UNITED NATIONS SECRETARIAT
Statistics Division

ESA/STAT/AC.93/2
03 November 2003

English only

Expert Group Meeting to
Review the Draft Handbook on
Designing of Household Sample Surveys
3-5 December 2003

D R A F T

Sampling strategies[*]

by

Anthony G. Turner[**]

[*] This document is being issued without formal editing.

[**] The views expressed in this paper are those of the author and do not imply the expression of any opinion on the part of the United Nations Secretariat.

Table of contents

Chapter Two: Sampling strategies

1. Virtually all sample designs for household surveys, both in developing and developed countries, are complex because of their multi-stage, stratified and clustered features. In addition national-level household sample surveys are often general-purpose in scope, covering multiple topics of interest to the government and this also adds to their complexity. The handbook naturally focuses on multi-stage sampling strategies. A review of basic sampling techniques is presented in the annex.

2. Analogous to a symphonic arrangement, a good sample design for a household survey must combine, harmonically, numerous elements in order to produce the desired outcome. The sample must be selected in *stages* to pinpoint the locations where interviews are to take place and to choose the households efficiently. The design must be *stratified* in such a way that the sample actually selected is spread over geographic sub-areas and population sub-groups properly. The sample plan must make use of *clusters* of households in order to keep costs to a manageable level, while simultaneously avoiding being overly *clustered* because of the latter's damaging effects on reliability. The *size* of the sample must take account of competing needs so that costs and precision are optimally maintained, while also addressing the urgent needs of users who desire data for sub-populations or sub-areas – *domains*. The design must seek maximum accuracy by construction of a *sample frame* that is asymptotically complete, correct and current and by use of selection techniques that minimize unintentional bias sometimes caused by the implementers. The design should also be self-evaluating in the sense that *sampling errors* can be and are estimated to guide users in gauging the reliability of the key results.

3. This chapter and the following one discuss in detail each of the features that go into designing a proper sample design for a household survey. Because of the crucial importance of sample frames in achieving good sample practice a separate chapter is devoted to it – chapter 3.

2.1. Probability sampling versus other sampling methods for household surveys

4. While it is beyond the scope of this handbook to provide a discussion of probability theory, it is important to explain how probability methods play an indispensable role in sampling for household surveys. A brief description of probability sampling, its definition and why it is important are given in this section. Other methods such as judgmental or purposive samples, random "walk," quota samples and convenience sampling that do not meet the conditions of probability sampling are briefly mentioned and why such methods are not generally recommended for household surveys.

2.1.1. Probability sampling

5. Probability sampling in the context of a household survey refers to the means by which the elements of the target population - geographic units, households and persons - are selected for inclusion in the survey. The requirements for probability sampling are that each element must have a known mathematical chance of being selected and that chance must be greater than

zero and numerically calculable. It is important to note that the chance of each element being selected need not be equal but can vary in accordance with the objectives of the survey.

6. It is this mathematical nature of probability samples that permits scientifically-grounded estimates to be made from the survey. More importantly it is the foundation upon which the sample estimates can be inferred to represent the total population from which the sample was drawn. A crucial feature and by-product of probability sampling in surveys is that sampling errors can be estimated from the data collected from the sample cases, a feature that is not mathematically justifiable when non-probability sampling methods are used.

2.1.1.1. Probability sampling in stages

7. As implied in the first paragraph of this subsection, probability sampling must be used at each stage of the sample selection process in order for the requirements to be met. For example, the first stage of selection generally involves choosing geographically-defined units such as villages while the last stage involves selecting the specific households or persons to be interviewed. Each stage must utilize probability methods for proper sampling. To illustrate, a simplified example is given below.

- ***Example***

Suppose a simple random sample, *SRS*, of 10 villages is selected from a total of 100 villages in a rural province. Suppose further that for each sample village a complete listing of the households is made from which a systematic selection of 1 in every 5 is made for the survey interview, no matter what the total number is in each village. This is a probability sample design, selected in two stages with probability at the first stage of 10/100 and at the second stage of 1/5. The overall probability of selecting a particular household for the survey is 1/50, that is, 10/100 multiplied by 1/5.

8. Though not particularly efficient, the sample design of the example nevertheless illustrates how both stages of the sample utilize probability sampling. Because of that the survey results can be estimated in an *unbiased* way by properly applying the probability of selection at the data analysis stage of the survey operation (see discussion of survey weighting in chapter 5).

2.1.1.2. Calculating the probability

9. The example in the preceding subsection also illustrates two other requirements for probability sampling. First, each village in the province was given a *non-zero* chance of being selected. By contrast, if one or more of the villages had been ruled out of consideration for whatever reason such as security concerns, the chance of selection of those villages would be zero and the probability nature of the sample would thus be violated. The households in the above example were also selected with non-zero probability, but if some of them had been purposely excluded due to, say, inaccessibility, they would have had zero probability and the sample implementation would then revert to a non-probability design.

10. Second, the probability of selecting both the villages and the households can actually be *calculated* based on the information available. In the case of selecting villages, both the sample

size (10) and the population size (100) are known, and those are the parameters that define the probability, 10/100. For households, calculation of the probability is slightly different because we do not know, in advance of the survey, how many households are to be selected in each sample village. We are simply instructed to select 1 in 5 of all of them; so that if there is a total of 100 in Village A and 75 in Village B we would select 20 and 15 respectively. Still, the probability of selecting a household is 1/5, irrespective, of the population size or the sample size (20/100 = 1/5 but so does 15/75).

11. Referring still to the illustration above, the second-stage selection probability could be calculated as a cross-check after the survey is completed when m_i and M_i are known, where m_i and M_i are, respectively, the number of sample households and total households in the i^{th} village. The probability would be equal to m_i/M_i, and there would be 10 such probabilities – one for each sample village. As was noted, however, this ratio is always 1/5 for the design specified and so it would be superfluous to obtain the counts of sample and total households for the sole purpose of calculating the second-stage probability. For quality control purposes it would nevertheless be useful to obtain those counts to ensure that the 1 in 5 sampling rate was applied accurately.

2.1.1.3. When target population is ill-defined

12. Sometimes the conditions for probability sampling are violated because of loose criteria in defining the *target population* that the survey is intended to cover. For example, the desired target population may be all households nationwide, yet when the survey is designed and/or implemented certain population sub-groups are intentionally excluded. This occurs, for example, if a country excludes such groups as nomadic households, boat people or whole areas that are inaccessible due to the terrain. Other examples occur when a target population is intended to cover a restricted, special population such as ever-married women or young people under 25 years old, but yet important sub-groups are omitted for various reasons. For example a target population may be intended to cover youth under 25 but those in the military, jail or otherwise institutionalized are excluded.

13. Whenever the actual target population that the survey covers departs from the one intended, the survey team should take care to re-define the target population for more accurately, not only for clarification to users of the survey results but also in order to meet the conditions of probability sampling. In the example of the youth under 25 above, the target population should be more precisely described and re-defined as *civilian, non-institutional youth under 25 years old.* Otherwise, survey coverage should be expanded to include the omitted sub-groups.

14. Thus, it is important to define the target population very carefully to cover only those members that will actually be given a *chance of selection* in the survey. In cases where sub-groups are intentionally excluded, it is of course crucial to apply probability methods to the actual population that is in scope for the survey. Furthermore, it is incumbent upon the survey directors to clearly describe to the users, when results are released, which segments of the population the survey includes and which are excluded.

2.1.2. Non-probability sampling methods

15. In contrast to probability sampling, there is no statistical theory to guide the use of non-probability samples, which can only be assessed through subjective evaluation. Failure to use probability techniques means, therefore, the survey estimates will be biased. Moreover, the magnitude of these biases and often their direction (under-estimates or over-estimates) will be unknown. As mentioned earlier, the precision of sampling estimates, that is, their standard errors, can be estimated when probability sampling is used. This is necessary in order for the user to gauge the reliability of the survey estimates. As mentioned above, unbiased estimates can also be made with probability sampling, though this is not always the best choice when the estimation procedures are developed (see further discussion on this point in chapter 5).

16. In spite of their theoretical deficiencies, non-probability samples are frequently used in various settings and situations, mainly because of cost, convenience and even apprehension on the part of the survey team that a "random" sample may not properly represent the target population. In the context of household surveys we will briefly discuss various types of non-probability samples, chiefly by way of example, and indicate some of the problems with using them.

2.1.2.1. Judgmental samples

17. Judgmental sampling is a method that relies upon "experts" to choose the sample elements. Supporters claim that the method avoids the potential when random techniques are used of selecting a "bad" or odd sample, such as one in which all the sample elements unluckily fall in, say, the northwest region.

- ***Example***

An example of judgmental sampling applied to a household survey would be a group of experts who choose, purposively, the geographic districts to use as the first stage of selection in its sample plan – their decision being based on opinions of which districts are typical or representative in some sense or context.

18. The main difficulty with this type of sample is that what constitutes a representative set of districts is subjective and, ironically, highly dependent on the *choice* of experts themselves. With probability sampling, by contrast, the districts would first be stratified using, if necessary, whatever subjective criteria the design team wanted to impose; and then a random sample (selected in any of a variety of ways) of districts would be chosen *from each stratum*. Note that stratification decreases greatly the likelihood of selecting an odd sample such as the one alluded to above and this is the reason stratification was invented. With the stratified sample, every district would have a known, non-zero chance of selection that is unbiased and unaffected by subjective opinion. On the other hand, the judgmental sample affords neither the mechanism for ensuring that each district has a non-zero chance of inclusion nor of even calculating the probability of those that do happen to be selected.

2.1.2.2. Random walk and quota sampling

19. Another type of non-probability sampling that is widely used is the so-called "random walk" procedure at the last stage of a household survey, even if the prior stages of the sample were selected with legitimate probability methods. The illustration below shows a type of sampling that is a combination of random walk and quota sampling, another non-probability technique in which interviewers are given quotas of certain types of persons to interview.

- ***Example***

To illustrate the method, interviewers are instructed to begin the interview process at some geographic point in, say, a village, and follow a specified path of travel to select the households to interview. It may entail either selecting every n^{th} household or screening each one along the path of travel to ascertain the presence of a special target population such as children under 5 years old. In the latter instance each qualifying household would be interviewed for the survey until a pre-determined quota has been reached.

20. This methodology is often justified as a way to avoid the costly and time-consuming expense of listing all the households in the sample area - village or cluster or segment - as a prior stage before selecting the ones to be interviewed. It is also justified on the grounds that non-response is avoided since the interviewer continues beyond non-responding households until he/she obtains enough responding ones to fulfil the quota. Furthermore, its supporters claim the technique is unbiased as long as the starting point along the path of travel is determined randomly and that probabilities of selection can be properly calculated as the number of households selected divided by the total number in the village, assuming that the latter is either known or can be closely approximated.

21. Theoretically, given the conditions set forth in the last sentence above, a probability sample is thus attainable. In practice, however, it is dubious whether it is ever actually achieved. This is attributable to (a) interviewer behaviour and (b) the treatment of non-response households including those that are potentially non-response. It has been shown in countless studies that when interviewers are given control of sample selection in the field, biased samples result. For example, the average size (number of persons) of the sample households is usually smaller than the population of households.[1] It is basic human nature for an interviewer to avoid a household that may be perceived to be difficult in any way. For this reason, it is simpler to bypass a household with a threatening dog or one that is heavily gated and not easily accessible in favor of a household next door which does not present such problems.

