geographically in a serpentine manner. Table B.2.1 shows the allocation of the 372 PSUs to the five divisions.

---

[1] In Bangladesh, the SMAs are extensions of the divisions' headquarters and include rural areas surrounding these headquarters. According to BBS, these "rural" areas, because of their proximity to the cities, are closer to having urban characteristics than rural. Consequently, the SMAs are considered urban in designing the master sample for the IMPS. However, areas can still be identified as urban or rural within the SMAs.

Table B.2.1 IMPS Distribution of PSUs

Division	Total	SMA	Municipality	Rural
Barisal	36	0	10	26
Chittagong	86	16	10	60
Dhaka	114	35	10	69
Khulna	48	11	8	29
Rajshahi	88	8	12	68
Total	372	70	50	252

As such, the design for the IMPS gave more weights to Barisal division which represented only about 7 percent of the population in 1991, thus improving estimates at the divisional level. While the rural allocation was close to proportionality, the urban sample was being shifted between SMA and municipalities within division so as to allow for better urban estimates. In Table B.2.2, the proportional allocation of the 372 PSUs is presented to allow comparison between the two allocation schemes:

Table B.2.2 Proportional distribution of PSUs

Division	Total	SMA	Municipality	Rural
Barisal	3	0	3	20
Chittagong	8	18	8	63
Dhaka	132	50	10	72
Khulna	44	9	6	29
Rajshahi	84	4	12	68
Total	372	81	39	252

B.3 CHARACTERISTICS OF THE BDHS SAMPLE

The sample for the BDHS was selected in two effective stages from the IMPS master sample. In the first stage, 304 PSUs were selected in each stratum with equal probabilities. Since the PSUs in the IMPS master sample were selected with PPS from the sampling frame, equal probability selection of a subsample of these PSUs for the BDHS was equivalent to selection with PPS from the entire sampling frame. A complete listing of the households in the selected PSUs was carried out. The list of households obtained was used as the frame for the second-stage sampling, which was the selection of the households visited by the BDHS interviewing teams during the main survey fieldwork. Ever-married women under age 50 were identified in these households and interviewed.

An intermediate sampling stage was introduced after the first stage due to the relatively large size of the PSUs. This intermediate sampling stage consisted of selecting one urban block or village with PPS within each selected PSU. This was not considered an *effective* sampling stage since the purpose was only to reduce the size of the PSU (hence, reducing the household listing workload) and only where it was feasible, i.e., only in PSUs that were made up of more than one urban block or village.

B.4 SAMPLE ALLOCATION

Table B.4.1 shows the distribution of the total number of households in Bangladesh to the 14 strata according to the 1991 census:

Table B.4.1 Household distribution (1991 census)

Division	Total	SMA	Municipality	Rural
Barisal	1,425,051	0	70,288	1,354,763
Chittagong	4,949,766	418,909	185,428	4,345,429
Dhaka	6,376,882	1,191,566	239,611	4,945,705
Khulna	2,382,235	217,413	144,399	2,020,423
Rajshahi	5,053,185	98,462	286,725	4,667,998
Total	20,187,119	1,926,350	926,451	17,334,318

A proportional allocation of 10,000 women to the 14 strata would yield the sample distribution in Table B.4.2.

Table B.4.2 Proportional sample distribution

Division	Total	SMA	Municipality	Rural
Barisal	706	0	35	671
Chittagong	2,452	208	92	2,153
Dhaka	3,159	590	119	2,450
Khulna	1,180	108	72	1,001
Rajshahi	2,503	49	142	2,312
Total	10,000	955	460	8,587

While the samples for the four largest divisions would be sufficiently large for providing reliable estimates, it was not the case for Barisal division. For this reason, it was necessary to increase the sampling rate for Barisal relative to the other divisions. Results of other demographic and health surveys show that a minimum sample of 1,000 women is required in order to obtain estimates of fertility rates at an acceptable level of sampling errors. The sample allocation in Table B.4.3 was proposed after taking into account four different factors: (1) a minimum sample of 1,000 completed interviews for Barisal; (2) a minimum sample of 1,000 completed interviews for each of the two urban domains (SMAs and municipalities); (3) close to it is expected that about 25 completed interviews of urban women and 35 completed interviews of rural a proportional allocation within each rural, SMA, and municipality domain; and (4) the number of PSUs available in the IMPS master sample.

