Department of Economic and Social Affairs

Statistics Division

Studies in Methods Series F No.98

Designing Household Survey Samples: Practical Guidelines

Logo

United Nations

New York, 2005

The Department of Economic and Social Affairs o f the United Nations Secretariat is a vital interface between global policies in the economic, social and environmental spheres and national action. The department works in three main interlinked areas: (i) it compiles, generates and analyses a wide range of economic, social and environmental data and information on which States Members of the United Nations draw to review common problems and to take stock of policy options; (ii) it facilitates the negotiations of Members States in many intergovernmental bodies on joint courses of action to address ongoing or emerging global challenges; and (iii) it advises interested Governments on the ways and means of translating policy frameworks developed in United Nations conferences and summits into programmes at the country level and, through technical assistance, helps build national capacities.

NOTE

Symbols of the United Nations documents are composed of capital letters combined with figures.

ST/ESA/STAT/SER.F/98

UNITED NATIONS PUBLICATION
Sales No.

ISBN

Copyright @ United Nations 2005
All rights reserved

Preface

The main purpose of the handbook is to include in one publication the main sample survey design issues that can conveniently be referred to by practicing national statisticians, researchers and analysts involved in sample survey work and activities in countries. Methodologically sound techniques that are grounded in statistical theory are used in the handbook, implying the use of probability sampling at each stage of the sample selection process. A well designed household survey which is properly implemented can generate necessary information of sufficient quality and accuracy with speed and at a relatively low cost.

The contents of the handbook can also be used, in part, as a training guide for introductory courses in sample survey design at various statistical training institutions that offer courses in applied statistics, especially survey methodology.

In addition, the handbook has been prepared to complement other publications dealing with sample survey methodology issued by the United Nations, such as the recent publication on Household Surveys in Developing and Transitional Countries and the series under the National Household Survey Capability Programme (NHSCP).

More specifically, the objectives of the Handbook are to:

a. Provide, in one publication, basic concepts and methodologically sound procedures for designing samples for, in particular, national-level household surveys, emphasizing applied aspects of household sample design;
b. Serve as a practical guide for survey practitioners in designing and implementing efficient household sample surveys;
c. Illustrate the interrelationship of sample design, data collection, estimation, processing and analysis;
d. Highlight the importance of controlling and reducing *nonsampling errors* in household sample surveys.

While having a sampling background is helpful in using the handbook, other users with a general knowledge of statistical and mathematical concepts should also be able to use and apply the handbook with little or no assistance. This is because one of the key aims of the handbook is to present material in a practical, hands-on format as opposed to stressing the theoretical aspects of sampling. Theoretical underpinnings are, however, provided when necessary. It is expected that a basic understanding of algebra is all that is needed to follow the presentation easily and to apply the techniques. Accordingly, numerous examples are provided to illustrate the concepts, methods and techniques.

In the preparation of this handbook the United Nations Statistics Division was assisted by Mr. Anthony Turner, Sampling Consultant, who drafted chapters 3 to 5 and finally reviewed the consolidated document. Mr. Ibrahim Yansaneh, Deputy Chief of Cost of living Division of the International Civil Service Commission, drafted chapters 6 and 7 and Mr. Maphion Jambwa, Technical Adviser, Southern African Development Community Secretariat, drafted chapter 9. The draft chapters were reviewed by an Expert Group Meeting organized, in New York, by the

United Nations Statistics Division from 3 to 5 December 2003. List of experts is given in the Appendix.

Preface i
Chapter 1 1
Sources of Data for Social and Demographic Statistics 1
1.1 Introduction 1
1.2 Data sources 1
1.2.1 Household surveys 1
1.2.2 Population and housing censuses 4
1.2.3 Administrative records 5
1.2.4 Complementarities of the three data sources 6
1.2.5 Concluding remarks 7
References and further reading 8
Chapter 2 9
Planning and Execution of Surveys 9
2.1 Planning of surveys 9
2.1.1 Objectives of a survey 9
2.1.2 Survey universe 10
2.1.3. Information to be collected 10
2.1.4 Survey budget 11
2. 2. Execution of surveys 16
2.2.1 Data collection methods 16
2.2.2 Questionnaire design 18
2.2.3 Tabulation and analysis plan 20
2.2.4 Implementation of field work 21
References and further reading 26
Chapter 3 27
Sampling Strategies 27
3.1 Introduction 27
3.1.1 Overview 27
3.1.2 Glossary of sampling and related terms 28
3.1.3 Notations 31
3. 2 Probability sampling versus other sampling methods for household surveys 32
3.2.1 Probability sampling 32
3.2.2 Non-probability sampling methods 34
3.3 Sample size determination for household surveys 36
3.3.1 Magnitudes of survey estimates 37
3.3.2 Target population 37
3.3.3 Precision and statistical confidence 38
3.3.4 Analysis groups - domains 39
3.3.5 Clustering effects 41
3.3.6 Adjusting sample size for anticipated non-response 42
3.3.7 Sample size for master samples 42
3.3.8 Estimating change or level 43
3.3.9 Survey budget 43
3.3.10 Sample size calculation 44
3.4 Stratification 46

3.4.1 Stratification and sample allocation ... 47
3.4.2 Rules of stratification ... 48
3.4.3 Implicit stratification ... 49
3.5 Cluster sampling ... 50
3.5.1 Characteristics of cluster sampling ... 51
3.5.2 Cluster design effect ... 51
3.5.3 Cluster size ... 53
3.5.4 Calculating *deff* ... 54
3.5.5 Number of clusters ... 54
3.6 Sampling in stages ... 54
3.6.1 Benefits of sampling in stages ... 54
3.6.2 Use of dummy stages ... 55
3.6.3 The two-stage design ... 58
3.7 Sampling with probability proportional to size (PPS) ... 60
3.7.1 *PPS* sampling ... 60
3.7.2 *PPES* sampling (probability proportional to estimated size) ... 63
3.8 Options in sampling ... 65
3.8.1 *Epsem*, *PPS*, fixed-size, fixed-rate sampling ... 65
3.8.2 Demographic and Health Survey (DHS) ... 69
3.8.3 Modified Cluster Design - Multiple Indicator Cluster Surveys (MICS) ... 70
3.9 Special topics – two-phase samples and sampling for trends ... 72
3.9.1 Two-phase sampling ... 72
3.9.2 Sampling to estimate change or trend ... 74
3.10 When implementation goes wrong ... 77
3.10.1 Target population definition and coverage ... 77
3.10.2 Sample size too large for survey budget ... 78
3.10.3 Cluster size larger or smaller than expected ... 78
3.10.4 Handling non-response cases ... 79
3.11 Summary guidelines ... 79
References and further reading ... 81
Chapter 4 ... 83
Sampling Frames and Master Samples ... 83
4.1 Sampling frames in household surveys ... 83
4.1.1 Definition of sample frame ... 83
4.1.2 Properties of sampling frames ... 84
4.1.3 Area frames ... 86
4.1.4 List frames ... 87
4.1.5 Multiple frames ... 88
4.1.6 Typical frame(s) in two-stage designs ... 89
4.1.7 Master sample frames ... 90
4.1.8 Common problems of frames and suggested remedies ... 90
4.2 Master sampling frames ... 94
4.2.1 Definition and use of a master sample ... 94
4.2.2 Ideal characteristics of PSUs for a master sample frame ... 94
4.2.3 Use of master samples to support surveys ... 95
4.2.4 Allocation across domains (administrative regions, etc.) ... 97

4.2.5 Maintenance and updating of master samples 98
4.2.6 Rotation of PSUs in master samples 98
4.2.7 Country examples of master samples 100
4.3 Summary guidelines 106
References and further reading 108
Chapter 5 109
Documentation and Evaluation of Sample Designs 109
5.0 Introduction 109
5.1 Need for, and types of, sample documentation and evaluation 109
5.2 Labels for design variables 110
5.3 Selection probabilities 112
5.4 Response rates and coverage rates at various stages of sample selection 112
5.5 Weighting: base weights, non-response and other adjustments 113
5.6 Information on sampling costs 114
5.7 Evaluation – limitations of survey data 115
5.8 Summary guidelines 117
References and further reading 118
Chapter 6 119
Construction and Use of Sample Weights 119
6.1 Introduction 119
6.2 Need for sampling weights 119
6.2.1 Overview 120
6.3 Development of sampling weights 120
6.3.1 Adjustments of sample weights for unknown eligibility 121
6.3.2 Adjustments of sample weights for duplicates 122
6.4 Weighting for unequal probabilities of selection 123
6.4.1 Case study in construction of weights: Vietnam National Health Survey 2001 127
6.4.2 Self-weighting samples 128
6.5 Adjustment of sample weights for non-response 128
6.5.1 Reducing non-response bias in household surveys 129
6.5.2 Compensating for non-response 129
6.5.3 Non-response adjustment of sample weights 130
6.6 Adjustment of sample weights for non-coverage 132
6.6.1 Sources of non-coverage in household surveys 133
6.6.2 Compensating for non-coverage in household surveys 134
6.7 Increase in sampling variance due to weighting 135
6.8 Trimming of Weights 136
6.9 Concluding Remarks 138
References and further reading 140
Chapter 7 141
Estimation of Sampling Errors for Survey Data 141
7.1 Introduction 141
7.1.1 Sampling error estimation for complex survey data 141
7.1.2 Overview of the chapter 142
7.2 Sampling variance under simple random sampling 143
7.3 Other measures of sampling error 149