22. By substituting non-responding households or those that might be with responding ones, the sample is biased toward cooperative, readily available households. Clearly there are differences in the characteristics of households depending on their willingness and availability to participate in the survey. With the quota sample approach, persons who are difficult to contact or unwilling to participate are more likely to be underrepresented than would be the case in a

[1] In many survey organizations it is now standard practice to make sure that the designation of the households to be selected for the sample is carried out as an office operation, where it is more easily controlled by supervision, by someone who either was not involved in creating the list of households prior to sample selection or is otherwise unfamiliar with the actual situation on the ground.

probability sample, since interviewers for the latter are generally required to make several callbacks to households where its members are temporarily unavailable and, moreover, to make extra efforts to convince reluctant households to agree to be interviewed.

2.1.2.3. Convenience samples

23. Convenience sampling is also widely used because of its simplicity of implementation. There are many examples of a convenience sample, though it is not applied often in household surveys. One example is conducting a survey of school youth in a purposively chosen sample of schools that are easily accessible and known to be cooperative, that is, convenient. Another that is currently in vogue is the instant poll that is administered on Internet sites, wherein persons who login are asked their opinions on various topics. It is perhaps obvious why samples of this type are inherently biased and should not be used to make inferences about the general population.

2.2. Sample size determination for household surveys

24. Considerable detail is devoted to this subsection because of the importance of sample size to the entire operation and cost of a survey. Not only is it important in terms of how many households are interviewed but how many geographic areas (*PSU*s – primary sampling units) are sampled, how many interviewers are hired, how big the workload if for each interviewer, etc. The factors and parameters that must be considered in determining the sample size are many but they revolve chiefly around the measurement objectives of the survey. We will discuss sample size determination in terms of the key estimates desired, target population, number of households that must be sampled to reach the requisite target populations, precision and confidence level wanted, estimation domains, whether measuring level or change, clustering effect, allowance for non-response and available budget. Clearly, sample size is the pivotal feature that governs the overall design of the sample.

2.2.1. Magnitudes of survey estimates

25. In household surveys, whether general-purpose or devoted to a certain topic such as health or economic activity, every estimate (often referred to as *indicator*) to be generated from the survey requires a different sample size for reliable measurement. The size of the sample depends on the size of the estimate, that is, its expected proportion of the total population. For example, to estimate, reliably, the proportion of households with access to safe water requires a different sample size than estimating the proportion of adults not currently working.

26. In practice the survey itself can have only one sample size. To calculate the sample size a choice must be made from among the many estimates to be measured in the survey. For example, if the key indicator is the unemployment rate, that would be the estimate upon which to calculate the sample size.[2] When there are many key indicators a convention sometimes used is to calculate the sample size needed for each and use the one that yields the largest sample.

[2] It is somewhat paradoxical that in order to calculate the sample size, its formula requires knowing the approximate value of the estimate to be measured. The value may be “guessed,” however, in various ways such as by using data from a census or similar survey, a neighboring country, a pilot survey and so forth.

Generally this is the indicator for which the base population is the smallest sub-target population, in terms of its proportion of the total population. The desired precision must of course be taken into account (precision is discussed further below). By basing the sample size on such an estimate, each of the other key estimates is therefore measured with the same or better reliability.

27. Alternatively, the sample size can be based on a comparatively small proportion of the target population instead of specifying a particular indicator. This may likely be the best approach in a general purpose household survey that focuses on several, unrelated subjects, in which case it may be impractical or imprudent to base the sample size on an indicator that pertains to a single subject. The survey managers may decide, therefore, to base the sample size on being able to measure, reliably, a characteristic held by 5 percent (or 10 percent) of the population – the exact choice dependent upon budget considerations.

2.2.2. Target population

28. Sample size depends also on the target population that the survey will cover. Like indicators, there are often several target populations in household surveys. A health survey, for example, may target (1) households to assess access to safe water and sanitation while targeting (2) all persons to estimate chronic and acute conditions, (3) women 14-49 for reproductive health indicators and (4) children under 5 for anthropometric measurements of height and weight.

29. Calculation of the sample size must therefore take into consideration each of the target populations. With multiple target populations – each of equal interest with respect to the measurement objectives of the survey – it is plausible, again, to focus on the smallest one in determining sample size. For example, if children under 5 years old are an important target group for the survey, the sample size should be based of that group. Utilizing the concept described in the preceding subsection, the survey management team may decide to calculate the sample size to estimate a characteristic held by 10 percent of children under 5. The resulting sample size would be considerably larger than that needed for a target group comprised of all persons or all households.

2.2.3. Precision and statistical confidence

30. Above it is suggested that the estimates, especially those for the key indicators, must be reliable. Sample size determination depends, critically, on the degree of precision wanted for the indicators. The more precise or reliable the survey estimates must be the greater the sample size must be – and by orders of magnitude. Doubling the reliability requirement, for example, necessitates *quadrupling* the sample size. Survey managers must obviously be cognizant of the impact that overly stringent precision requirements have on the sample size and hence the cost of the survey, while simultaneously being careful not to use such a small sample size that the main indicators would be too unreliable for informative analysis or meaningful planning.

31. Similarly, the sample size increases as the degree of statistical confidence wanted increases. The 95 percent confidence level is almost universally taken as the standard and the sample size necessary to achieve it is calculated accordingly. The confidence level describes the confidence interval around the survey estimate. For example, if a survey estimate is 15 percent

and its standard error is 1.0 percentage point the confidence interval at the 95-percent level is given as 15 percentage points plus or minus 2 percentage points (twice the standard error gives the 95 percent level of confidence), that is, ¦ 13 - 17¦ .

32. Taking account of the indicators, a convention that is used in many well-conceived surveys is to use, as the precision requirement, a margin of *relative* error of 10 percent at the 95 percent confidence level on the key indicators to be estimated. By this is meant that the standard error of a key indicator should be no greater than 5 percent of the estimate itself. This is calculated as (2 * 0.05x, where x is the survey estimate). For example, if the estimated proportion of persons in the labour force is 65 percent its standard error should be no larger than 3.25 percentage points, that is, 0.65 multiplied by .05. Two times .0325, or .065, is the relative margin of error at the 95 percent confidence level.

33. The sample size needed to achieve the criterion of 10 percent margin of relative error is thus one-fourth as big as one where the margin of relative error is set at 5 percent. A margin of relative error of 20 percent is generally regarded as the maximum allowable for important indicators because the confidence interval surrounding estimates with greater error tolerances are too wide to achieve meaningful results for most analytical or policy needs.

2.2.4. Analysis groups - domains

34. Another significant factor that has a large impact on the sample size is the number of domains, defined as the analytical sub-groups for which *equally* reliable data are wanted. The sample size is increased, approximately,[3] by a factor equal to the number of domains wanted, because sample size for a given precision level does not depend on the size of the population itself, except when it is a significant percentage, say, 5-percent or greater, of the population (rarely the case in household surveys). Thus, the sample size needed for a single province (if the survey were to be confined to only one province) would be the same as that needed for an entire country.

35. Thus, when only national-level data are wanted there is a single domain, and the sample size calculated thus applies to the sample over the entire country. If, however, it were decided that equally reliable results should be obtained for urban and rural areas, separately, then the calculated sample size must apply to each domain, which, in this case, would require doubling it. Moreover, if domains were defined as, say, the 5 major regions of a country, then 5 times the calculated sample size would be necessary, again if equally reliable data for each region takes precedence over the national estimate.

2.2.4.1. Over-sampling for domain estimates

36. An important implication of the equal reliability requirement for domains is that disproportionate sampling rates must be used. Thus, when the distribution is not 50-50, as will likely be the case for urban-rural domains, deliberate over-sampling of the urban sector will most likely be necessary in most countries to achieve equal sample sizes and thus equal reliability.

[3] This is the case whenever the same degree of reliability is wanted for each of the domains.

37. It is important to note that deliberate over-sampling of sub-groups, whether for domains or strata, not only necessitates the use of compensating survey weights to form the national-level estimates but the latter are somewhat less reliable than would be the case if the sample were distributed proportionately among the sub-groups.

2.2.4.2. Choosing domains

38. Geographic sub-areas are important of course and there is always pressure to treat them as domains for estimation purposes. For example, in a national-level survey, constituent users often want data not only for each major region but often for each province. Clearly, the number of domains has to be carefully considered and the type of estimation groups comprising those domains prudently chosen. The most plausible strategy is to decide which estimation groups, despite their importance, would not require equal reliability in the survey measurement. The estimation groups would be treated in the analysis as major tabulation categories, but the sample sizes for each one would be considerably less than if they were treated as domains; consequently, their reliability would be less as well.

- ***Example***

An example is shown of how the sampling would be done and what its effect on the reliability would be if urban-rural were treated as tabulation groups rather than domains. Suppose the population distribution is 60-percent rural and 40-percent urban. If the calculated sample size was determined to be 8000 households, then 16,000 would have to be sampled if urban and rural were separate domains – 8000 in each sector. Instead, by treating them as tabulation groups, the sample of 8000 households would be selected, proportionately, by rural and urban, yielding 4800 and 3200 households, respectively, in each of the two sectors. Suppose, further, the anticipated standard error for a 10-percent characteristic, based on the sample of 8000 households, is 0.7 percentage points. This is the standard error that applies to the national estimate (or to urban and rural separately if 8000 households were sampled in each domain). For a national sample of 8000 households selected proportionately by urban-rural, the corresponding standard error for rural would be approximately, 0.9, calculated as the square root of the ratio of sample sizes times the standard error of the national estimate, or ([v8000/4800]*0.7). For urban the standard error would be about 1.1, or ([v8000/3200]*0.7). Another way of evaluating the effect is that standard errors for all rural estimates would be about 29-percent higher (v8000/4800) than those for national estimates; for urban they would be about 58 percent higher, (v8000/3200).

39. Note that the last sentence of the example applies no matter what the standard error is at the national level. It is thus possible to analyze the impact on reliability for various sub-groups that might be considered as domains prior to sampling, in order to help decide whether they should be treated as tabulation groups, with its implications for proportionate rather than equal sampling, instead. For example, if a national survey is planned for a country that is only 20 percent urban the sample size in the urban area would be only 20 percent of the total sample size. Thus, sampling error for the urban estimates would be twice (square root of *0.8n/0.2n*) as big as those for the rural estimates and about two and a quarter times larger than the national estimates (square root of *n/0.2n*). In such case, survey managers might decide it is necessary to over-sample the urban sector, effectively making separate urban and rural domains.

40. Similarly, analysis of the relationship between standard errors and domains versus tabulation groups can be made to guide the decision-making process on whether to use regions or other sub-national geographic units as domains and if so, how many. With equal samples sizes necessary for domains, 10 regions requires 10 times the national sample size, but this is reduced by half if only 5 regions can be suitably identified to satisfy policy needs. Likewise, if regions are treated as tabulation groups instead, the national sample would be distributed proportionately among them. In that case, the *average* region would have standard errors approximately 3.2 times larger than the national estimates if there are 10 regions but only twice as large if there are 5.

2.2.5. Clustering effects

41. A more detailed discussion of cluster sampling is provided later in this chapter, but here we discuss how determination of the sample size is affected. The degree to which a household survey sample is *clustered* affects the reliability or precision of the estimates and therefore the sample size. Cluster effects in household surveys come about in several ways – (1) from the penultimate sampling units, generally referred to as the "clusters," which may be villages or city blocks, (2) from the sample households, (3) from the size and/or variability of the clusters and (4) from the method of sampling households within the selected clusters. Clustering as well as the effects of stratification are measured numerically by the so-called design effect, or *deff*, that expresses how much larger the sampling variance (square of the standard error) for the stratified, cluster sample is compared to a simple random sample of the same size. Stratification tends to improve the sampling variance to a small degree, so that *deff* indicates, primarily, how much clustering there is in the survey sample.

42. Efficient sample design requires that clusters be used to control costs but also that the design effect be kept as low as possible in order for the results to be usably reliable. Unfortunately, *deff* is not known before a survey is undertaken and can only be estimated afterwards, from the data themselves. Unless previous surveys have been conducted or similar ones in other countries so that proxy estimates of *deff* can be utilized, a default value of 1.5 to 2.0 for *deff* is typically used by the sampling practitioner in the formula for calculating the sample size.