Table B.4.3 Disproportional sample allocation

Division	Total	SMA	Municipality	Rural
Barisal	1,000	0	100	900
Chittagong	2,400	200	200	2,000
Dhaka	3,000	650	250	2,100
Khulna	1,150	100	150	900
Rajshahi	2,450	50	300	2,100
Total	10,000	1,000	1,000	8,000

The number of households to be selected so as to yield the above target sample is calculated as follows:

$$Number\ of\ households = \frac{Number\ of\ women}{Number\ of\ women\ per\ household \times Overall\ response\ rate}$$

According to the *1991 Bangladesh Contraceptive Prevalence Survey* there were 1.07 ever-married women per rural household and 1.13 ever-married women per urban household. The overall response rate found in the same survey was around 92 percent. Using a slightly lower expected overall response rate of 90 percent, the number of households to be selected for the BDHS were distributed as shown in Table B.4.4.

Table B.4.4 Number of households to be selected

Division	Total	SMA	Municipality	Rural
Barisal	1,033	0	98	935
Chittagong	2,471	197	197	2,077
Dhaka	3,066	639	246	2,181
Khulna	1,180	98	147	935
Rajshahi	2,525	49	295	2,181
Total	10,275	983	983	8,309

The number of PSUs selected for each stratum was calculated by dividing the number of households to be selected by the average take in the PSU. Analytical studies of surveys of the same nature suggest that the optimum number of women to be interviewed is around 20-25 in each urban PSU and 30-35 in each rural PSU. If on average 25 households were selected in each urban PSU and 37 households in each rural PSU, it is expected that about 25 completed interviews of urban women and 35 completed interviews of rural women (considering an overall response rate of 90 percent, and the number of ever-married women of 1.07 and 1.13 in rural and urban areas, respectively) would be obtained. Table B.4.5 shows the final distribution of PSUs.

Table B.4.5 Number of PSUs

Division	Total	SMA	Municipality	Rural
Barisal	29	0	4	25
Chittagong	72	8	8	56
Dhaka	95	26	10	59
Khulna	35	4	6	25
Rajshahi	73	2	12	59
Total	304	40	40	224

B.5 SYSTEMATIC SELECTION OF PSUs

The 304 PSUs were selected from the IMPS master sample in a systematic manner, with equal probability, and independently in each stratum. The selection interval was calculated as follows:

$$I_h = \frac{A_h}{a_h}$$

where A_h was the number of PSUs that were in the IMPS master sample frame for the h^{th} stratum, and a_h was the number of PSUs to be selected for BDHS.

If a selected PSU was made up of more than one urban block or village, then the r^{th} urban block/village was selected for the BDHS, where r was a random number between 1 and the number of urban blocks/villages that make up the selected PSU.[2]

B.6 SAMPLING PROBABILITIES

The sampling probabilities were calculated separately for each sampling stage and for each stratum. The following notations were used:

P_{ihi}: Sampling probability for the i^{th} PSU in the h^{th} stratum according to the IMPS frame.[3]

---

[2] This is the simplified procedure for selecting *one* unit with PPS. The selection probability is m_{ij} / M_i where m_{ij} is the size of the j^{th} unit in the i^{th} selected PSU and M_i is the total size of the i^{th} selected PSU.

[3] In the IMPS, the sampling probability for the i^{th} PSU in the h^{th} stratum was calculated as:

$$P_{ihi} = \frac{A_h M_{hi}}{\sum_i M_{hi}}$$

where A_h was the number of PSUs in the IMPS for the h^{th} stratum, M_{hi} is the size of the ith selected PSU, and ΣM_{hi} is the size of the h^{th} stratum.

P_{2hi}: Sampling probability for the i^{th} PSU in the h^{th} stratum for BDHS.
P_{3hi}: Sampling probability for the household in the i^{th} PSU of the h^{th} stratum.