7.3.1 Standard error 149
7.3.2 Coefficient of variation 149
7.3.3 Design effect 150
7.4 Calculating sampling variance for other standard designs 150
7.4.1 Stratified sampling 150
7.4.2 Single-stage cluster sampling 153
7.5 Common features of household survey sample designs and data 154
7.5.1 Deviations of household survey designs from simple random sampling 154
7.5.2 Preparation of data files for analysis 154
7.5.3 Types of Survey Estimates 155
7.6 Guidelines for presentation of information on sampling errors 156
7.6.1 Determining what to report 156
7.6.2 How to report sampling error information 157
7.6.3 Rule of thumb in reporting standard errors 157
7.7 Methods of variance estimation for household surveys 158
7.7.1 Exact methods 158
7.7.2 Ultimate cluster method 159
7.7.3 Linearization approximations 163
7.7.4 Replication 165
7.7.5 Some replication techniques 167
7.8 Pitfalls of using standard statistical software packages to analyze household survey data 172
7.9 Computer software for sampling error estimation 174
7.10 General comparison of software packages 177
7.11 Concluding remarks 177
References and further reading 179
Chapter 8 181
Nonsampling Errors in Household Surveys 181
8.1 Introduction 181
8.2 Bias and variable error 182
8.2.1 Variable component 184
8.2.2 Systematic error (bias) 185
8.2.3 Sampling bias 185
8.2.4 Further comparison of bias and variable error 185
8.3 Sources of nonsampling error 186
8.4 Components of nonsampling error 186
8.4.1 Specification error 186
8.4.2 Coverage or frame error 187
8.4.3 Non-response 189
8.4.4 Measurement error 190
8.4.5 Processing errors 191
8.4.6 Errors of estimation 191
8.5 Assessing nonsampling error 192
8.5.1 Consistency checks 192
8.5.2 Sample check/verification 192
8.5.3 Post-survey or re-interview checks 193
8.5.4 Quality control techniques 193

8.5.5 Study of recall errors 194
8.5.6 Interpenetrating sub-sampling 194
8.6 Concluding remarks 195
References and further reading 196
Chapter 9 197
Data Processing for Household Surveys 197
9.1 Introduction 197
9.2 The household survey cycle 197
9.3 Survey planning and the data processing system 199
9.3.1 Survey objectives and content 199
9.3.2 Survey procedures and instruments 199
9.3.3 Design for household surveys data processing systems 202
9.4 Survey operations and data processing 206
9.4.1 Frame creation and sample design 206
9.4.2 Data collection and data management 208
9.4.3 Data preparation 209
References and further reading 226
Software options for different steps of survey data processing 230
Annex: Overview of sample survey design 233
A.1 Sample design 233
A. 2 Basics of probability sampling strategies 235
A.2.1 Simple random sampling 235
A.2.2 Systematic sampling 240
A.2.2.1. Linear systematic sampling 241
A.2.2.2 Circular systematic sampling 241
A. 2.2.3. Estimation in systematic sampling 242
A. 2.2.4 Advantages of using systematic sampling 244
A. 2.2.5. Disadvantages of Systematic Sampling 246
A.2.3 Stratification 246
A.2.3.1. Advantages of stratified sampling 246
A. 2.3.2. Weights 248
A. 2.3.3. Sample values 248
A. 2.3.4. Proportional allocation 249
A. 2.3.5. Optimum allocation 249
A. 2.3.6 Determination of within stratum sample sizes 251
A.2.4 Cluster sampling 252
A.2.4.1. Some reasons for using cluster sampling 253
A.2.4.2. Single stage cluster sampling 253
A.2.4.3. Sample mean and variance 254
Appendix 255
List of E Experts 255

Chapter 1

Sources of Data for Social and Demographic Statistics

1.1 Introduction

1. Household surveys are among three major sources of social and demographic statistics in many countries. It is recognized that population and housing censuses are also a key source of social statistics but they are conducted, usually, at long intervals of about ten years. The third source is administrative record systems. For most countries this source is somewhat better developed for health and vital statistics, however, than for social statistics. Household surveys provide a cheaper alternative to censuses for timely data and a more relevant and convenient alternative to administrative record systems. They are used for collection of detailed and varied socio-demographic data pertaining to conditions under which people live, their well-being, activities in which they engage, demographic characteristics and cultural factors which influence behaviour, as well as social and economic change. This, however, does not preclude the complementary use of data generated through household surveys with data from other sources such as censuses and administrative records.

1.2 Data sources

2. As mentioned in the introductory section, the main sources of social and demographic data are population and housing censuses, administrative records and household sample surveys. These three sources, if well planned and executed, can be complementary in an integrated programme of data collection and compilation. Social and demographic statistics are essential for planning and monitoring socio-economic development programmes. Statistics on population composition by age and sex including geographical distribution are among the most basic data necessary to describe a population and/or a sub-group of a population. These basic characteristics provide the context within which other important information on social phenomena, such as education, disability, labour force participation, health conditions, nutritional status, criminal victimization, fertility, mortality and migration, can be studied.

1.2.1 Household surveys

3. Household sample surveys have become a key source of data on social phenomena in the last 60-70 years. They are among the most flexible methods of data collection. In theory almost any population-based subject can be investigated through household surveys. It is common for households to be used as second-stage sampling units in most area-based sampling strategies (see chapters 3 and 4 of this handbook). In sample surveys part of the population is selected from which observations are made or data are collected and then inferences are made to the whole population. Because in sample surveys there are smaller workloads for interviewers and a longer time period assigned to data collection, most subject matter can be covered in greater detail than in censuses. In addition, because there are far fewer field staff needed more qualified individuals can be recruited and they can be trained more intensively than is possible in a census operation.

The reality is that not all the data needs of a country can be met through census-taking; therefore, household surveys provide a mechanism for meeting the additional and emerging needs on a continuous basis. The flexibility of household surveys makes them excellent choices for meeting data users' needs for statistical information which otherwise would not be available and insufficient.

1.2.1.1 Types of household surveys

4. Many countries have in place household survey programmes that include both periodic and *adhoc* surveys. It is advisable that the household survey programme be part of an integrated statistical data collection system of a country. In the area of social and demographic statistics intercensal household surveys can constitute part of this system.

5. The National Household Survey Capability Programme (NHSCP) was a major effort to help developing countries establish the statistical and survey capabilities to obtain requisite socio-economic and demographic information from the household sector. The NHSCP was implemented for nearly 14 years from 1979 to 1992. By the time of its conclusion, 50 countries had participated in the programme. Its major achievement was the promotion and adoption by countries of continuous multi-subject, integrated household surveys. In addition, the programme fostered sample survey capacity-building, especially in African countries.

6. There are different types of household surveys that can be conducted to collect data on social and demographic statistics such as specialized surveys, multi-phase surveys, multi-subject surveys and longitudinal surveys. The selection of the appropriate type of survey is dependent on a number of factors including subject matter requirements, resources and logistical considerations.

7. Specialized surveys cover single subjects or issues such as time-use or nutritional status. The surveys may be periodic or ad-hoc.

8. Multi-phase surveys entail collecting statistical information in succeeding phases with one phase serving as a precursor to the next. The initial phase usually constitutes a larger sample than subsequent phases. It is used to screen sample units based on certain characteristics to ascertain the eligibility of such units to be used in the subsequent phases. These surveys are a cost-effective way of reaching the target population in the latter phases to collect detailed information on a subject of interest. The study of such topics as disability and orphanhood are among those suited to this approach.

9. In multi-subject surveys, different subjects are covered in a single survey. This approach is generally more cost-effective than conducting a series of single subject surveys.