43. To keep the design effect as low as possible, the sample design should follow these general principles:

a. Use as many clusters as is feasible.
b. Use the smallest cluster size in terms of number of households that is feasible.
c. Use a constant cluster size rather than a variable one.
d. Select a systematic sample of households at the last stage, geographically dispersed, rather than a segment of geographically contiguous households.

44. Thus, for a sample of 12,000 households it is better to select 600 clusters of 20 households each than 400 clusters of 30 households each, because *deff* is much lower in the former. Moreover, *deff* is reduced if the households are chosen systematically from all the households in the cluster rather than selecting them in contiguous geographic sub-segments.

When these rules-of-thumb are followed the design effect is likely to be reasonably low, and if a sample design is so based the default value of the design effect to use in calculating the sample size would tend more toward 1.5 than 2.0.

2.2.6. Adjusting sample size for anticipated non-response

45. It is common practice in surveys to increase the sample size by an amount equal to the anticipated non-response rate. This assures that the actual number of interviews completed in the survey will closely approximate the target sample size.

46. In many countries non-response is handled by substituting households that are available or willing to respond. In this case it would be superfluous to adjust the sample size to account for non-response since the targeted sample size is likely to be met anyhow. The substitution procedure, however, is biased in the same way as non-response cases are. There is nothing to be gained in *statistical accuracy* by substitution over what would be achieved by over-sampling to account for non-response in the first place. The substitution method does, however, achieve the targeted sample size, thus rendering it unnecessary to over-sample for non-response. In this way *sampling reliability* is the same for the two methods of dealing with non-response, because the ultimate sample size in terms of completed interviews is the same.

47. The degree of non-response in surveys varies widely by country. In the calculation exercise below, we allow the anticipated non-response rate to be 10 percent. Countries should of course use the figure that more accurately reflects their recent experience with national surveys.

2.2.7. Sample size for master samples

48. Master samples are discussed in detail in the next chapter but here we focus on the sample size for a master sample plan. Briefly, a master sample is a large sample of *PSU*s for countries that have major and continuing integrated survey programmes. The large sample is intended to provide enough “banked” sample cases to support multiple surveys over several years without having to interview the same respondents repeatedly.

49. With many surveys and hence many substantive subjects being accommodated by the master sample, there are of course numerous target populations and key estimates to be served. In that regard, most countries establish the sample size based on budgetary considerations and the anticipated sample sizes of the individual surveys that might be used over the time interval that the master sample will be used – often as long as ten years between population censuses. Thus, plausible sample sizes for master samples are very large, ranging as high as 50,000 households or even more. Plans for utilization of the entire bank of households are carefully formulated.

- ***Example***

Suppose the master sample in country A comprises 50,000 households. The master sample is intended to be used in three surveys that have already been planned, as well as, potentially, in two others not yet planned. One of the surveys is for household income and expenditures which is to be repeated three times during the decade – in year 1, year 5 and year 8. That survey is designed to survey 8000 households in each of the three years of its operation, but year 5 will have a replacement sample of 4000 households for half of the 8000 interviewed in year 1. Similarly, year 8 will replace the remaining 4000 households from year 1 with 4000 new ones. Thus, a total of 16,000 households will be used for the income and expenditure survey. The second survey being planned is a health survey in which it is expected about 10,000 households will be used, while the third survey on labour force participation will use about 12,000 households. Altogether, 38,000 households are reserved for these three surveys. Altogether, 12,000 households still remain that can be used for other surveys if necessary.

2.2.8. Estimating change or level

50. In surveys that are repeated periodically, a key measurement objective is to estimate changes that occur between surveys. In statistical terms, the survey estimate obtained on the first occasion provides the *level* for a given indicator, while the difference between that and the estimate of level on the second occasion is the estimated *change*. Estimating change generally requires a substantially larger sample size to draw reliable conclusions compared to that which is needed to estimate level only, especially when small changes are being measured. There are, however, certain sampling techniques that serve to reduce the sample size (and hence the cost) when estimating change, and these are discussed in the subsection on sampling to estimate change or trend.

2.2.9. Survey budget

51. Perhaps it goes without saying that the survey budget cannot be ignored when determining an appropriate sample size for a household survey. While the budget is not a parameter that figures in the mathematical calculation of sample size, it does figure prominently at a practical level.

52. When the sample size is calculated by the statistician, taking account of each of the parameters discussed in this chapter, it is often the case that it may be larger than the survey budget can support. When this occurs, the survey team must either seek additional funds for the survey or modify its measurement objectives. The objectives may be altered by reducing either the precision requirements or the number of domains.

2.2.10. Sample size calculation

53. In this subsection we provide the calculation formula for determining the sample size, taking account of the previously-discussed parameters. Because we are focusing on household surveys the sample size is calculated in terms of the number of households that must be selected. Illustrations are also given.

54. The estimation formula for the sample size, n_h, is

$$n_h = (z^2)\ (r)\ (1\text{-}r)\ (f)\ (k)/\ (p)\ (n)\ (e^2), \text{ where} \qquad [1]$$

n_h is the parameter to be calculated and is the sample size in terms of number of households to be selected;

z is the statistic that defines the level of confidence desired;

r is an estimate of a key indicator to be measured by the survey;

f is the sample design effect, *deff*, assumed to be 2.0 (default value);

k is a multiplier to account for the anticipated rate of non-response;

p is the proportion of the total population accounted for by the target population and upon which the parameter, r, is based;

n is the average household size (number of persons per household);

e is the margin of error to be attained.

55. Recommended values for some of the parameters are as follows:

The z-statistic to use should be 1.96 for the 95-percent level of confidence (as opposed to, say, 1.645, for the 90-percent level). The default value of f, the sample design effect, should be set at 2.0 unless there is supporting empirical data from previous or related surveys that suggest a different value. The non-response multiplier, k, should be chosen to reflect the country's own experience with non-response – typically under 10 percent in developing countries; a value of 1.1 for k, therefore, would be a conservative choice. The parameter, p, can usually be taken from the most recent census, although a reasonable rule of thumb is to use 0.03 for each year of age that the target population represents; for example, if the target population is children under 5 years old, p would be equal to 0.15 (5 * 0.03). The parameter, n, is often about 6.0 in most developing countries, but the exact value should be used - usually available from the latest census. For the margin of error, e, it is recommended to set the level of precision at 10 percent of r; thus $e = 0.10r$; a smaller sample size can be gotten with a less stringent margin of error, $e = 0.15r$, but the survey results would be much less reliable of course.

Substituting these recommended values gives

$$n_h = (3.84)\ (1\text{-}r)\ (1.2)\ (1.1)/\ (r)\ (p)\ (6)\ (.01). \qquad [2]$$

Formula [2] reduces further to

$$n_h = (84.5)\ (1\text{-}r)/\ (r)\ (p). \qquad [3]$$

- ***Example***

In country B it is decided the main survey indicator to measure is the unemployment rate which is thought to be about 10 percent of the civilian labour force. Civilian labour force is defined as the population 14 and older, which makes up about 65 percent of the country's total population. In this case, r = 0.1 and p = 0.65. Suppose we wish to estimate the unemployment rate with 10-percent margin of relative error at the 95-percent level of confidence; then e = 0.10r (that is, 0.01 standard error) as recommended above. Furthermore, the values for the expected non-response rate, design effect and average household size are the ones we have recommended. Then we can use formula [3], which yields 1170 households (84.5*0.9)/(0.1*0.65). This is a fairly small sample size, primarily because the base population comprises such a large proportion of the total, that is, 65 percent. Recall that the sample size calculated is for a single domain – in this case the national level. If the measurement objectives include obtaining equally reliable data for urban and rural areas the sample size would be doubled assuming all the parameters of formulas [2] and [3] pertain for both urban and rural. To the extent they differ (for example, the average household size may be different between urban and rural households as might the expected non-response rate), the more accurate values should be used to calculate the sample size for urban and for rural separately. The results would be different of course.

56. The next example is for a smaller base population – children under 5 years old.

- ***Example***

In country C the main survey indicator is determined to be the mortality rate among children under 5 years old, thought to be about 5 percentage points. In this case, r = 0.05 and p is estimated to be about 0.15, or 0.03*5. Again we wish to estimate the mortality rate with 10-percent margin of relative error; then e = 0.10r (or 0.005 standard error) and the values for the expected non-response rate, design effect and average household size are again the ones we have recommended. Then formula [3] gives nearly 10,800 households (84.5*0.95)/(0.05*0.15), a sample so large that most countries do not seek to measure child mortality rates as the main indicator in a survey, with its concomitant requirement for fairly precise estimation. Again, the primary reason for this is related to the size of the base population, that is, children under 5, who comprise only 15 percent of the total. The estimated parameter, r, is also small and that together with a small p combines to force a large size sample.

57. The final example is for a case where the total population is the target population. In that case, p = 1 and can be ignored, but still formulas [2] or [3] may be used if the recommended values of the parameters are utilized.

- ***Example***

In country D the main survey indicator is determined to be the proportion of persons in the entire population that suffered an acute health condition during the preceding week. That proportion is thought to be between 5 and 10 percent, in which case the smaller value is used because it will give a larger sample size – the conservative approach. In this case, r = 0.05 and p is of course 1.0. Again we wish to estimate the malnutrition rate with 10-percent margin of relative error; then e = 0.10r[4] (or 0.005 standard error) and the values for the expected non-response rate,

[4] Since r applies to the entire population in this case, it is equal to p, so that e also equals 0.10p.

design effect and average household size are again the ones we have recommended. Then formula [3] gives a little over 1600 households (84.5*0.95)/(0.05).

58. As mentioned earlier the sample size for the survey may ultimately be determined by calculating sample sizes for several key indicators and basing the decision on the one which gives the largest sample size. In addition, the number of survey domains must also be considered as well as the survey budget before reaching a final determination.

59. For countries in which one or more of the assumptions discussed above do not hold, simple substitutions may easily be made in formula [1] to arrive at more accurate figures on sample size. For example, the average household size may be larger or smaller than 6.0; non-response may be expected around 5 percent instead of 10; and the value of p for a particular country can generally be more precisely computed from census figures than by using the convention of multiplying 0.03 times the number of years of age a target population comprises.

60. It is recommended however that no change be made for the z-statistic value of 1.96, which is the international standard. The design effect, f, should also be left at 2.0 unless recent survey data from another source suggests otherwise, as already mentioned. It is also recommended that e be defined as $0.10r$ except in cases where budgets cannot sustain the sample size that results. In that case it might be increased to $0.12r$ or $.15r$. Such increases in the margin of error will yield much higher sampling errors however.

2.3. Stratification

61. In designing a household survey, stratification of the population to be surveyed prior to sample selection is a commonly used technique. It serves to classify the population into sub-populations – strata – on the basis of auxiliary information that is known about the full population. Sample elements are then selected, independently, from each stratum in a manner consistent with the measurement objectives of the survey.

2.3.1. Use of stratified samples

62. With stratified sampling the sample sizes within each stratum are controlled by the sampling technician rather than by random determination through the sampling process. A population divided into strata can have exactly n_s units sampled from each, where n_s is the desired number of sample units in the s^{th} stratum. By contrast, a non-stratified sample would yield a sample size for the s^{th} sub-population that varies somewhat from n_s.