P_{2hi} is calculated as follows:

$$P_{2hi} = \frac{a_h}{A_h} \times \frac{m_{hij}}{M_{hi}}$$

The first term in the above equation is the inverse of the selection interval from section B.5. In the second term, m_{hij} is the size of the j^{th} unit selected in the i^{th} PSU; if there is no sub-selection in the i^{th} PSU, then this second term is equal to 1. M_{hi} is the size of the i^{th} selected PSU.

It should be noted that field segmentation (and subsequent selection of one segment) may be necessary in the PSUs that are large in size but are not made up of more than one urban block/village. In the case of segmentation, P_{2hi} is calculated as above with m_{hij} being the size of the segment selected.

In order for the sample to be self-weighting within each stratum, the stratum overall probability $f_h = P_{1hi}P_{2hi}P_{3hi}$ must be the same for each household in the sample. That is,

$$P_{3hi} = \frac{f_h}{P_{1hi}\ P_{2hi}} \quad with$$

$$f_h = \frac{n_h}{N_h}$$

where n_h was the number of households selected in the h^{th} stratum and N is the projected number of households in the year of the survey (1993) for the h^{th} stratum.

The selection of households was systematic with equal probabilities, and the household sampling interval I_{hi} in the i^{th} cluster of the h^{th} stratum was calculated as:

$$I_{hi} = \frac{1}{P_{3hi}}$$

For each cluster, a list of households was obtained for BDHS prior to the main survey fieldwork, and the interval just stated was applied to the list in order to select the households.

To compensate for the disproportionate allocation of the sample among strata, design stratum weights were calculated as follows:

$$W_h = \frac{f}{f_h}$$

where f_h was as stated earlier, and where f was the overall sampling fraction, calculated as follows:

$$f = \frac{n}{N}$$

where n was the number of households selected for the BDHS, and N was the estimated number of households in 1993 for Bangladesh.