10. In longitudinal surveys, data is collected from the same sample units over a period of time. The interval can be monthly, quarterly or annually. The purpose for conducting such surveys is to measure changes in some characteristics for the same population over a period of time. The major problem with this type of surveys is the high attrition rate of respondents. There is also the problem of conditioning effect.

1.2.1.2 Advantages and limitations compared to censuses

11. While household surveys are not as expensive as censuses they can, nevertheless, become quite costly if results have to be produced separately for relatively lower administrative domains such as provinces or districts. Unlike a census where data are collected for millions of households, a sample survey is typically limited to a sample of several thousand households due to cost constraint, severely limiting its use to produce reliable data for small areas. See more about the relationship of sample size on data reliability for small areas and domains in succeeding chapters.

Some advantages of household surveys compared to censuses are as follows:

a. As mentioned above, the overall cost of a survey is generally lower compared to a census as the latter requires large amounts of manpower, financial, logistical and material resources. From a probability sample, properly selected and implemented, accurate and reliable results can be a basis for making inferences on the total population. Consequently for some estimates such as total fertility rate, there is no compelling need for a census.
b. In general sample surveys produce statistical information of better quality because, as stated earlier, it is more feasible to engage better and well-trained interviewers. It is also easier to provide better supervision because supervisors are usually well trained and the supervisor/interviewer ratio can be as high as 1 to 4. In addition it is possible to use better technical equipment for taking physical measurements in surveys when such measurements are needed. In a census data quality is, in some cases, compromised because of the massive nature of the exercise, which is prone to lapses in, and neglect of, quality assurance at various stages, resulting in high nonsampling errors.
c. There is greater scope and flexibility in a sample survey than in a census with respect to the depth of investigation and number of items in the questionnaire. Information of a more specialized type may not be collected in a census because of the prohibitive number of specialists or equipment necessary to carry out the study. An example is the weighing of food and other measurements in a nutrition study. It is likewise not feasible to subject every person in the population to a medical examination to collect health information such as the incidence of HIV/AIDS infection. On the other hand, it is possible to add items in a household sample survey that would be relatively complex for the census.

12. Sample surveys are better suited for the collection of national and relatively large geographic domain level data on topics that need to be explored in depth such as the multi-dimensional aspects of disability, household expenditure, labour-force activities and criminal victimization. This is in contrast to censuses that collect and are a source of relatively general information covering small domains.

13. In general, the strengths of household survey statistical operations include the flexibility of data collection instruments to accommodate a larger number of questions on a variety of

topics and also the possibility of estimating parameters comparable to those measured in population and housing censuses.

1.2.2 Population and housing censuses

14. A population census, henceforth referred to as census, is the total process of collecting, compiling, evaluating and disseminating demographic, social and other data at a specified time covering all persons in a country or in well-delimited part(s) of a country. It is a major source of social statistics, with its obvious advantage of providing reliable data – that is, unaffected by sampling error - for small geographic units. A census is an ideal method for providing information on size, composition and spatial distribution of the population in addition to socio-economic and demographic characteristics. In general the census collects information for each individual in households and each set of living quarters, usually for the whole country or well-defined parts of the country.

1.2.2.1 Basic features of a traditional population and housing census

a. Individuals in the population and each set of living quarters are enumerated separately and the characteristics thereof are recorded separately.
b. The goal of a census is to cover the whole population in a clearly defined territory. It is intended to include every person present and/or usual residents depending on whether the type of population count is *de facto* or *de jure*. In the absence of comprehensive population or administrative registers, censuses are the only source that can provide small area statistics.
c. The enumeration over the entire country is generally as simultaneous as possible. All persons and dwellings are enumerated with respect to the same reference period.
d. Censuses are usually conducted at defined intervals. Most countries conduct censuses every 10 years while others every five years. This facilitates the availability of comparable information at fixed intervals.

1.2.2.2 Uses of census results

a. Censuses provide information on size, composition and spatial distribution of population together with demographic and social characteristics.
b. Censuses are a source of small area statistics.
c. Census enumeration areas are the major source of sampling frames for household surveys. Data collected in censuses are often used as auxiliary information for stratifying samples and for improving the estimation in household surveys.

1.2.2.3 Main limitations of censuses

15. Because of its unparalleled geographical coverage it is usually a major source of baseline data on the characteristics of the population. It is not feasible, therefore, to cover many topics

with appreciable detail. The census may not be the most ideal source of detailed information, for example on, economic activity. Such information requires detailed questioning and probing.

16. Because the census interview relies heavily on proxy respondents it may not always capture accurate information on characteristics which only an individual might know, such as occupation, hours worked, income, etc.

17. Population censuses have been carried out in many countries during the past few decades. For example, about 184 countries and areas have conducted censuses during the 2000 round (1995-2004).

1.2.3 Administrative records

18. Many types of social statistics are compiled from various administrative records as by-products of the administrative processes. Examples include health statistics compiled from hospital records, employment statistics from employment exchange services, vital statistics compiled from the civil registration system and education statistics from enrolment reports of the ministries of education. The reliability of statistics from administrative records depends on the completeness of the administrative records and the consistency of definitions and concepts.

19. While administrative records can be very cost-effective sources of data, such systems are not well established in most developing countries. This implies that in a majority of cases such data are inaccurate. Even if the administrative recording processes are continuous for purposes of administration, the compilation of statistics is, in most cases, a secondary concern for most organizations and, as a result, the quality of the data suffers. Statistical requirements that need to be maintained such as standardization of concepts and definitions, adhering to timeliness and complete coverage are not usually considered or adhered to.

20. For most countries, information from administrative records is often limited in content as their uses are more for legal or administrative purposes. Civil registration systems are examples of administrative systems that many countries have developed. However, not all countries have been successful in this effort. Countries with complete vital registration systems are able to produce periodic reports on vital events, such as number of live births by sex; date and place of births; number of deaths by age; sex; place of deaths and cause of death; marriages and divorces; etc.

21. A population register maintains life databases for every person and household in a country. The register is updated on a continuous basis when there are changes in the characteristics of an individual and/or a household. If such registers are combined with other social registers they can be a source of rich information. Countries which have developed such systems include Denmark, Norway, the Netherlands, Germany and Sweden. For most of these countries censuses are based on the registration system.

22. In many developing countries, while administrative records for various social programmes can be cost-effective data source and an attractive proposition, they are not well

developed. Administrative records are often limited in content and do not usually have the adaptability of household surveys from the standpoint of concepts or subject detail. In this case their complementary use with other sources is a big challenge because of lack of standardized concepts, classification systems coupled with selective coverage and under-coverage.

1.2.4 Complementarities of the three data sources

23. The sub-sections above, in this chapter, have alluded to various ways in which censuses, surveys and administrative record systems can be used in concert. This sub-section provides more detail on the subject of combining information from different data sources in a complementary fashion. The interest in this area is driven by the necessity to limit census and survey costs and to lower response burden, to provide data at lower domains, which may not be covered by survey data for instance, and to maximize the use of available data in the country.

24. Because censuses cannot be repeated frequently, household surveys provide a basis for updating some census information especially at national and other large domain levels. In most cases only relatively simple topics are investigated in a census and the number of questions is usually limited. Census information can therefore be complemented by detailed information on complex topics from the household surveys, taking advantage of their small size and potential flexibility.

25. Censuses and household surveys have, in many instances, been complementary. Collecting information on additional topics from a sample of the households during the census is a cost-effective way to broaden the scope of the census to meet the expanding demands of social statistics. The use of sampling methods and techniques makes it feasible to produce urgently needed data with acceptable precision when time and cost constraints would make it impractical to obtain such data through complete enumeration.

26. The census also provides a sampling frame, statistical infrastructure, statistical capacity and benchmark statistics that are needed in conducting household surveys. It is common to draw a sample of households within a census context, to collect information on more complex topics such as, disability, maternal mortality, economic activity and fertility.

27. Censuses support household surveys by providing sampling frames; the census provides an explicit list of all area units, such as enumeration areas, commonly used as first stage units in household sample surveys selection process. Moreover, some auxiliary information available from a census can be used for efficient design of surveys. Furthermore, auxiliary information from censuses can be used to improve sample estimates through regression and ratio estimates, thereby improving the precision of survey estimates.

28. In order to achieve integration of data sources there is need to clearly identify units of enumeration and adopt consistent geographic units in collecting and reporting statistics through the various sources. In addition, it is essential to adopt common definitions, concepts and classifications across different sources of data including administrative records.