- ***Example***

Suppose the sample design of a survey is to consist of two strata – urban and rural. Information from the population census is available to classify all the geographic administrative units into either urban or rural, thus allowing the population to be stratified by this criterion. It is decided to select a proportionate sample in each stratum (as opposed to disproportionate) because the population is distributed 60-percent rural and 40-percent urban. If the sample size is 5000 households, independent selection of the sample by stratum will ensure that 3000 of them will be rural and 2000 urban. If the sample were selected randomly without first setting up strata the distribution of households in the sample would vary from the 3000-2000 split, although that would be its expected distribution. The non-stratified sample could, by unlucky chance, produce a sample of, say, 3200 rural households and 2800 urban ones.

63. Thus one reason for stratification is to reduce the chance of being unlucky and having a disproportionately large (or small) number of the sample units selected from a sub-population that is considered significant for the analysis. Stratification is done to ensure proper representation of important sub-population groups without biasing the selection operation. It is important to note, however, that proper representation does not imply proportionate sampling. In many applications one or more of the strata may also be estimation domains (discussed above), in which case it might be necessary to select equal-sized samples in the affected strata, thus producing a disproportionate sample by stratum. Hence, both proportionate and disproportionate *allocation* of the sample units among the strata are legitimate design features of a stratified sample, and the choice depends on the measurement objectives of the survey.

2.3.2. Rules of stratification

64. There are two basic rules when stratifying a population. One of the rules is required while the other should be observed, although little damage is done to the sample design in its breach. The required rule is that at least one sample unit must be selected from each stratum that is created. The strata are, essentially, independent and mutually exclusive subsets of the population; every element of a population must be in one and only one stratum. Because of this characteristic, each stratum *must* be sampled in order to represent the whole population and calculate an unbiased estimate of the population mean. Since each stratum can theoretically be treated independently in the sample design, creation of the strata need not be done using objective criteria; if desired, subjective criteria may be used as well.

65. The second rule for stratification is that each stratum created should, ideally, be as different as possible. Heterogeneity *among* strata with homogeneity *within* strata is thus the primary feature that should guide the establishment of strata. It can easily be seen from this feature why urban and rural areas are often established as two of the strata for a household survey. Urban and rural populations are different from each other in many ways (type of employment, source and amount of income, average household size, fertility rates, etc.) while being similar within their respective sub-groups.

66. The heterogeneity feature is a useful guide in determining how many strata should be created. There should be no more strata than there are identifiable sub-populations for the particular criterion being used to define strata. For example, if a country is divided into 8 geographic regions for administrative purposes but two of the regions are very much alike with respect to the subject-matter of a proposed survey, an appropriate sample design could be accomplished by creating 7 strata (combining the two similar regions). Nothing is gained by using, for example, 20 strata if 10 can accomplish the same heterogeneous sub-groupings.

67. An important point to note about *proportionate* stratification is that the resulting sample is at least as precise as a simple random sample of the same size. Thus, stratification provides gains in precision, or reliability, of the survey estimates and the gains are greatest when the strata

are maximally heterogeneous. This feature of stratified sampling is the reason that even poor stratification[5] does not damage the survey estimates in terms of their reliability.

68. Another important point to note is that while a single unit selected from each stratum suffices to meet the theoretical requirements of stratified sampling, a minimum of two must be chosen if the sample is to be used to calculate sampling errors for the survey estimates.

2.3.3. Implicit stratification

69. As mentioned, the choice of information available to create strata is determined by the measurement objectives of the survey. For household surveys that are large-scale and multipurpose in content a particularly useful method is so-called *implicit* stratification. Its essential criterion is geographic, which generally suffices to spread the sample properly among the important sub-groups of the population, such as urban-rural, administrative regions, ethnic subpopulations, socio-economic groups, etc.

70. To be applied correctly, implicit stratification requires using systematic selection at the first stage of sampling. The procedure is simple to implement and entails (1) arranging the file of *PSU*s in geographic sequence such as urban by province and within province by district, followed by rural by province and within province by district and (2) systematically selecting *PSU*s from the sorted file. The systematic selection is done either by equal probability (*epsem*) or probability proportionate to size (*pps* sampling). *PPS* sampling is discussed further in section 2.6 below.

71. An important advantage of implicit stratification is that it eliminates the need to establish explicit geographic strata. That, in turn, does away with the need to allocate the sample to those strata, especially when proportionate sampling is used. Another advantage is simplicity since the method only requires file sorting and application of the sampling interval(s). Disproportionate sampling may also be easily applied at the first level of the geographic sort. For example, if urban-rural is the first level, applying different sampling rates to the urban and rural portions is a straightforward operation.

72. An illustration of an implicit stratification scheme with systematic *pps* sampling is shown in Illustration 1 below.

2.4. Cluster sampling

73. The term, cluster sampling, was coined originally to refer to sample designs in which all members of a group were sampled and the groups themselves were defined as the clusters. For example, a sample of schools might be selected at the first stage and classrooms at the second; if all members of each classroom are surveyed then we would have a cluster sample of classrooms. In household surveys an example of the original notion of cluster sampling would be the selection of city blocks in which all the residents of the block would be interviewed for the survey. In recent years, however, cluster sampling has been broadly used to refer more generally

[5] Poor stratification can occur when strata are unnecessarily created or when some of the population elements are miss-classified into the wrong strata.

to surveys in which there is a penultimate stage of sampling that selects (and defines) the clusters, such as villages, census enumeration areas or city blocks; the final stage of sampling consists of a sub-sample of the households within each selected cluster (as opposed to surveying all of them). The latter use of the term is generally employed in this handbook.

Illustration 1. **Arrangement of Administrative Areas for Implicit Stratification**

- Urban
    - Province 01
        - District 01
            - EA 001
            - EA 002
            - EA 003
            - EA 004
        - District 02
            - EA 005
            - EA 006
            - EA 007
    - Province 02
        - District 01
            - EA 008
            - EA 009
        - District 02
            - EA 010
            - EA 011
            - EA 012
    - Province 03, *etc.*
- Rural
    - Province 01
        - District 01
            - EA 101
            - EA 102
            - EA 103
            - EA 104
        - District 02
            - EA 105
            - EA 106
            - EA 107
    - Province 02
        - District 01
            - EA 108
            - EA 109
            - EA 110
            - EA 111
        - District 02
            - EA 112
            - EA 113
            - EA 114
    - Province 03, *etc.*

74. In household surveys the sample design will invariably utilize some form of cluster sampling, of necessity, in order to ensure that the survey costs are contained. As mentioned earlier, it is much cheaper to carry out a survey of, say, 1000 households in 50 locations (20 households per cluster) than 1000 selected randomly throughout the population. Clustering of

the sample, unfortunately, decreases the reliability of the sample due the likelihood that people living in the same cluster tend to be homogeneous or to have the same characteristics. This so-called clustering effect has to be compensated in the sample design by increasing the sample size commensurately.

2.4.1. Cluster versus stratified sampling

75. Cluster sampling differs importantly from stratified sampling in two ways.[6] For the latter all strata are represented in the sample, since a sample of units is selected from each stratum. In cluster sampling a selection of the clusters themselves is made; thus, those that are in sample represent those that are not. That distinctive difference between stratified and cluster sampling leads to the second way in which they differ. As previously mentioned, strata should be created, ideally, to be internally homogeneous but externally heterogeneous with respect to the survey variables to be measured. The opposite is true for clusters. It is more advantageous in terms of sample precision for clusters to be as internally heterogeneous as possible.

76. In household surveys, clusters are virtually always defined as geographical units such as villages or parts of villages, which means, unfortunately, that heterogeneity within the cluster is not generally achieved to a high degree. Indeed, geographically-defined clusters are more likely to be internally homogeneous than heterogeneous with respect to such variables as type of employment (farmers for example), income level and so forth. The degree to which clusters are homogeneous, for a given variable, thus determines how “clustered” a sample is said to be. The more clustering there is in the sample the less reliable it is.

2.4.2. Cluster design effect

77. The clustering effect of a sample is partially measured by the design effect, *deff*, previously mentioned, although *deff* also reflects the effects of stratification. It is incumbent upon the sample design team to ensure that the sample plan achieves an optimum balance that seeks to minimize costs while maximizing precision. The latter is achieved by minimizing or controlling the design effect as much as possible. It is useful to look at the mathematical definition of the clustering component of *deff* in order to then see how it may be minimized or controlled.

$$deff = f \,\tilde{}\, 1 - d\,(\tilde{n}-1), \text{ where} \qquad [4]$$

f is the shortened symbol for *deff*,

d is the intra-cluster correlation, that is, the degree to which two units in a cluster are likely to have the same value compared to two at random in the population,

$\tilde{n}$ is the number of units of the target population in the cluster.

[6] It is important to note that stratification and cluster sampling are not competing alternatives in sample design, because both are invariably used in household survey sampling.

78. Formula [4] is not strictly the expression for *deff* because stratification is ignored as well as another factor that pertains when the clusters are not uniform in size. Still, the clustering component is the predominant factor in *deff* and that is why we can use it as an approximate form, which will serve to show how clustering affects sample design and what might be done to control it.

79. In the expression above it can be seen that *deff* is a multiplicative function of two variables, the intra-cluster correlation, d, and the size of the cluster, *ñ*. Thus, *deff*, increases as both d and *ñ* increase. While the sampler can exercise no control over the intra-cluster correlation for whatever variable is under consideration, he/she can adjust the cluster size up or down in designing the sample and thus control the design effect to a large extent.

- ***Example***

Suppose a population has an intra-cluster correlation of 0.03, which is fairly small, for chronic health conditions. Suppose further the sample planners are debating whether to use clusters of 10 households or 20, with an overall sample size of 5000 households. Suppose further that all households are the same size, 5 persons, just to simplify the illustration. The value of *ñ* is then 50 for 10 households and 100 for 20 households. Simple substitution in expression [4] shows an approximate value of *deff* of (1 + 0.03[49]), or 2.5, for the 10-household cluster design but 4.0 for the 20-household design. Thus the design effect is roughly 60 percent greater for the larger cluster size. The survey team would then have to decide whether it is better to sample twice as many clusters (500) by using the 10-household option in order to keep the reliability within a more acceptable level or to choose the cheaper option of 250 households at the expense of increasing the sampling variance dramatically. Of course other options between 10 and 20 households may be considered also.

80. There are several ways of interpreting the design effect. One is that it is the factor by which the sampling variance of the actual sample design (to be) used in the survey is greater than that of a simple random sample, *SRS*, of the same size. Another is simply how much worse the actual sample plan is over that of the *SRS* in terms of its precision. A third interpretation is how many more sample cases would have to be selected in the planned sample design compared to a *SRS* in order to achieve the same level of sampling variance. For example, *deff* of 2.0 means twice as many cases would have to be selected to achieve the same reliability that a *SRS* would produce. Clearly, therefore, it is undesirable to have a sample plan with *deff*s much larger than 2.5-3.0 for the key indicators.

2.4.3. Cluster size

81. It was noted that the sampler cannot control the correlations. Moreover, for most survey variables there is little if any empirical research that has attempted to estimate the value of those correlations. The intra-cluster correlation can vary, theoretically, between -1 to +1, although it is difficult to conceive of many household variables where it is negative. The only possibility the sampler has, therefore, in keeping *deff* to a minimum is to urge that cluster sizes be as small as the budget can allow. Table 1 displays *deffs* for varying values of the intra-cluster correlation and a constant cluster size.