APPENDIX C: EXAMPLES OF dBASE PROGRAMS FOR SAMPLE SELECTION

```
*  Example 1.  ESELECT.PRG
*  Lines preceded by * are comments
*
*  SELECTION OF CLUSTERS WITH EQUAL PROBABILITY
*
*  The sampling frame FRAME.DBF contains the following data for each cluster::
*  1.  REGION     Character  Length 2     [Region code]
*  2.  DISTRICT   Character  Length 2     [District code]
*  3.  CLUSTER    Character  Length 3     [Cluster code]
*  4.  POP        Numeric    Length 9     [Population size of the cluster]
*  5.  HH         Numeric    Length 4     [Number of households in the cluster]
*
*  Reminder: Modify structure to add the following fields to the database before running
*  ESELECT.PRG.  This must be done interactively.
*  6.  SELECTED   Character  Length 6     [to mark selected cluster]
*  7.  SERIALNO   Numeric    Length 6     [serial number of cluster]
*
*  Variable names in capital letters indicate fields in the database; variable names in small letters
*  indicate working variables.
*
set talk off
use FRAME
go bottom
records=recno()
n=0
do while n < records
    n=n+1
    goto n
    replace SERIALNO with n
enddo
*
*  The values of xrandom and units must be provided before running the program; xrandom is
*  the random start; units is the number of units to be selected.
*
xrandom = .1267
units = 10
set talk off
go bottom
records=recno()
clustotal=SERIALNO
interval=round(clustotal/units,2)
```

```
ran=int(xrandom*interval)
s=0
xselect=ran
n=1
goto n
do while (xselect <= clustotal).and. (n <= records)
    do while (SERIALNO <= int(xselect)) .and. (n <= records)
        n = n+1
        goto n
    enddo
    s=s+1
    replace SELECTED with 'S '+str(s,3)
    xselect=xselect+interval
    n=n+1
    goto n
enddo
*
*   Copy selected units to SAMPLE.DBF and examine selected units.
*
copy to SAMPLE.DBF for len(trim(SELECTED)) > 0
set talk on
use SAMPLE
browse

*   Example 2.  PSELECT.PRG
*   Lines preceded by * are comments
*
*   SELECTION OF CLUSTERS WITH PROBABILITY PROPORTIONAL TO SIZE
*   CALCULATION OF SELECTION PROBABILITIES
*
*   In this example, it is assumed that the measure of size used for PPS selection is the population
*   size, hence POP and CUMPOP are used.  However, if the number of households in each cluster
*   is used as the measure of size, then HH and, consequently CUMHH, must be used.
*
*   The sampling frame FRAME.DBF contains the following data for each cluster:
*   1.  REGION      Character  Length 2      [Region code]
*   2.  DISTRICT    Character  Length 2      [District code]
*   3.  CLUSTER     Character  Length 3      [Cluster code]
*   4.  POP         Numeric    Length 9      [Population size of the cluster]
*   5.  HH          Numeric    Length 4      [Number of households in the cluster]
*
*   Reminder: Modify structure to add the following fields to the database before running
*   PSELECT.PRG.  This must be done interactively.
*   6.  SELECTED    Character  Length 6      [to mark selected cluster]
*   7.  SERIALNO    Numeric    Length 6      [serial number of cluster]
```

```
*   8.  CUMPOP      Numeric   Length 9                [Cumulative population size of cluster]
*   9.  P1          Numeric   Length 9   Decimal 6    [Selection probability of selected cluster]
*
*   Variable names in capital letters indicate fields in the database; variable names in small letters
*   indicate working variables.
*
use FRAME
set talk off
go bottom
records=recno()
n=0
cum=0
do while n < records
    n=n+1
    goto n
    cum=cum+POP
    replace CUMPOP with cum
    replace serialno with n
enddo
*
*   The values of xrandom and units must be provided before running the program; xrandom is
*   the random start; units is the number of units to be selected.
*
xran=.2967
units=60
set talk off
go bottom
records=recno()
cumtot=CUMPOP
interval=int(cumtot/units)
ran=int(xran*interval)
s=0
xselect=ran
n=1
goto n
do while (xselect <= cumtot) .and. (n <= records)
    do while (CUMPOP < xselect) .and. (n <= records)
        n = n+1
        goto n
    enddo
    s=s+1
    xp1=units*POP/cumtot
    replace selected with 'R  '+str(s,3)
    replace P1 with xp1
    xselect=xselect+interval
```

```
    n=n+1
    goto n
enddo
*
*   Copy selected units to SAMPLE.DBF and examine selected units.
*
copy to SAMPLE.DBF for len(trim(selected)) > 0
set talk on
use SAMPLE
browse

*   Example 3. STRAT.PRG
*   Lines preceded by * are comments.
*
*   STRATIFICATION
*   Dbase can also be used to stratify the clusters according to some criteria.  In this example, we
*   will stratify the clusters in the sampling frame FRAME.DBF into regions and type of residence.
*   Note that, generally, selection of sample points is done independently in each stratum once
*   stratification is achieved.
*
*   The sampling frame FRAME.DBF contains the following data for each cluster:
*   1.  DISTRICT    Character  Length 2    [District code]
*   2.  CLUSTER     Character  Length 3    [Cluster code]
*   3.  POP         Numeric    Length 9    [Population size of the cluster]
*   4.  HH          Numeric    Length 4    [Number of households in the cluster]
*
*   Reminder: Modify structure to add the following fields to the database before running
*   PSELECT.PRG.  This must be done interactively.
*   5.  REGION      Character  Length 2    [Newly created region]
*   6.  TYPE        Numeric    Length 1    [Type of residence; 1=urban/2=rural]
*
*   Variable names in capital letters indicate fields in the database; variable names in small letters
*   indicate working variables.
*
use FRAME
set talk off
go bottom
records=recno()
n=0
do while n < records
    n=n+1
    goto n
    do case
        case val(DISTRICT)=1.or.val(DISTRICT)=4.or.val(DISTRICT)=7
        xregion=1
```

```
        case val(DISTRICT)=2.or.val(DISTRICT)=6
        xregion=2
        case val(DISTRICT)=3.or.val(DISTRICT)=5
        xregion=3
        case val(DISTRICT) > 7
        xregion=4
  endcase
  replace REGION with xregion
  do case
    case val(CLUSTER)<600
    xtype=1
    case val(CLUSTER)>600
    xtype=2
  endcase
  replace TYPE with xtype
enddo
set talk on
```