29. Data from household surveys can also be used to check census coverage and content. The aim is to determine the size and direction of such errors. Post enumeration surveys were, for instance, used for this purpose during the 2000 round of censuses in Zambia and in Cambodia to evaluate coverage errors. Likewise census data can be used to evaluate some survey results.

30. Small area estimation, which has received a lot of attention due to growing demand for reliable small area estimators, is an area where data from surveys and administrative records are used to produce estimates concurrently. Traditional area-specific direct estimators do not provide adequate precision because sample sizes in small areas are seldom large enough. Small area estimation is based on a range of statistical techniques used to produce estimates for areas when traditional survey estimates for such areas are unreliable or cannot be calculated. The techniques involve models that borrow strength over space and time or from auxiliary information from administrative records or censuses. The basic idea of small area procedures is, therefore, to borrow and combine the relative strength of different sources of data in an effort to produce more accurate and reliable estimates.

31. In countries with well-developed civil registration systems, census and survey data can be successfully used together with data from administrative records. For example in the 1990 population census in Singapore, interviewers had pre-filled basic information, from administrative records, for every member of the household. This approach reduced interviewing time and enumeration costs. Since the register-based census provides only the total count of the population and basic characteristics of the population, detailed socio-economic characteristics are collected on a sample basis.

32. Data from administrative records can be used to check and evaluate results from surveys and censuses. For instance in countries with complete vital registration systems, data on fertility and mortality from censuses can be cross-checked with that from the registration system.

1.2.5 Concluding remarks

33. In conclusion household, surveys, censuses and administrative sources should be viewed as complementary. This implies that, whenever possible, in planning for censuses and surveys common concepts and definitions should be used. Administrative procedures should also be checked periodically to make sure that common concepts and definitions are being used.

34. The household survey programme should be part of an integrated statistical data collection system within a country, including censuses and administrative records so that the overall needs for socio-demographic statistics can be adequately met.

References and further reading

Ambler, R. et al. (2001), "Combining Unemployment Benefits Data and LFS to Estimate ILO Unemployment for Small Areas: An Application of Modified Fay-Herriot Method," Invited Paper, International Statistical Institute Session, Seoul.

Banda, J (2003), Current Status of Social Statistics: An overview of Issues and Concerns, United Nations Expert Group Meeting in collaboration with the Siena Group on Social Statistics, New York, 6-9 May 2003.

Bee-Geok, L. and Eng-Chuan, K. (2001), "Combining Survey and Administrative Data for Singapore's Census of Population 2000," Invited Paper, International Statistical Institute Session, Seoul.

Kiregyera, B. (1999), *Sample Surveys: with Special Reference to Africa*, PHIDAM Enterprises, Kampala.

Rao, J.N.K. (1999), "Some Recent Advances in Model-based Small Area Estimation," *Survey Methodology*, Vol.25, No.2, pp. 175-186, Statistics Canada, Ottawa.

Singh, R. and Mangat, N. (1996), *Elements of Survey Sampling*, Kluwer Academic Publishers, Boston.

Statistics Canada. (2003) *Survey Methods and Practices*, Ottawa.

United Nations Statistics Division (1982), *Nonsampling Errors in Household Surveys: Sources, Assessment and Control, National Household Survey Capability Programme*, United Nations, New York.

_____ (1984), *Handbook of Household Surveys*, revised edition ST/ESA/SER.F/31, United Nations, New York.

_____ (1998), *Principles and Recommendations for Population and Housing Censuses*, Revision 1, ST/ESA/STAT/SER.M/67/REV.1, New York

_____ (2001), *Principles and Recommendations for a Vital Statistics System*, Revision 2, ST/ESA/STAT/SER.M/19/Rev. 2, New York.

United Nations (2002), *Technical Report on Collection of Economic Characteristics in Population Censuses*, Statistics Division, Department of Social and Economic Affairs and Bureau of Statistics, International Labour Office, New York and Geneva.

Whitfold, D. and Banda, J. (2001), "Post Enumeration Surveys: Are they Worth it?" United Nations Symposium on Global Review of 2000 Round of Population and Housing Censuses: Mid –Decade Assessment and Future Prospects, New York, 7-10 August.

Chapter 2

Planning and Execution of Surveys

1. While the emphasis of this handbook is on the sampling aspects of household surveys, it is necessary to provide an overview of household survey planning, operations and implementation in order to fit the sampling chapters and sections into proper context. There are many textbooks, handbooks and manuals that deal, in considerable detail, with the subject of household survey planning and execution and the reader is urged to refer to those for more information. Many of the main points, however, are highlighted and briefly described in this chapter. A key feature of planning and execution is sample design and selection but those topics are discussed in chapters 3 and 4.

2.1 Planning of surveys

2. For a survey to yield desired results, there is need to pay particular attention to the preparations that precede the field work. In this regard all surveys require careful and judicious preparations if they have to be successful. However, the amount of planning will vary depending on the type of survey, materials and information required. The development of an adequate survey plan requires sufficient time and resources and a planning cycle of two years is not uncommon for a complex survey (for detailed discussion on survey planning see United Nations Handbook of Household Surveys ST/ESA/Ser.F/31, 1984).

2.1.1 Objectives of a survey

3. It is imperative that the objectives of a survey be clearly spelled out from the start of the project. There should be a clear statistical statement on the desired information, giving a clear description of the population and geographical coverage. It is also necessary at this stage to stipulate how the results are going to be used. The given budget of the survey should guide the survey statistician in tailoring the objectives. Taking due cognizance of the budgetary constraints will facilitate successful planning and execution of the survey.

4. In some cases objectives of surveys are not explicitly stated. For instance, a survey organization may be called upon to carry out a study on the activities of the informal sector. If the purpose is not clearly stated, it is for the statistician or survey manager to define the informal sector in operational terms for survey-taking, outlining in detail the particular economic activities that most closely reflect the requirements of the sponsoring agency. It should be mentioned that a survey which has ambiguous and vague objectives is very much susceptible to high nonsampling errors.

5. It is very important that stakeholders, thus various users and producers of statistics, be involved in defining the objective of the survey as well as its scope and coverage. The consultations help to come up with consensus or compromises on what data are needed, the form in which data are required, levels of disaggregation, dissemination strategies and frequency of data collection.

6. Some of the surveys conducted by survey organizations have clear objectives. For example, the 1983 Zambian Pilot Manpower Survey had the following objectives:

a. To collect information on the size and composition of currently working population in the formal sector;
b. To assess manpower demand and supply;
c. To serve as a basis for making manpower projections for particular occupations.
d. To assist in planning for the expansion of education in fields that are crucial to economic development.

7. It should be noted that having clearly stated objectives is the first step in forming the basis of what questions to ask in the survey for which statistical answers are required.

2.1.2 Survey universe

8. When planning to carry out a survey, it is necessary to define the geographical areas to be covered and the target population. In a household Income and Expenditure Survey, for instance, the survey may cover the urban areas and perhaps exclude rural areas.

9. In defining the universe, the exact population to be sampled should be identified. In the above-mentioned survey the universe of first stage units would be enumeration areas (*EA*s) in urban areas and the second-stage would be households in selected *EA*s.

10. It should be pointed out that in practice, however, the target population is somewhat smaller than the population forming the universe. It is usual to restrict the target population for a number of reasons. For instance, in some surveys, some military households in barracks may be excluded from the survey. In labour-force surveys, children below a specified age may be shown as members of households surveyed, but would not be part of the labour-force.

11. It is important to note that when the actual population differs from the target population, the results will apply to the particular population from which a sample was drawn. As discussed in chapter 4, comprehensive and mutually exclusive frames should be constructed for every stage of selection.

2.1.3. Information to be collected

12. From the list of questions requiring statistical answers, a list of items that could provide factual information bearing on issues under investigation can be produced. It is always important to bear in mind that some of the required data could be available from existing sources. In producing the list of items, provision should be given for the inclusion of supplementary items that are correlated with the main items. In a survey of Employment and Earnings, for example, supplementary information on age, sex and education may be gathered. Such information would give additional insight into related questions and thus enrich the analysis.

13. We may add that a tabulation plan should be produced at the time of planning the survey. The blank tables should be circulated for comments and improvement.