Table 1. Comparison of Clustering Component of Design Effect for Varying Intra-cluster correlations, d, and cluster sizes, *ñ*

ñ	**d**						
	.02	**.05**	**.10**	**.15**	**.20**	**.35**	**.50**
5	1.08	1.2	1.4	1.6	1.8	2.4	3
10	1.18	1.45	1.9	2.35	2.8	4.15	5.5
20	1.38	1.95	2.9	3.85	4.8	6.65	10.5
30	1.58	2.45	3.9	5.35	6.8	11.15	15.5
50	1.98	3.45	5.9	8.35	10.8	18.15	25.5
75	2.48	4.7	8.4	12.1	15.8	26.9	38

82. From Table 1 it is clearly seen that cluster sizes above 20 will give unacceptable *deff*s (greater than 3.0) unless the intra-cluster correlation is quite small. In evaluating the numbers in the table it is important to remember that *ñ* refers to the number of units in the target population, not the number of households. In that respect the value of *ñ* to use is equal to the number of households in the cluster multiplied by the average number of persons in the target group. If the target group, for example, is women 14-49, there is typically about one per household for this group, in which case a cluster size of *b* households will have approximately that same number of women 14-49. In other words *ñ* and *b* are roughly equal for that target group and Table 1 applies as it stands. Following is an example when the number of households and target population in the cluster are not equal.

- ***Example***

Suppose the target population is all persons, which would be the case in a health survey to estimate acute and chronic conditions. Suppose, further, the survey is intended to use clusters of 10 households. The value of *ñ* in that case is 10 times the average household size; if the latter is 5.0 then *ñ* is 50. Thus, 50 is the value of *ñ* that must be viewed in Table 1 to assess its potential *deff*. Table 1 reveals that *deff* is very large except when d is about 0.02. This suggests that a cluster sample designed to use as little as 10 households per cluster would give very unreliable results for a characteristic such as contagious conditions, since the latter would likely have a large d.

83. The example illustrates why it is so important to take into account the cluster size when designing a household survey, particularly for the key indicators to be measured. Moreover, it must be kept in mind that the stated cluster size, in describing the sample design, will generally refer to the number of households, while the cluster size for purposes of assessing design effects must consider, instead, the target population(s).

2.4.4. Calculating *deff*

84. Actual *deff*s, for survey variables that the analysts specify, can be calculated after the survey has been completed. It requires estimating the sampling variance for the chosen variables (methods are discussed in chapter 4) and then computing, for each variable, the ratio of its variance to that of simple random sample of the same overall sample size. This calculation is an estimate of the “full” *deff* including stratification effects as well as variability in cluster sizes, rather than only the clustering component.

85. The square root of the ratio of variances gives the ratio of standard errors, or *deff* as it is called, and this is often calculated in practice and presented in the technical documentation of a survey such as the Demographic and Health Surveys (DHS).

2.4.5. Number of clusters

86. It is important to bear in mind that the size of the cluster is significant beyond its effect on sampling precision. It also matters in relation to the overall sample size, because it determines the number of different locations that must be visited in the survey. That of course affects survey costs significantly, which is why cluster samples are used in the first place. Thus, a 10,000-household sample with clusters of 10 households each will require 1000 clusters, while 20-household clusters will require only 500. As emphasized previously, it is crucial that both the costs and precision be taken into account to reach a decision on this feature of the sample design.

2.5. Sampling in stages

87. On a theoretical level the perfect household survey sample plan is to select the sample, *n*, of households randomly from among appropriately identified strata encompassing the entire population, *N*, of households. The stratified random sample so obtained would provide maximum precision. Practically, however, a sample of this type is far too expensive to undertake feasibly,[7] as we have previously noted in the discussion of the cost benefits that cluster sampling affords.

2.5.1. Benefits of sampling in stages

88. Selecting the sample in *stages* has practical benefits in the selection process itself. The process permits the sampler to isolate, in successive steps, the geographic locations where the survey operations, including listing households and administering interviews, will take place. When listing must be carried because of an obsolete sampling frame, a stage of selection can be introduced to limit the size of the area to be listed.

89. With cluster sampling there is a minimum of two stages to the selection procedure – first, selection of the clusters and second, selection of the households. The clusters in household surveys are always defined as geographical units of some kind. If those units are sufficiently small, both geographically and in population, and a current, complete and accurate list of them is available from which to sample, then two stages can suffice for the sample plan. If the smallest geographical unit available is too large to be efficiently used as a cluster, three stages of selection would be necessary.

- ***Example***

Suppose a country wishes to define its clusters as census enumeration areas, or *EA*s, because this is the smallest geographical unit that exists administratively. It is complete because the entire country is divided into *EA*s, accurate because every household lives, by definition, in one and only one *EA* and current in the sense that it is based on the most recent census and there have

[7] There are one or two exceptions and they would be for countries that are very small geographically, such as Kuwait or Singapore, where a random sample of households would incur very little travel costs.

been no changes after the census in defining *EA*s. Suppose, further, the census is two years old and so it is determined that it will be necessary to compile a current list of households in the sample *EA*s rather than using the two-year old census list of households. The average size of an *EA* is 200 households, yet the desired cluster size for interviewing is intended to be 15 households per cluster. The survey team calculates that the cost of listing 200 households for every 15 that are ultimately sampled (ratio of over 13 to 1) is too great an expense. Instead, the sampler decides to implement a cheaper field operation by which each sample *EA* is divided into quadrants of approximately equal size of about 50 households each. The sample plan is then modified to select one quadrant, or *segment*, from each sample *EA* in which to conduct the listing operation, thus reducing it by three-fourths. In this design we have three stages – first stage selection of *EA*s, second-stage selection of *EA*-segments and third-stage selection of households.

2.5.2. Use of dummy stages

90. Often, so-called "dummy" stages are used in sample selection as a method to avoid having to sample from an enormous file of units in the penultimate stage. The file may contain so many units and be so unwieldy that it cannot be realistically managed through tedious, manual selection. Even if the file is computerized it may still be so large that it cannot be managed efficiently for sample selection. Dummy stages allow one to narrow the sub-universes to more manageable numbers by taking advantage of the hierarchical nature of administrative subdivisions of a country.

91. For rural surveys in Bangladesh, for example, villages are often designated as the next-to-last stage of selection. There are more than 100,000 villages in Bangladesh, which is far too many to manage efficiently for sample selection. If a sample plan is designed to select 600 villages at the penultimate stage, for example, only 1 in about 167 would be selected in Bangladesh. To cut down on the size of the files for sample selection it might be decided to select the sample in stages using the hierarchy of geographical units in which Bangladesh is divided – thanas, unions and villages. The sample selection would proceed in steps by first selecting 600 thanas, using probability proportionate to their sizes (this method is discussed in detail in section 2.6). Next, exactly one union would be selected from each sample thana, again using probability proportionate to size; thus there are 600 unions in the sample. Third, one village would be selected *pps* from each sample union, again resulting in 600 villages. Finally, the sample of households would be selected from each sample village; this would generally be a systematic sample of all the households in each sample village.

92. The sample selection methodology described above is in effect a two-stage sample of villages and households, although two dummy stages were initially utilized to select the thanas and unions from which the villages are selected. To illustrate the dummy character of the first two stages, it is useful to examine the probabilities at each selection stage and the overall probability.

First stage of selection: thanas

93. Thanas are selected with probability proportionate to size, *pps* (illustration of *pps* sampling is in section 2.6). The probability at that stage is given by:

$P_1 = (a\ m_t)/?\,m_t$, where [5]

P_1 is the probability of selecting a given thana,

a is the number of thanas to be selected (600 in this illustration),

m_t is the number of rural households[8] in the t^{th} thana according to the sampling frame used (for example, the most recent population census).

94. The factor, $?\,m_t$, is the total number of rural households over all thanas in the country. It should be noted that the actual number of thanas selected may be less than 600. This can occur whenever one or more thanas is selected twice, which is a possibility for any thana for which its measure of size exceeds the sampling interval. The sampling interval for selecting thanas is given by $?\,m_t \div a$. Thus, if the sampling interval is, say, 12500 and the thana contains 13800 households it will automatically be selected once and have a chance of 1300/12500 of being selected twice (the numerator is equal to 13800-12500).

Second stage of selection: unions

95. At the second stage, one union is selected from each sample thana, again with *pps*. Practically, this is accomplished by listing all the unions in the selected thana, cumulating their measures of size, m_u, and choosing a random number between 1 and m_t, the measure of size for the sample thana. The cumulant whose value which is the smallest number equal to or greater than the random number identifies the selected union; or, an equivalent convention is used to identify the selected union. If a thana was selected more than once in the first stage the same number of unions would then be selected from it. The probability at the second stage is given by:

$P_2 = (1)\ (m_u)/m_t$, where [6]

P_2 is the probability of selecting a given union in the sample thana,

(1) signifies that only one union is selected,

m_u is the number of households in the u^{th} union according to the frame.

Third stage of selection: villages

96. At the third stage, one village is selected from each sample union with *pps*. The probability at the third stage is given as:

$P_3 = (1)\ (m_v)/m_u$, where [7]

P_3 is the probability of selecting a given village in the sample union,

[8] This is the measure of size and it may, instead, be the population of the thana, provided the number used is consistent for all measures of size at every stage.

(1) signifies that only one village is selected,

m_v is the number of households in the v^{th} village according to the frame.

Fourth stage of selection: households

97. At the fourth stage we will assume that the frame list of households is available for each selected village, so that the sample of households can be systematically selected from those lists. A fixed number of households is selected from each sample village – that number being the pre-determined cluster size. The probability at the fourth stage is given as:

$P4 = (b)/mv$, where [8]

P_4 is the probability of selecting a given household in the sample village and

b is the fixed number of households selected in each village.

Overall probability of selection

98. The overall probability is the product of probabilities at each stage, as follows:

$$P = P_1\, P_2\, P_3\, P_4.$$

Substituting, we have:

$$\begin{aligned} P &= [(a\, m_t)/?\, m_t]\; [(1)\, (m_u)/m_t]\; [(1)\, (m_v)/m_u]\; [b/m_v], \\ &= [(a)\, (b)]/\; ?\, m_t. \qquad [9] \end{aligned}$$

99. Note that P_2 and P_3 cancel out completely, demonstrating the dummy nature of the "four" stage selection process. Thus the thanas and unions, though physically "selected," nevertheless serve merely to pin down where the sample villages are located.

2.5.3. The two-stage design

100. Recently, much attention has been given to the use of two-stage sample designs in developing countries. It is the sample design of choice for the Multiple Indicator Cluster Surveys (MICS) that UNICEF has carried out in over 100 countries since the mid-1990s. They are also used predominantly in the Demographic and Health Surveys (DHS).

101. Typically the two-stage design consists, simply, of a *pps* sample of several hundred geographical units at the first stage. A current listing of households may be developed in the first-stage sample units, depending upon the availability of information regarding the address and/or location of the households and whether that information is current. This is followed by a systematic sample of a fixed number of households at the second stage. The geographical units, commonly referred to as the "clusters" are usually defined as villages or census enumeration areas (*EA*s) in rural areas and city blocks in urban areas.

102. The two-stage design described above is appealing in many ways but chiefly because of its simplicity. It is always advantageous in sample design to strive more toward simplicity than complexity in order to reduce the potential for nonsampling error in sample implementation. Some of the useful features of the two-stage design that make it comparatively simple and desirable are as follows:

- As described, the sample design is self-weighting (all the households in the sample are selected with the same probability) or approximately self-weighting – see the following two subsections for the distinction between samples selected *pps* versus *ppes* (probability proportionate to estimated size). There are advantages with self-weighting designs both in terms of sampling reliability and ease of implementation at the data analysis phase.
- Clusters defined in terms of *EA*s or city blocks are a convenient size – not too big - in most countries, especially if a fresh listing of households must be made before the final stage of selection.
- *EA*s, city blocks and most villages are usually mapped, either for census operations or other purposes, with well-delineated boundaries.

2.6. Sampling with probability proportional to size (PPS)

103. In section 2.5 an illustration was presented in which sampling with probability proportionate to size featured prominently in the selection of the clusters for the sample. This section discusses *pps* sampling in greater detail.