2.1.4 Survey budget

14. The survey budget indicates the financial requirements of the survey which is to be conducted. The budget is necessary to support and guide the implementation of the survey and the construction of the timetable for producing the survey results. Cost estimates must be as detailed as possible. It is therefore necessary to understand all the detailed steps involved in the survey operation. The budget shows cost of personnel, equipment and all other items of expense. If there is a pre-determined ceiling (which is usually the case) of funds available, the overall survey budget must be within the pre-determined framework. It is also advisable to follow the general guidelines of the financing agency in preparing the budget. This may facilitate the approval of the budget estimates. If there is need to depart from the prescribed budget, authority must be sought from the relevant organization(s). The financial requests of the survey should be prepared at an early stage. In general, the budget will depend largely on the survey design, precision required and geographical coverage. Table 2.1 below illustrates a possible survey budget.

TABLE 2.1. Illustration of a cost worksheet for a household survey programme

Activity	*Estimated units of work (person-months except where otherwise indicated)*	*Unit cost (relevant unit of currency per person-month, except where otherwise indicated)*	*Estimated total cost relevant unit of currency)*
I. PLANNING AND PREPARATORY ACTIVITIES			
A. *Initial planning and subsequent monitoring (senior staff)* ……			
B. *Selection and specification of subject-matter*			
1 Subject-matter planning			
2 Preparation of tabulation plans			
3 Secretarial and other services			
C *Development of survey design*			
1. Initial design planning: survey structure, population coverage, sampling procedures, data collection methods etc. (professional staff)			
2. Development of sampling materials:			
(a) Cartographic materials (assumes census materials available):			
Personnel costs			
Maps and supplies			
(b) Field household listings (2,000 enumeration areas):			
Personnel costs (mainly interviewers)			
Travel costs			
(c) Sample selection and preparation from field lists			
D. *Design and printing of questionnaires and other*			
1 Professional staff			
2 Secretarial and other services			
3 Printing costs (after pre-tests)			
E. *Pre-testing*			
1 Professional staff planning:			
(a) Initial preparations			
(b) Analysis of results and revision of materials			
2. Field supervisor:			
(a) Personnel costs			
(b) Travel costs			
3. Interviewers:			
(a) Personnel costs			
(b) Travel costs			
F. *Preparation of instructional and training materials for field use*			
1. Professional staff			
2. Secretarial and other services			
3. Reproduction costs			
G. *Miscellaneous planning activities (for example, public relations and publicity)*			
H. *Subtotal components*			
I. Senior staff			
2. Professional staff			
3. Technical staff			
4. Service staff			
5. Travel			
6. Printing			
7. Cartography and miscellaneous			
SUBTOTAL			
II. FIELD OPERATIONS			
A. *Training of field supervisors*			
1. Personnel costs			
2. Lodging and meals			
3. Travel costs			

B. Training of interviewers
1. Supervisor costs .....
2. Interviewer costs:
   (a) Personnel costs .....
   (b) Travel costs .....

C. Data collection (including quality control)
1. Supervisor costs:
   (a) Personnel costs .....
   (b) Travel costs .....
2. Interviewer costs .....

D. Field administration
1. Field direction .....
2. Travel .....
3. Other costs (for example, control and shipment of materials) .....

E. Subtotal components
1. Professional staff .....
2. Technical staff .....
3. Service staff .....
4. Travel .....
5. Travel subsistence .....
6. Interviewing .....
7. Miscellaneous .....

SUBTOTAL

III. DATA PROCESSING

A. Systems planning

B. Computer programming

C. Clerical coding
1. Initial coding .....
2. Quality control .....
3. Supervision .....

D. Key-to-disk operations
1. Initial keying .....
2. Quality control .....
3. Supervision .....

E. Computer time (including operator and maintenance costs)

F. Miscellaneous processing costs (supplies etc.)

G. Subtotal components
1. Professional staff
2. Technical staff
3. Quality control staff
4. Service staff
5. Computing
6. Miscellaneous

SUBTOTAL

IV. DATA REVIEW AND PUBLICATION

A. Professional time .....

B. Publication costs .....

V. SURVEY DIRECTION AND CO-ORDINATION (continuing oversight over all activities) .....

VI. SUBTOTAL .....

VII. EVALUATION STUDIES AND METHODOLOGICAL RESEARCH (may be estimated at 10 per cent of cumulative total)

VIII. GENERAL OVERHEAD (may be estimated at 15 per cent of cumulative total for administrative costs, space rental, general supplies and the like) .....

IX. TOTAL

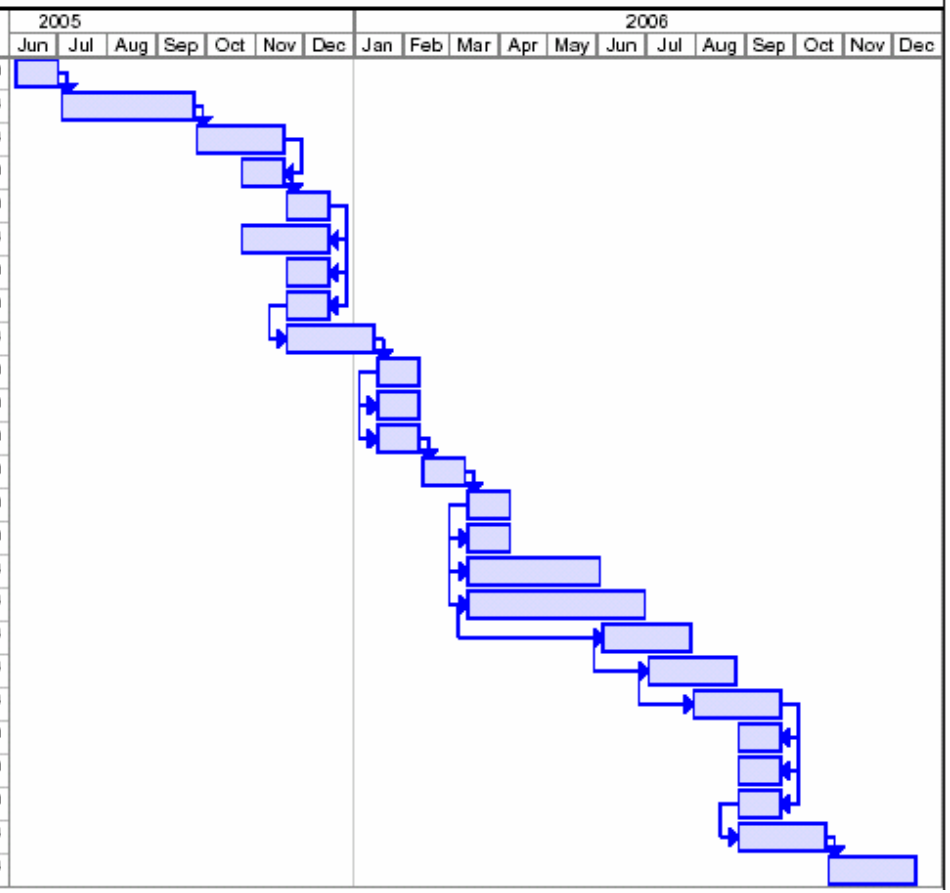
Sourrce: Handbook of household surveys (revised edition) St/ESA/StTAT/SER.F/31, United Nations, New York 1984

15. It is essential that an effective cost control system be established in the organization that is conducting the survey. In most large scale survey operations, chances are high of losing control of monitoring the disbursement of funds once field work starts. In such circumstances a large amount of funds tend to be channeled in areas unrelated to the major survey operations. Judicious cost control helps to monitor actual expenditures in relation to estimated costs and actual work accomplished. It is imperative that management responsible for the survey should ensure accountability of funds. This greatly enhances the credibility of the survey organization.

The following graphic is an illustration of a timetable.

Time-table of Household Survey Activities for Country X

ID	Task Name	Duration
1	Meeting with stakeholders	1 mon
2	Preparatory activities	3 mons
3	Initial questionnaire design	2 mons
4	Send questionnaires to user committee members	1 mon
5	Include in questionnaire agreed suggestions	1 mon
6	Draft interviewer's & supervisor's manuals	2 mons
7	Print questionnaires & manuals	1 mon
8	Make plans for pre-test	1 mon
9	Train interviewers and supervisors	2 mons
10	Conduct the pre-test	1 mon
11	Revise questionnaire (if need be)	1 mon
12	Revise manuals (if need be)	1 mon
13	Sample design	1 mon
14	Design & test data entry programme(s)	1 mon
15	Design & test data cleaning programme(s)	1 mon
16	Data collection	3 mons
17	Data entry	4 mons
18	Design & test tabulation plan	2 mons
19	Data cleaning (eg. range and consistency checks)	2 mons
20	Assignment of chapters to authors	2 mons
21	Carry out the processing according to plan	1 mon
22	Calculate sampling errors among other estimates	1 mon
23	Distribution of tabulations to authors	1 mon
24	Analysis and report writing	2 mons
25	Compilation & tabulation of final report	2 mons



2. 2. Execution of surveys

2.2.1 Data collection methods

16. There are a number of methods used in data collection, among them, direct observation and measurement; mail questionnaire; telephone and personal interview.