2.6.1. PPS sampling

104. Use of *pps* sampling permits the sampler to exercise greater control over the ultimate sample size in cluster surveys. In situations where the clusters are all the same size, or approximately so, there would no advantage to using *pps* sampling. For example, if every block in a particular city contained exactly 100 households and one wanted a sample of 1000 households spread among a sample of 50 city blocks, the obvious sample plan would be to select a *SRS* sample of 50 blocks, that is, with equal probability or *epsem*, and then systematically select exactly 1 in 5 of the households from each block – also an *epsem* sample. The result would be a sample of precisely 20 households per block or 1000 altogether. The selection equation in this case is expressed as

$P = [50/M]\ [1/5]$, where

P is the probability of selecting a household,

$[50/M]$ is the probability of selecting a block,
M is the total number of blocks in the city and

$[1/5]$ is the probability of selecting a household within a given sample block.

105. P reduces to $10/M$. Since M is a constant, the overall probability of selection for each sample household is equal to 10 divided by the number of blocks, M.

106. In real situations, however, blocks or other geographical units that might be used as clusters for household surveys are seldom so unvarying in their sizes. For the example above, they may range in size, say, from 25 to 200. An *epsem* sample of blocks could result in an “unlucky” selection of mostly small ones or mostly large ones, with the result that the overall sample size could be drastically different from the desired 1000 households discussed in the example. One method of reducing the potential for widely variable sample sizes is to create strata based on the size of the clusters and select a sample from each stratum, but that method may reduce or complicate the use of other stratification factors in the sample design. *PPS* sampling permits greater control over the ultimate sample size without the need for stratification by size.

107. To illustrate *pps* sampling we start with the selection equation, mentioned above, but expressed more formally for a two-stage design[9] as follows:

$P(aß) = P(a)\ P(ß¦\ a)$, where [10]

$P(aß)$ is the probability of selecting household ß in cluster a,

$P(a)$ is the probability of selecting cluster a,

$P(ß¦\ a)$ is the conditional probability of selecting household ß in the second stage given that cluster a was selected at the first stage.

108. To fix the overall sample size in terms of number of households we want an *epsem* sample of n households out of the population of N households. Thus, the overall sampling rate is n/N which is equal to P(aß). Further, if the number of clusters to be sampled is specified as a then ideally we need to select b households from each cluster no matter the sizes of the selected clusters. If we define m_i as the size of the i^{th} cluster, then, we need P(ß¦ a) to be equal to b/m_i. Then,

$$P(aß) = [P(a)]\ [b/m_i].$$

Since $n = ab$, we have

$$ab/N = [P(a)]\ [b/m_i].$$

Solving the latter equation for $P(a)$, we get

$P(a) = (a)(m_i)/N.$ [11]

109. Note that $N = S\ m_i$, so that the probability of selecting a cluster is therefore proportional to its size. The selection equation for a *pps* sample of first-stage units in which the ultimate units are, nevertheless, selected with equal probability is therefore

$P(aß) = [(a)(m_i)/\ S\ m_i,]\ [b/m_i]$ [12]

$= [(ab)/\ S\ m_i].$ [13]

[9] See (Kalton, 1984, 38-47) for development of this notation and additional discussion on *pps* sampling.

110. The sample design so achieved is self-weighting, as can be seen from [13], because all the terms of the equation are constants; recall that while m_i is a variable the summation, S m_i, is a constant equal to N.

Below is an illustration of how to select a sample of clusters using *pps*.

Illustration 2. ***Illustration of Systematic pps Selection of Clusters***

Cluster/PSU No.	Measure of Size (Number of HHs)	Cumulative	Sample Selection
001	215	215	
002	73	288	
003	89	377	311.2
004	231	608	
005	120	728	
006	58	786	
007	99	885	878.8
008	165	1050	
009	195	1245	
010	202	1447	1446.4
011	77	1524	
012	59	1583	
013	245	1828	
014	171	1999	
015	99	2098	2014.0
016	88	2186	
017	124	2310	
018	78	2388	
019	89	2477	
020	60	2537	
021	222	2759	2581.6
022	137	2896	
023	199	3095	
024	210	3305	3149.2
025	165	3470	
026	274	3744	3716.8
027	209	3953	
028	230	4183	
029	67	4250	
030	72	4322	4284.4
031	108	4430	
032	111	4541	

SAMPLE INSTRUCTIONS: Select 8 PSUs (clusters) from 32 in the universe using *pps*; Selection Interval (*I*) therefore equals 4541/8, OR 567.6, where 4541 is total cumulated measures of size for all clusters and 8 is the number of clusters to select; Random Start (*RS*) is random number between 0.1 and 567.6 chosen from a random number table; in this illustration, RS = 311.2.

111. To physically select the sample, note that in the illustration above the sampling interval, *I*, is successively added to the random start, *RS*, seven times (or *a*-1 times, where *a* is the number of clusters to be selected). The resulting selection numbers are 311.2 (which is *RS*), 878.8, 1446.4, 2014, 2581.6, 3149.2, 3716.8 and 4284.4. The cluster that is sampled for these 8 selection numbers is, in each case, the one whose cumulated measure of size is the smallest value

equal to or greater than the selection number. Thus cluster 03 is selected because 377 is the smallest cumulant equal to or greater than 311.2 and cluster 26 is selected because 3744 is the smallest cumulant equal to or greater than 3716.8.

112. Although the illustration does not conclusively demonstrate it (because only 8 clusters were selected), *pps* sampling tends to select larger rather than smaller clusters. This is perhaps obvious since from formula [11] it is seen that the probability of selecting a cluster is proportionate to its size; thus a cluster containing 200 households is twice as likely to be selected as one containing 100 households. Because of this, it should be noted that the same cluster may be selected more than once if its measure of size exceeds the sampling interval, *I*. None of the clusters in the illustration, however, fit that condition. If it should happen, however, the number of households to select in such a cluster is double for two "hits," triple for three "hits" and so forth.

2.6.2. PPES sampling (probability proportionate to estimated size)

113. The *pps* sampling methodology described in the previous subsection is somewhat ideal and may not be realizable in practical applications in most cases. That is because the measure of size used to establish the probability of selection of the cluster, at the first stage, is often not the *actual* measure of size when the sample of households is selected at the second stage.

114. In household surveys the measure of size generally adopted for the first stage selection of *PSU*s or clusters is the count of households (or population) from the most recent census. Even if the census is very recent, the actual number of households at the time of the survey is likely to be different, if only by a small amount. There is an exception, however, and that is when the second-stage selection of households is taken directly from the same frame as the one used to establish the measures of size (see more discussion about sampling frames in the next chapter).

- ***Example***

Suppose a household survey is taken 3 months after the conclusion of the population census. The survey team decides to use the census list of households at the second stage of a two-stage sample. This is in lieu of making a fresh listing of households in the selected clusters because it is plausibly assumed that the census list is, for all practical purposes, current and accurate. At the first stage a sample of villages is selected using the census count of households as the measure of size for the village. For each sample village the measure of size, m_i, is identical to the actual number of households from which the sample is to be selected. Thus, if village A is selected and it contained 235 households according to the census, the list from which the sample of households will be selected for the survey also contains 235 households.

115. In many household survey applications that are based on census frames, however, the survey is conducted many months and sometimes years after the census was taken (see further discussion in the following chapter of up-dating sampling frames). Under those circumstances it is often decided to conduct a field operation in order to prepare a fresh list of households in clusters that are selected into the sample at the first stage. From the fresh listing a sample of households is then selected for the survey.

116. The measure of size, m_i, used to select the cluster is the census count of households, discussed in the example above. However, the actual list from which the sample of households will be selected will be different and of course have a different measure of size to some degree depending upon the length of time between when the census was taken and the survey listing is conducted. Differences will occur because of migration into and out of the cluster, construction of new housing or demolition of old, establishment of separate households when marriage occurs (sometimes within the same dwelling unit of the parental household) and death. When the sample is selected *pps*, its probability, from the selection equation, is

$$P(a\text{ß}) = [(a)(m_i)/ \text{ S } m_i,]\ [b/m'_i], \text{ where} \qquad [14]$$

m'_i is the count of households according to the listing operation and the other terms are defined as previously.

117. Since m'_i and m_i are likely to be different for most, if not all sample clusters, calculation of the probability of selection (and hence the weight, that is, the inverse of the probability) should take the difference into account. As [14] shows, each cluster would have a different weight, thus precluding a self-weighting sample design.

118. By using the exact weights that compensate for differences between census and survey measures of size, the resulting survey estimates will be unbiased. Failure to adjust the weights accordingly produces biased estimates whose magnitudes undoubtedly increase the longer the interval between the census and the survey. It should be noted, however, that when there are minor differences between m'_i and m_i, the sample is virtually self-weighting and it may, under some circumstances,[10] be prudent to generate the survey estimates without weighting because the biases would be negligible. Before deciding upon this course of action, however, it is essential to examine m'_i and m_i, cluster by cluster to gauge empirically if the differences are minor.

119. There is an alternative strategy which may be used to select households at the last stage whenever *ppes* sampling is employed – one in which the sample is actually self-weighting. It involves selecting households at a variable rate within each cluster depending upon its actual size. See the next section for more discussion.

2.7. Options in sampling

120. This section discusses some of the many options that may be considered in designing an appropriate sample for a general-purpose household survey. We focus primarily on strategies at the penultimate and final stages of selection since they are the stages where several alternatives are available. We examine the choice of *epsem* or *pps* sampling of clusters at the penultimate stage together with fixed-rate versus fixed-size sampling of households in the final stage. The section, to some extent, summarizes prior subsections with respect to issues of controlling sample size, self-weighting versus nonself-weighting designs plus other issues such as interviewer workloads. In addition we also review particular designs that are currently being

[10] In surveys where the estimates are restricted to proportions, rates or ratios this would be an appropriate strategy; for surveys where estimated totals or absolutes are wanted weighting must be used irrespective of whether the sample is self-weighting, approximately self-weighting or not self-weighting.

widely used such as those for the Demographic and Health Survey and UNICEF's Mid-decade Indicator Cluster Survey. Those designs provide additional options that are useful to consider including the use of compact (take-all) and non-compact clusters.

2.7.1. Epsem, pps, fixed-size, fixed-rate sampling

121. It is useful to look at a chart of potential designs to provide a framework for discussing the procedures, conditions, advantages and disadvantages of various sample plans.

Chart 1. Alternative Sample Plans – Last Two Stages of Selection

Selection of penultimate units	Fixed cluster size (number of households)	Fixed rate of selection in each cluster
PPS	Plan 1	Plan 2 [not recommended]
PPES	Plan 3	Plan 4 [not recommended]
Epsem	Plan 5	Plan 6

122. We have discussed how *pps* sampling of *PSU*s or clusters is a means of controlling the ultimate sample size more accurately than *epsem* sampling. That is its chief advantage, especially if the clusters are widely variable in the number of households each contains. Control of the sample size is important not only for its cost implications but also for permitting the survey manager to accurately plan interviewer workloads in advance of survey operations. *Epsem* sample selection, on the other hand, is simpler to carry out than *pps* and makes sense when the measure of size, *MOS*, of each cluster is approximately equal or little different from each other. As a practical matter *ppes* sampling must be used in lieu of *pps* whenever the actual *MOS* is different from the *MOS* as given on the frame.

123. Selection of a fixed number of households in each sample cluster has the very important advantages of (1) controlling the sample size precisely and (2) providing the means for the survey manager to assign exact workloads to interviewers and to equalize those workloads if she so chooses. Fixed-size sampling, however, is somewhat complicated as it requires the calculation of different sampling intervals for each cluster. Applying different sampling intervals can be confusing and error-prone, although there is a built-in quality control check since the number of households to be selected is known in advance; but the complications can create inefficiency from time lost through errors in selection and then correcting them.

124. Fixed-size sampling requires, by definition, a listing of households upon which the selected households can be designated and identified. Most often that listing is a currently prepared one undertaken as part of pre-survey field operations. It is useful to ensure that selection of the sample households be done in a central office and preferably by someone other than the lister himself, in order to minimize the possibility of bias in the selection procedure.

125. Alternatively, households may be sampled at a fixed rate in every cluster, which is simpler to select and less error-prone. An advantage in the field is that the sampling can be done at the time the interviewer is canvassing the cluster to obtain a current listing of households. This is accomplished by designing the listing form to show pre-designated lines for identifying

the sample households. Thus, listing and sampling can be done in a single visit, which has obvious advantages in cost. There are, however, some important disadvantages.

126. One disadvantage of fixed-rate sampling is that it provides little control over the sample size or the interviewer workloads, unless the *MOS* for each cluster is approximately the same. Another, more serious disadvantage is that when interviewers are entrusted with actually selecting the households for the sample, that is, identifying the ones that are to be listed on the sample lines, biased selection often results. Countless studies have been conducted that show that the households which are selected when the interviewers are in control tend to be smaller in size. This may be a conscious or even sub-conscious action on the part of interviewers to choose households that have fewer respondents so that the amount of work is decreased.

127. As mentioned earlier, self-weighting designs are advantageous in terms of both data analysis and reliability of the estimates. A design is self-weighting or not, depending upon the particular mix of sampling procedures at each stage. Thus in a two-stage design *pps* sampling of clusters coupled with fixed-size sampling of households is a self-weighting design while the combination of *pps* and fixed-rate is not. In the discussion below it is pointed out which of Plans 1-6 of Chart 1 are self-weighting.

Plan 1 – *pps*, fixed cluster size

Conditions

- Variable *MOS* for universe of clusters
- Households selected from same lists (example, census) that are used for *MOS*

Advantages

- Control of total sample size and hence cost
- Control of interviewer workloads
- Self-weighting

Disadvantages

- *PPS* somewhat more difficult than *epsem* to apply
- Different selection rates for choosing households from each cluster with its potential for errors

Plan 2 – *pps*, fixed rate

128. There are no plausible conditions under which this design would be used. If the clusters are variable in size then *pps* sampling together with a fixed cluster size is the proper plan to use. If clusters are of approximately equal size then fixed rate sampling is appropriate but the clusters themselves should be selected *epsem*.

Plan 3 – *ppes*, fixed cluster size

Conditions

- Variable *MOS* for universe of clusters
- Households selected from fresh listings up-dated from those used from the frame to establish the original *MOS*

Advantages

- Control of total sample size and hence cost
- Control of interviewer workloads
- More accurate than *pps* for a given frame because the household listings are current

Disadvantages

- *PPES* somewhat more difficult than *epsem* to apply
- Different selection rates for choosing households from each cluster with its potential for errors
- Not self-weighting

Plan 4 – *ppes*, fixed rate

There are no plausible conditions for utilizing plan 4 for the same reasons stated above for plan 2.

Plan 5 – *epsem*, fixed cluster size

Conditions

- *MOS* for universe of clusters is approximately equal or minimally variable

Advantages

- Control of total sample size (but somewhat less than plan 1) and hence cost
- Control of interviewer workloads but again somewhat less than plan 1
- *Epsem* easier to apply than *pps* or *ppes*

Disadvantages

- Different selection rates for choosing households from each cluster with its potential for errors
- Not self-weighting

Plan 6 – *epsem*, fixed rate

Conditions

- *MOS* for universe of clusters is virtually equal

Advantages

- Self-weighting
- Very simple to select sample at both stages

Disadvantages

- Poor control of total sample size with cost and reliability consequences especially if current MOS is substantially different from frame MOS and reliability consequences if sample is much smaller than targeted
- Little control of interviewer workloads

2.7.2. Demographic and Health Survey (DHS)

129. While the focus of DHS is women of child-bearing age its sample design is apt for general-purpose surveys.

130. The DHS, which has been widely applied in scores of developing countries since 1984, promotes the use of the *standard segment design*, for its convenience and practicality, in its sampling manual. A standard segment is defined in terms of its size, usually 500 persons. Each areal unit of the country that comprises the sampling frame is assigned a measure of size calculated as its population divided by 500 (or whatever standard segment size is decided upon for the country in question). The result, rounded to the nearest integer, is the number of standard segments in the areal unit.

131. A *pps* sample of areal units is selected using the number of standard segments as the measure of size, *MOS*. Since the areal units which are used for this stage of the sample are typically enumeration areas (*EA*s), city blocks or villages, the *MOS* for a large proportion of them is equal to one or two. For any selected areal unit with an *MOS* greater than one a mapping operation is organized in which geographic segments are created, the number of such segments equaling the *MOS*. Thus, a sample areal unit with *MOS* of 3 will be mapped to divide the unit into 3 segments of roughly equal size, to the extent that natural boundaries will allow, in terms of the number of persons in each segment (as opposed to its geographic size).

132. Each areal unit with *MOS* of one is automatically in sample, and, in each of the others, one segment is selected at random, *epsem*. All sample segments, including those automatically chosen, are then canvassed to obtain a current listing of households. A fixed fraction (rate) of households is selected systematically from each sample "cluster" for the DHS interview. Because the segments are all of approximately the same size, the sampling procedure yields a two-stage *epsem* sample of segments and of households.

133. The DHS standard segment design is close to plan 6 above – *epsem* selection of clusters and fixed-rate selection of households within sample clusters (also *epsem*). It avoids, however, the serious disadvantages noted above for plan 6 through its standard segmentation procedure; the overall sample size is controlled almost precisely as well as interviewer workloads.

134. A disadvantage of the standard segment design is that mapping operations must be conducted for segments with an *MOS* greater than one – a stage of *operations* if not a stage of sampling. Mapping can be tedious and costly, requires careful training and is subject to errors. Often natural boundaries are not well-defined so that segments can be reasonably delineated within the areal unit. That deficiency makes it difficult for interviewers who visit the segment later to locate exactly where the selected household might be. The latter problem can be

ameliorated somewhat, however, by including the name of the head of the household at the listing stage, in which case poor boundaries are less troublesome.

2.7.3. Modified Cluster Design in Multiple Indicator Cluster Surveys (MICS)

135. A common complaint among survey practitioners is the expense and time required to list households in the sample clusters that are selected in the penultimate stage. Listing is generally required in most surveys including the standard segment design method of DHS in order to obtain a current list of households from which to select those for the survey interview. It is especially crucial when the sampling frame is more than a year or so old. The listing operation is a significant survey cost and process that is often overlooked in both budgetary planning and survey scheduling. A separate visit to the field, apart from that required to conduct the survey interviews, must be made to effectuate listing. Moreover, it is frequently the case that the ratio of households to be listed is as much as five to ten times the number to be selected. For example, suppose the sample plan is to select 300 *PSU*s with cluster sizes of 25 households for a total of 7500 to be interviewed. If the average penultimate *PSU* contains 150 households, then 45,000 households must be listed.

136. The sampling strategy used for the EPI (Expanded Programme on Immunization) Cluster Survey [WHO, 1991] was developed by the Centers for Disease Control and World Health Organization partly to avoid the expense and time involved in listing. The EPI Cluster Survey is intended to estimate immunization coverage of children and it has been widely used in scores of developing nations for more than two decades. An important statistical issue [Turner etal 1997] is the sampling methodology. The cluster survey methodology utilizes a quota sample at the second stage of selection, even though the first stage units (villages or neighborhoods) are usually selected in accordance with the tenets of probability sampling. The quota sample method that is often used, although there are variations, is to commence the survey interviewing at some central point in the selected village, and then to proceed in a randomly determined direction, while continuing to interview households until a particular quota is met. Under the EPI Cluster Survey variation, enough households are visited until seven children in the target age group are found. While there is no intentional bias with the utilization of these kinds of techniques, various criticisms have been registered by many statisticians over a long period of time, including Kalton [1987], Scott [1993] and Bennett [1993]. The chief criticism is that the methodology does not produce a probability sample.

137. A variation of the EPI Cluster Survey method, the so-called modified cluster survey (MCS) design, was developed in response to the need to have a sampling strategy that avoids listing operations but is nevertheless grounded in probability sampling. Various applications of the MCS, as well as other designs, have received use around the globe in the Multiple Indicator Cluster Surveys (MICS) sponsored by UNICEF to monitor certain Child Summit goals and targets relating to the situation of children and women [UNICEF, 2000].

138. The MCS design is a minimalist sampling strategy. It uses a simple two-stage design, employs careful stratification plus quick canvassing and area segmentation. There is no listing operation. Essential features of the MCS sample design are as follows:

- Selection of a first stage sample of areal units such as villages or urban blocks using *pps*, or equal probability depending upon how variable the *PSU*s are with respect to their measures of size. Old measures of size can be used, even if the census frame is a few years old, though the frame must fully cover the population of interest - whether national or localized.
- Visits to each sample areal unit for quick canvassing plus area segmentation using existing maps or sketch maps, with the number of segments being predetermined and equal to the census measure of size divided by the desired (expected) cluster size. The segments which are created are approximately equal in population size in their expected value.
- Selection of one area segment with equal probability from each sample *PSU*.
- Conduct of interviews with all the households in each selected segment.

139. Use of segmentation without listing is the key advantage of MCS. This differs from the standard segment design of DHS that requires each segment to be listed. The segmentation operation also partially compensates for using a frame which may be out of date. While that has the advantage of producing an unbiased estimate, it also has the disadvantage of having less control over the ultimate sample size (a segment selected could be much bigger than the frame indicated due to growth) and a corresponding decrease in sampling precision.

140. Mapping, however, is required for MCS just as it is for the standard segment design of DHS, with all of the disadvantages this entails as mentioned for the DHS method. In addition, the creation of small segments, the size of which is the cluster size, that are accurately delineated can be difficult when natural boundaries are absent for small areas. A final disadvantage is that the segment interviewed is a compact cluster – all the households are geographically contiguous – and this has a greater design effect than the non-compact clusters of the standard segment design.

2.8. Special topics – two-phase samples and sampling for trends

141. This subsection covers two special topics on sample design in household surveys. One is the subject of two-phase sampling in which the first phase is used for a short interview in order to screen the household residents for persons who comprise the target population. The second phase of sampling entails selection of a sample of those who fit the criteria. The second topic discusses the sampling methodology when a survey is repeated for the purpose of estimating change or trend.

2.8.1. Two-phase sampling

142. A special type of sample design is needed in household surveys where not enough information is available to efficiently select a sample of the target population of interest. This need arises generally when the survey target population is a sub-population – often a rare one – whose members are not present in most households. Examples would be members of a particular ethnic group, orphans and persons with income above or below a specified level. Careful stratification can often be used to identify, for example, areal units where an ethic group of interest or high income persons are concentrated. But when such groups are dispersed fairly

randomly throughout the population or when the target group is rare such as orphans, then stratification is an insufficient strategy and other techniques must be used for sampling them.

143. One technique often used is two-phase sampling – also referred to as post-stratified sampling or double sampling. It involves four steps:

a. selection of a "large" sample of households,
b. conducting a short, screening interview to identify households where members of the target population reside,
c. post-stratifying the large sample into two categories based upon the screening interview and
d. selection of a sub-sample of households from each of the two strata for a second, longer interview with the target group.

144. The objective of the two-phase approach is to save costs by having a short, screening interview in the initial, large sample followed by the more extensive interview at a later date in the qualifying households. For that reason the initial sample is often one that was chosen for another purpose and the screening interview is appended as a "rider" to the parent survey. The procedure thus allows most resources to be allocated to the second-phase sampling and interviewing with only a modest budget required for the screening phase.