Direct observation and measurement: is the most ideal method as it is usually more objective. It is free from memory lapse and subjectivity of both respondents and interviewers. Examples of areas where direct observation has been used are:

a. Some aspects of food consumption surveys;
b. Price collection exercises, where enumerators can purchase the produce and record prices.

17. This method, though useful, has a snag of being expensive both in terms of resources and time. In most cases, interviewers have to use some equipment. Experience has shown that the method of direct observation and measurement tends to be useful and practical when the sample sizes or populations are relatively small.

Mail questionnaires: the use of mail questionnaires is fairly cheap and quick. The major cost component at the data collection stage is postage. After the questionnaire is designed and printed it is mailed to respondents. In this case the respondents are assumed to be literate as they are expected to fill the questionnaire on their own. This may be an erroneous assumption especially in developing countries where literacy levels are still low. The major weakness of this method is the high non-response rates associated with it. This may be due to the complexity of questionnaires used. However, apathy cannot be completely ruled out. In some cases there is good questionnaire response but high item non-response.

18. In trying to improve the response rate, some reminders have to be sent to non-respondents. However, it is advisable to select a sub-sample of the non-respondents and cover them by the personal interview method. This may be necessary because the characteristics of the non-responding units may be completely different from those that responded. In this case the responding and non-responding units are treated as two post-strata that have to be differentially weighted when preparing the estimates (see more about survey weighting in subsequent chapters particularly chapter 6). In order to increase the response rate, the mailed questionnaires should be attractive, short and as simple as possible. Enclosing stamped and addressed returns may help to improve the response rate.

19. In order to use this method satisfactorily, there must also be a sampling frame that is as current as possible. Thus addresses of the respondents must be up-to-date. The survey organization must also be convinced that respondents are capable of completing the questionnaires on their own.

20. Here is a summary of some of the advantages and limitations of mail questionnaire surveys:

Advantages:

a. It is cheaper;
b. Sample can be widely spread**;**
c. Interviewer bias is eliminated;
d. It is quick**.**

Limitations:

a. Non-response is usually high;
b. The answers to the questions are taken at their face value as there is no opportunity to probe;
c. If it is an attitude survey, it is difficult to ascertain whether the respondent answered the questions unaided;
d. The method is useful only when the questionnaires are fairly simple, and, therefore, it is not a suitable method for complex surveys.

Personal interview method: this method is the most common in collecting data through large scale sample surveys in developing countries. Apart from the usually high response rate resulting from personal interviews, the method is appropriate because of the prevailing high illiteracy rates in some of these countries. The method entails interviewers going to selected respondents collecting information by asking questions. The main advantage of this method is that the interviewers can persuade (through motivation) respondents to answer questions and can explain the objectives of the survey. Further, in using the personal interview method, there is greater potential for collecting statistical information on conceptually difficult items which are likely to yield ambiguous answers in a mailed questionnaire.

21. The following are some of the limitations in using the personal interview method:

a. Different interviewers may give different interpretations to the questions, thereby introducing bias in the survey results as very few interviewers consistently refer to the instructions manual;
b. In the process of probing, some interviewers may suggest answers to respondents.
c. Personal characteristics of interviewers may influence attitudes of respondents, for example, age, sex at times even race;
d. Interviewers may read questions wrongly because of the divided attention of interviewing and recording.

22. Collectively the limitations listed above are the main sources of so-called interviewer bias, studies of which have shown can cause serious nonsampling errors in surveys.

23. The following points should be taken into consideration when asking questions to respondents:

a. The interviewer should clearly understand the purpose of each question as explained in the interviewers' manual. It is important that interviewers constantly refer to the manual.
b. Experience has shown that it is better for the interviewer to follow the sequence of questions in the questionnaire. In most questionnaires careful thought is given in the ordering of questions, taking into consideration motivation of respondents, linkage of topics, facilitating memory of the respondent's past events, and careful of the most sensitive questions.
c. Interviewers should by all means refrain from suggesting answers to respondents.
d. All questions should be asked. In this way, item non-response is minimized. Further, no item in the questionnaire should have a blank space unless it satisfies the skip pattern. If a question is not relevant to a particular respondent, then a comment should be included. Such an approach assures the survey manager that all questions included in the questionnaire have been administered.

2.2.2 Questionnaire design

24. Once the survey objectives and tabulation plan have been determined, the relevant questionnaire can be developed. The questionnaire plays a central role in the survey process in which information is transferred from those who have it (the respondents) to those who need it (the users). It is the instrument through which the information needs of the users are expressed in operational terms as well as the main basis of input for the data processing system for the particular survey.

25. The size and format of the questionnaire need very serious consideration. It is advisable to design questionnaires at the time of planning for the survey. If the questionnaires have to be mailed to respondents, they have to be attractive and simple. This may increase the response rate. On the other hand a questionnaire to be used in the field for recording responses by interviewers should be sturdy to survive the field handling.

26. The questionnaire so designed should be ideal in facilitating the collection of relevant and accurate data. In order to enhance accuracy in the survey data, special consideration should be made in ordering the sequence and in the wording of items in the questionnaire. The respondent has to be motivated. The questionnaire has to be well spread out to facilitate easy reading of questions either by the respondent or the interviewer. We cannot overemphasize that every questionnaire should have clear instructions.

27. Special care, therefore, should be taken by the survey team in giving precise definitions of the data to be collected and the translation of data requirements and related concepts into operational questions. In this connection pre-testing of the questionnaire becomes a usual and generally a necessary activity to undertake, unless it has been fully validated in prior surveys.

28. In summary a good questionnaire should:

a. Enable the collection of accurate information to meet the needs of potential data users in a timely manner;
b. Facilitate the work of data collection, data processing and tabulation;
c. Ensure economy in data collection, that is, avoid collection of any non-essential information.
d. Permit comprehensive and meaningful analysis and purposeful utilization of the data collected.

29. This implies that survey questionnaires must be developed so as to yield information of the highest quality possible with special emphasis on relevance, timeliness and accuracy. This must be accomplished efficiently, minimizing the cost and burden involved in the provision of the necessary information.

2.2.2.1 Question construction

30. Open and closed-ended questions are used in sample survey questionnaires. In an open-ended question, the respondent gives his/her own answer to a question. In an attitudinal survey we may ask respondents to define what they consider is good quality of life. Obviously different respondents will define in their own way what constitutes quality of life. On the other hand, a closed-ended question restricts the respondent to select answers from a list already given by the survey team. The following are examples of closed-ended questions:

Do you have any permanent mental disabilities that limit you in daily activities?

Yes ☐ No ☐

How do you evaluate your capacity to see (even with glasses or contact lenses, if used)?

1. Unable
2. Severe permanent difficulty
3. Some permanent difficulty
4. No difficulty

31. The advantages of using closed-ended questions are that: (i) they yield more uniform responses and (ii) they are easy to process. The main limitation of such questions is that the designer of the survey has to structure the possible answers. In such a case, important possible responses may be overlooked. In most surveys, complex issues and questions pertaining to attitudes and perceptions that may not be known are best handled by open-ended questions.

2.2.2.2 Question wording

32. The questions should be clear, precise and unambiguous. The respondent should not be left to guess what the interviewer wants out of him/her. The use of definitions and concepts may seem obvious to the survey manager while this *may* not be so to the respondent. This way, a

respondent may use discretion when answering questions. The end result is a proliferation of nonsampling errors in survey results. Consider a simple example. The question, "What is your home address?," creates confusion in many African countries, especially for the urban population, unless "home" is clearly defined. There are respondents who take "home" to mean the village they originally come from.

2.2.2.3 "Loaded" questions

33. A so-called loaded question persuades a respondent to answer a question in a certain way. This means that the question tends to be biased in favour of a certain answer. Here is an example of a loaded question in a health survey: "How many days in a week do you drink more than two bottles of beer"? This question courts the respondent into admitting that he/she drinks beer, above all, not less than two bottles a day. Such questions tend to bias answers of respondents. It is important to be mindful of avoiding creating data but rather simply collecting data.

2.2.2.4 Relevance of questions

34. The purpose of a questionnaire is to enlist information that would be used in studying the situation. It is therefore imperative for the survey organization to ask relevant questions in order to obtain a true picture of a particular situation under study. The questions included in a questionnaire should be relevant to most respondents. For instance, it is pointless to administer a questionnaire cluttered with questions on individual achievement with regard to higher university education in a typical rural environment of most African countries today. Similarly, it is not appropriate in a fertility survey to include females of, say, age 10 or under, and ask them questions on number of children ever born, whether married, divorced or widowed. These questions would be relevant to females above a certain age, but not to girls who are less than child bearing age.

2.2.2.5 Question sequence

35. The order of items in a questionnaire should try to motivate and facilitate recall in the respondent and help to solicit accurate information. It is suggested that the first questions should be easy, interesting and not sensitive. This builds up the confidence of the respondent to carry through the interview which in most cases he/she provides voluntarily. It has also become fairly standard that a general sequence in household surveys begins with questions that identify the sample unit, such as address, followed by those which describe the household and the individuals in the household, such as demographic characteristics. Finally, the detailed questions that constitute the main subject of the survey are asked[1]. In general, sensitive questions must be among the last questions to ask. We should emphasize that there must be a logical link in questions, especially those that are contingent.

2.2.3 Tabulation and analysis plan

[1] Handbook of Household Surveys (revised edition) St/ESA/STAT/Ser.F/31, New York, 1984.

36. A useful technique to assist the survey designer in bringing precision to the user's need for information (set of questions or objectives of the survey) is to produce tabulation plans and dummy tables. Dummy tables are draft tabulations, which include everything except the actual data. As a minimum the tabulation outline should specify the table titles, column stubs, identifying the substantive variables to be tabulated, the background variables to be used for classification, and the population groups (survey objects or elements or units) to which the various tables apply. It is also desirable to show the categories of classification in as much detail as possible, though these may be adjusted later when the sample distribution over the response categories is better known.

37. The importance of a tabulation plan can be viewed from a number of perspectives. One is that the production of dummy tables will indicate if data to be collected will yield useable tabulations. They will not only point out what is missing, but also reveal what is superfluous. Furthermore, the extra time that is spent on producing dummy tables is usually more than compensated for at data tabulation stages by reducing time spent on the design and production of actual tables.

38. There is also the close relationship between the tabulation plan and the sampling design employed for a survey. For example, geographical breakdown in the tables is only possible if the sample is designed to permit such breakdown. Also, the sample size may make it necessary to limit the number of cells in the cross-tabulations to avoid tables which are too sparse. Sometimes the plan might have to be modified during the tabulation work. Categories might have to be combined in order to reduce the number of empty cells; or, interesting findings in the draft data may prompt new tables. More generally, the way in which the data collected in the household survey will be used to answer the questions (attain the objectives) can be referred to as the 'data analysis plan.' Such a plan explains in detail what data are needed to attain the objectives of the survey. Survey designers must refer to it constantly when working out the details of the survey questionnaire. It perhaps goes without saying that the analysis plan should also be the main reference point to guide the analysis of the survey results.

2.2.4 Implementation of field work

39. In most developing countries, the implementation of field work is often seriously constrained by lack of resources. However, if a survey is to be carried out field work should be properly organized and implemented in order to utilize the limited resources at the disposal of the survey team efficiently. For the survey operations to succeed, the conceptual aspects of the survey subject matter should be clearly understood by those involved in designing the survey operations. Further, interviewers must thoroughly master the practical procedures that may lead to the successful collection of accurate data. In order for the survey operations to be successfully realized, there is always a need to have a well organized and effective field organization.

2.2.4.1. Equipment and Materials

40. In many developing countries it is necessary, that well in advance, equipment such as vehicles, boats, bicycles, etc are available and in working condition. It is also necessary to have

some spare parts. Vehicles and bicycles facilitate quick mobility of team leaders and supervisors/interviewers, respectively.

41. Adequate materials, like folders, clipboards, pencils, pencil sharpeners, notebooks and fuel (for vehicles) should be available in adequate supplies for use during the survey operations.

2.2.4.2 Management of survey operations

42. A large scale sample survey is usually a demanding and complex operation. Therefore the need for judicious, effective and efficient management of activities at various levels cannot be overemphasized.

43. There must be a clear and well defined line of command from the survey manager to the interviewer. It should be noted that control forms for monitoring progress of the survey have been found useful.

2.2.4.3 Publicity

44. Some surveys have had limited success partly due to high non-response owing to refusals. It is, therefore, incumbent upon survey organizers to mount some publicity campaigns for the survey. Experience has shown that publicity plays an important role in soliciting cooperation from respondents, even though some funding organizations/agencies consider expenditures on publicity as a waste of resources.

45. Different approaches to publicity can be adopted depending on prevailing circumstances. For example in some countries, in the urban areas radio, television and news paper messages can complement posters. While in the rural areas, radio messages and posters could be used.

46. Further, it may be necessary to arrange meetings with local opinion leaders in selected areas. During such meetings people would be briefed on the objectives of the survey. In addition the leaders should be requested to persuade people in their respective areas to provide requisite information to the interviewers.

47. Before going into the field it is important that the relevant legal provision for conducting the survey be published. The announcement should, among other information, give the survey objectives, duration and topics to be covered.

2.2.4.4 Selection of interviewers

48. An interviewer is at the interface with the respondents. He/she is the representative of the survey organization who is always in contact with the respondent. This is a clear indication of why an interviewer's job is so crucial to the success of the survey programme. The selection of an interviewer should, therefore, be given great consideration and care. An interviewer should be capable of effectively communicating with the respondent. He/she should have qualities of enlisting all the information with accuracy within a reasonable given time.

49. Depending on the type of survey, an interviewer should have an adequate level of education. In addition, an interviewer should be able to record information honestly, without "cooking *figures"*. The selected interviewers should follow instructions and use definitions and concepts as provided for in interviewer's field manual.

50. The following selection procedures may assist in selecting suitable interviewers:

a. It is suggested that the prospective interviewers should complete an application, indicating his/her age, marital status, current address, educational attainment and employment history;
b. Those initially selected may be subjected to an intelligence test and an additional test in simple numerical calculations;
c. Apart from written tests, there is usually a need to interview the candidates. The interviews should be conducted by a panel who will independently rate the candidates. Some of the attributes to be considered in rating the candidates are the following: friendliness, interest in work, expression and alertness.

51. Field work can be tedious, with problems of travel over difficult terrain; therefore an interviewer so selected should be committed and prepared to work under difficult conditions.

2.2.4.5 Training of interviewers

52. The selected interviewers should be thoroughly trained before being sent into the field. The main purpose of a training programme is to bring about uniformity in the interviewing procedures of the survey. This is necessary of course to avoid differing interpretations of the definitions, concepts and objectives of the survey by interviewers and hence to minimize interviewer bias.

53. Qualified instructors should be responsible for the training. Such instructors must obviously be well versed in the aims and objectives of the survey. Preferably, they should be part of the survey team carrying out the survey.

54. The interviewers should be carefully instructed on the purposes of the survey and how the results are going to be used. In order for the interviewers to be properly apprised of the objectives of the survey, they have to be well trained in the concepts and definitions used in the questionnaire.

55. As part of the training process, the interviewers, in the presence of the instructor, should take turns in explaining to others the various items in the questionnaire. Practical sessions should be arranged both in class and in the actual field situation. For example, interviewers could take turns in asking questions to each other in a classroom setting, followed by a field trip in a nearby neighborhood, where a few households could be interviewed by the trainee interviewers. The instructor should always be present to guide and correct the interviewers. After the field interviews, the trainees should discuss the results under the guidance of the instructor. The training programme should result in a decision by the survey manager of which

trainees may require additional training and whether any of them are entirely unsuited for the job.

2.2.2.6 Field supervision

56. It is generally agreed that training is a precursor to effective and successful field work. However, training without proper supervision may not yield the desired results. The success of field work requires dedicated, continuous and effective supervision by superior staff that are more experienced and better qualified than interviewers. Supervisors should undergo training in all aspects of the survey. It cannot be overemphasized that the supervisor is an important link between the data gathering organization and the interviewer. The supervisor is supposed to organize work for interviewers by determining field assignments and locations. The supervisor reviews completed work and maintains a high level of commitment to the survey programme by the interviewers. We advocate that, if possible, there should be a relatively high ratio between the supervisory staff and the interviewers. The ratio of one supervisor to four or five interviewers has been suggested as ideal for most household surveys. However, this is just a guide.