▪ ***Example***

Suppose a survey is being planned of 800 orphaned children who reside in households of a surviving parent or other relatives (as opposed to orphans living in institutional settings). Suppose, further, it is estimated that 16,000 households would have to be sampled to locate 800 orphans – about one orphan in every 20 households. Because the expense of designing and administering a sample of 16,000 households is considered impractical for only 800 detailed interviews, it is decided to make use of a general-purpose survey on health that is also being planned. The health survey is designed for a sample of 20,000 households. The survey managers of the two surveys agree that a rider will be appended to the health survey consisting of a single question, "Is there anyone 17 years old or younger living in this household whose mother, father or both have died?" The results of the rider question would be expected to identify households containing about 1000 orphans. From that the orphan survey manager would then plan to sub-sample 80-percent of those households for the detailed interview.

145. The example above also serves to illustrate *when* two-phase sampling is an appropriate strategy. Note that the targeted sample size, in the illustration, is only 800 orphans but the sample size in terms of the number of households necessary to find that many orphans is 16,000. Thus, in calculating the latter (see formula [1]) the sampling technician and the survey manager would likely conclude that two-phase sampling is the most practical and cost-efficient design to use.

146. Post-stratification of the first-phase sample is important for two reasons. The screening question or questions would almost always be brief because they are appended to another survey which undoubtedly already has a lengthy interview. The survey manager of the parent survey is unlikely to agree to a very detailed set of screening questions. Hence it is likely that some

households, in the example above, for which orphans were identified will have no orphans and vice-versa. Such misclassification errors suggest that two strata be set up, one for households where the screener result was positive and the other where it was negative. Samples would be taken from each stratum for the full interview on the grounds that misclassification probably occurred to some degree. The sample rate in the "yes" stratum would be very high - up to 100-percent - while in the "no" stratum a much smaller fraction would be taken.

2.8.2. Sampling to estimate change or trend

147. In many countries household surveys are designed with the dual purpose of estimating (1) *baseline* indicators (their *levels*) on the first occasion in which the survey is administered and (2) *change* in those indicators on second and subsequent administrations. When the survey is repeated more than once, trends in indicators are also measured. For repeat surveying, sample design is affected in various ways that do not pertain when a one-time, cross-sectional survey is conducted. In particular, the issues that are of concern are reliability for estimating change and the proper mix of using the same or different households from one occasion to the next. Related to the latter point is concern about biases and respondent burden when the same households are interviewed repeatedly.

148. To examine the reliability issue, we start by looking at the variance of the estimated change, $d = p_1 - p_2$. It is expressed by:

$$s^2_d = s^2_{p1} + s^2_{p2} - 2\, s_{p1,p2} \qquad [15]$$
$$= s^2_{p1} + s^2_{p2} - 2\ ?\ s_{p1} s_{p2}\text{, where}$$

d is the difference between p_1 and p_2 where the p value is the proportion being estimated,

s^2_d is the variance of the difference,

s^2_p is the variance of p on the first or second occasion, denoted by 1 or 2,

$s_{p1,p2}$ is the covariance between p_1 and p_2,

? is the correlation between the observed values of p_1 and p_2 on the two occasions of the survey.

Whenever the estimated change is comparatively small, which is often the case, we have:

$$s^2_{p1} \sim s^2_{p2}$$

Then, $s^2_d = 2\, s^2_p - 2\ ?\ s^2_p$ (We can drop the subscripts, 1 or 2).

$$\text{So, } s^2_d = 2\, s^2_p\ (1 - ?). \qquad [16]$$

149. To evaluate [16] we note that an estimate of s^{2}_{p} for a cluster survey is that of a simple random sample, *SRS*, times the sample design effect, *deff*, the latter of which we denote as *f*. The correlation, ?, is highest when the same sample of households is used and it may be 0.8 or even higher. In that case, the estimator, $s^{2}_{d,}$ of s^{2}_{d} is given as:

$$s^{2}_{d} = 2\ [(p\ q)\ f/n]\ (0.2), \text{ or } 0.4\ (p\ q)\ f/n. \qquad [17]$$

150. If the same clusters are used but different households, ? is still positive but substantially smaller – perhaps on the order of 0.25 to 0.35. We would then have (for ? of 0.3):

$$s2d = 2\ [(p\ q)\ f/n]\ (0.7), \text{ or } 1.4\ (p\ q)f/n. \qquad [18]$$

151. Finally, with a completely independent sample on the second occasion using different clusters and different households, ? is zero and we have:

$$s^{2}_{d} = 2\ [(p\ q)\ f\ /n]. \qquad [19]$$

Using a typical value for *deff* of 2.0, formula [19] yields:

$$s^{2}_{d} = 4\ [(p\ q)\ /n]. \qquad [20]$$

152. For repeat surveys using partial overlap, such as 50 percent of the same clusters/households and 50 percent new ones, ? must be multiplied by a factor, *F*, equal to the proportion of the sample that overlaps. In that case, equation [16] becomes:

$$s^{2}_{d} = 2\ s^{2}_{p}\ (1 - F\ ?). \qquad [21]$$

153. Interesting points can be observed from the above. First, the estimated variance of a comparatively small estimated change between two surveys using the same sample of households is only about 40 percent of the variance of level, on either the first or second occasion. Using the same clusters but different households produces a variance estimate on change that is 40 percent *higher* than that for level. Independent samples produce an estimated variance that is *double* that of level.

154. Thus, there are powerful advantages to using the same households in repeat surveys in terms of the reliability that can be attained. Failing that, there are still very significant improvements in using either (a) a portion of the same households or (b) the same clusters but with different households compared to the least attractive option of completely independent samples.

155. Regarding the issue of nonsampling error, there are two negative respondent effects - more non-response and conditioned responses the more the same sample of households is repeated. Respondents not only become increasingly reluctant to cooperate, thereby increasing non-response in later survey rounds, but they are also affected by conditioning, so that the quality or accuracy of their responses may deteriorate with repeat interviews.

156. Associated with the conditioning aspect is the problem known as "time-in-sample" bias, the phenomenon that survey estimates from respondents reporting for the same time period but with different levels of exposure to a survey have different expected values. This phenomenon has been extensively studied and has been shown to exist for surveys about many topics - labour force, expenditures, income and crime victimization. In the United States, for example, where the labour force survey respondents are interviewed 8 times, the estimate for unemployment for first-time-in-sample respondents is consistently about 7 percent higher than respondents averaged over the entire 8 interviews, and this pattern has persisted over a number of years in the U.S. Various reasons[11] have been postulated by experts to account for this bias and they include the following:

- Interviewers do not provide the same stimulus to the respondent in later interviews as in the first one.
- Respondents learn that some responses mean additional questions, so they may avoid giving certain answers.
- The first interview may include events outside the reference period, whereas in later interviews the event is "bounded."
- Respondents may actually change their behaviour because of the survey.
- Respondents may not be as diligent in providing accurate responses in later interviews once they become bored with the survey process.

157. It should be noted that most of the reasons cited above apply to repeat interviews for the same survey, but when the same households are used for different surveys some of the same respondent behaviour pertains.

158. Much of the discussion in the remainder of this subsection is taken from the sampling chapter, written by one of the authors of this Handbook, in the World Bank's CWIQ Survey Manual [World Bank, 2000].

Thus there are competing effects associated with using:

a. The same sample of households on each occasion.
b. Replacement households for part of the sample.
c. A new sample of households each time the survey is administered.

159. Proceeding from (a) to (c), sampling error on estimates of change increases while nonsampling error tends to decrease. Sampling error is least when the same sample households are used on each occasion because the correlation between observations is highest. By contrast, use of the same households increases nonsampling bias. The opposite pertains when a new sample of households is used each time. Sampling error to measure change is highest, but respondent conditioning is nil while non-response is that which is associated with first-time interviews.

[11] See *Panel Surveys*, editors Daniel Kasprzyk et al., John Wiley & Sons, New York, 1989, Chapter 1.

160. Alternative (b) above is the option that is generally offered to reach a compromise in balancing sampling error and nonsampling bias. If part of the sample is retained year-to-year, sampling error is improved over (c) and nonsampling error is improved over (a). When a survey is conducted on only two occasions, option (a) is likely the best choice. The respondent effects are not likely to be too damaging on the total survey error when a sample is used only twice. Repeat surveying three or more times would be better served by option (b), however. A convenient strategy is to replace 50 percent of the sample on each occasion in a rotating pattern (see more about rotation sampling for master samples in the next chapter).

References and further reading

Bennett, S. (1993), "The EPI Cluster Sampling Method: A Critical Appraisal," Invited Paper, International Statistical Institute's 49th Session, Florence.

Cochran, W.G. (1977), *Sampling Techniques*, third edition, Wiley, New York.

Hansen, M.H., Hurwitz, W.N. and Madow, W.G. (1953), *Sample Survey Methods and Theory*, Wiley, New York.

Hussmans, R., Mehran, F. and Verma. V. (1990), *Surveys of Economically Active Population, Employment, Unemployment and Underemployment, An ILO Manual on Concepts and Methods*, Chapter 11, “Sample Design,” International Labour Office, Geneva.

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

Kalton, G. (1983), *Introduction to Survey Sampling*, Sage, Beverly Hills.

________ (1987), "An Assessment of the WHO Simplified Cluster Sampling Method for Estimating Immunization Coverage," report to UNICEF, New York.

________ (1993), *Sampling Rare and Elusive Populations, National Household Survey Capability Programme,* United Nations Statistics Division, New York.

Kasprzyk, D. et al. editors (1989), *Panel Surveys*, Chapter 1, John Wiley & Sons, New York.

Kish, L. (1965), *Survey Sampling*, Wiley, New York.

Krewski, D., Platek, R., and Rao, J.N.K. editors (1981), *Current Topics in Survey Sampling*, Academic Press, New York.

League of Arab States (1990), *Sampling Manual, Arab Maternal and Child Health Survey,* Basic Documentation 5, Pan Arab Project for Child Development (PAPCHILD), Cairo.

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

Namboodiri, N.K. editor (1978), *Survey Sampling and Measurement*, Academic Press, New York.

Raj, D. (1972), *Design of Sample Surveys,* McGraw-Hill, New York.

Scott, C. (1993), Discussant comments for session on "Inexpensive Survey Methods for Developing Countries," Invited Paper Session, International Statistical Institute's 49th Session, Florence.

Som, R. (1966), *Practical Sampling Techniques*, second edition, Marcel Dekker, Inc., New York.

Turner, A., Magnani, R., and M. Shuaib (1996), "A Not Quite as Quick but Much Cleaner Alternative to the Expanded Programme on Immunization (EPI) Cluster Survey Design," *International Journal of Epidemiology*, Vol.25, No.1, Liverpool.

United Nations Children’s Fund (2000), *End-Decade Multiple Indicator Survey Manual*, Chapter 4, “Designing and Selecting the Sample,” UNICEF, New York.

United Nations Statistics Division (1984), *Handbook of Household Surveys*, revised edition, United Nations, New York.

____________________________ (1986), *Sampling Frames and Sample Designs for Integrated Household Survey Programmes, National Household Survey Capability Programme,* United Nations, New York.

United States Bureau of the Census (1978), *Current Population Survey Design and Methodology*, Technical Paper 40, Bureau of the Census, Washington.

Verma, V. (1991), *Sampling Methods*. Training Handbook, Statistical Institute for Asia and the Pacific, Tokyo.

Waksberg, J. (1978), "Sampling Methods for Random Digit Dialing," The Journal of the American Statistical Association, 73, 40-46.

World Bank (1999), *Core Welfare Indicators Questionnaire (CWIQ) Handbook*, Chapter 4 "Preparing the CWIQ Sample Design," World Bank, Washington.

World Health Organization (1991), *Expanded Programme on Immunization, Training for Mid-level Managers: Coverage Survey,* WHO/EPI/MLM91.10, Geneva.