2.2.2.7 Follow-up of non-respondents

57. In most surveys, there are bound to be cases of non-response(refer to chapter on nonsampling error). Some respondents refuse to co-operate with the interviewers, while in some cases, certain items in the questionnaire are not attended to. When a non-responding unit has been reported to the supervisor, he has to contact the sample unit and try to solicit for the information, owing to his better qualifications and more experience. Since an operational goal in any survey is to achieve the highest possible response rate, it is recommended to collect information from a sub-sample of the initial non-respondents. In this case, the survey effort is then re-directed to the sub-sample preferably using supervisors as interviewers.

2.2.2.8 Reducing non-response

58. It is important in designing and executing a household survey to develop good survey procedures aimed at maximizing the response rate. We emphasize the importance of having procedures in place to reduce the number of refusals, such as arranging to return to conduct an interview at the convenience of the respondent. Also, the objectives and uses of the surveys should be carefully explained to reluctant respondents to help win their cooperation. Assurance of confidentiality can also help alleviate fear respondents may have about the use of their responses for purposes other than those stipulated for the survey.

59. Repeated callbacks should be made when no one is at home. These should be done at different times of the day. It is recommended that as many as four callbacks should be attempted.

60. It is also important to avoid the problem of inability to locate the selected sampling units, which can be an important source of non-response. This problem is best addressed by

using the most current sampling frame as possible, and that topic is discussed in detail in chapter 4.

References and further reading

Kiregyera, B. (1999), *Sample Surveys: with Special Reference to Africa*, PHIDAM Enterprises, Kampala.

Statistics Canada. (2003) *Survey Methods and Practices*, Ottawa.

United Nations Statistics Division (1982), *Non-sampling Errors in Household Surveys: Sources, Assessment and Control, National Household Survey Capability Programme*, United Nations, New York.

_____ (1984), *Handbook of Household Surveys*, revised edition ST/ESA/SER.F/31, United Nations, New York.

_____ (1998), *Principles and Recommendations for Population and Housing Censuses*, Revision 1, ST/ESA/STAT/SER.M/67/REV.1, New York

United Nations (2002), *Technical Report on Collection of Economic Characteristics in Population Censuses*, Statistics Division, Department of Social and Economic Affairs and Bureau of Statistics, International Labour Office, New York and Geneva.

Zanutto, E. and Zaslavsky, A. (2002), "Using Administrative Records to Improve Small Area Estimation: An Example from the U.S. Decennial Census," *Journal of Official Statistics*, Statistics Sweden, Vol. 18, No.4, pp.559-576.

Chapter 3

Sampling Strategies

3.1 Introduction

1. While the preceding chapter on survey planning gave a general overview of the various phases of household survey operations, this chapter is the first of several that concentrate solely on sampling aspects – the principal focus of the handbook. This chapter briefly discusses probability versus non-probability sampling and argues why the former should always be used in household surveys. Considerable attention is given to sample size, the many parameters that determine it and how to calculate it. Techniques for achieving sampling efficiency in household surveys are presented. They include stratification, cluster sampling and sampling in stages, with special emphasis on two-stage sample designs. Various sampling options are provided and two major sample designs that have been used in many countries are described in detail. The special topics of (a) sampling in two phases to reach "rare" populations and (b) sampling to estimate change or trend are also discussed. Finally, a concluding summary of recommendations is given at the end of the chapter.

3.1.1 Overview

2. 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, therefore, focuses on multi-stage sampling strategies.

3. 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. At the same time it must avoid 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 balanced. The sample size must also address the urgent needs of users who desire data for sub-populations or sub-areas – *domains*. The sample design must seek maximum accuracy in two important ways. First, the *sample frame* that is used (or constructed) must be as complete, correct and current as possible. Second, sample selection techniques that minimize unintentional bias sometimes caused by the implementers should be used. 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.

4. 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. In general the emphasis is on

national surveys, though all the techniques described can be applied to large, sub-national surveys such as those restricted to one or more regions, provinces, districts or cities. Because of the crucial importance of sample frames in achieving good sample practice a separate chapter, following, is devoted to it.

3.1.2 Glossary of sampling and related terms

5. We begin with a glossary of terms, in Table 3.1, used in this chapter and the next. The glossary is not intended to provide formal definitions of sampling terms, some of which are mathematical. Instead, it gives the use of terms in the context of this handbook, focussing of course on household survey applications.

Table 3.1. Glossary of Sampling Terminology Used in Chapter 3

TERM	USAGE
Accuracy [validity]	[See nonsampling error]
Area sampling	Selection of geographical area units that comprise sampling frame (may include selection of area *segments*, defined as mapped sub-divisions of administrative area)
Canvassing	Method of "covering" a geographical area to locate dwellings, usually applied in operations to up-date sample frame
Cluster sampling	Sampling in which next-to-last stage is geographically-defined unit such as census enumeration area (*EA)*
Cluster size	(Average) number of sampling units – persons or households – in cluster
Clustering; clustered	Refers to tendency of sample units – persons or households – to have similar characteristics
Confidence level	Describes degree of statistical confidence with which precision or margin of error around the survey estimate is obtained, 95 per cent generally being regarded as the standard
Complex sample design	Refers to use of multiple stages, clustering and stratification in household survey samples, as opposed to simple random sampling
Compact cluster	Sample cluster consists of geographically contiguous households
Design effect – *deff*	Ratio of variance from complex sample design to simple random sample of same sample size; *deft* is ratio of standard errors; sometimes referred to as *clustering effect* though *deff* includes effects of stratification as well as clustering

Domain	Geographical unit for which separate estimates are to be provided
Dummy selection stage	A pseudo-stage of selection intended to simplify the manual task of identifying sub-areas where sample clusters will ultimately be located
EPSEM sample	Sampling with equal probability
Implicit stratification	Means of stratifying through geographic sorting of sample frame, coupled with systematic *pps* sampling
Intraclass correlation	Degree to which two units in cluster have same value, compared to two units selected at random in population
List sampling	Selection from a list of the units that make up the frame
Master sample	A "super" sample intended to be used for multiple surveys and/or multiple rounds of the same survey, usually over 10-year time frame
Measure of size, *MOS*	In multi-stage sampling a count or estimate of the size (eg., number of persons) of each unit at a given stage
Non-compact cluster	Sample cluster consists of geographically dispersed households
Non-probability sampling	See text descriptions of examples of these methods: quota, judgmental, purposive, convenience, random walk sampling
Nonsampling error	Bias in survey estimate arising from errors in design and implementation; refers to *accuracy* or *validity* of an estimate as opposed to its reliability or precision
PPS sampling	Selection of first (second, etc.) stage units in which each is chosen with probability proportional to its measure of size; see also *ppes* sampling in text – probability proportional to *estimated* size
Primary sampling unit, *PSU*	Geographically-defined administrative unit selected at first stage of sampling
Probability sampling	Selection methodology whereby each population unit (person, household, etc.) has known, non-zero chance of inclusion in the sample
Quick counting	Refers to up-dating operation when dwellings are roughly counted to provide current measure of size; see also *canvassing*
Reliability [precision, margin of error]	Refers to degree of sampling error associated with a given survey estimate

Relative standard error [coefficient of variation]	Standard error as percentage of survey estimate, i.e. standard error divided by estimate
Sample size	Number of households or persons selected
Sample frame(s)	Set of materials from which sample is actually selected, such as a *list* or set of *areas*
Sampling error [standard error]	Random error in survey estimate due to the fact that a sample rather than entire population is surveyed; square root of sampling variance
Sampling in phases; also known as double sampling or post-stratified sampling	Selecting sample in (generally) two time periods, with second phase typically a sub-sample of first-phase sample; not to be confused with *trend sampling* (see below)
Sampling in stages	Means by which sample of administrative areas and households/persons is chosen in successive stages to pinpoint geographic locations where survey is conducted
Sampling variance	Square of standard error or sampling error
Segment	A delineated, mapped sub-division of a larger cluster
Self-weighting	Sample design where all cases have same survey weight
SRS	Simple random sample (rarely used in household surveys)
Stratified sampling	Technique of organizing sample frame into sub-groupings that are internally homogeneous and externally heterogeneous to ensure sample selection is "spread" properly across important population sub-groups
Sub-segmentation [chunking]	Usually, a field exercise in which unexpectedly large clusters are sub-divided to decrease listing workload
Systematic sampling	Selection from a list, using a random start and predetermined selection interval, successively applied
Target population	Definition of population intended to be covered by survey; also known as *coverage universe*
Trend sampling	Sample design to estimate change from one time period to another
Weight	Inverse of probability of selection; inflation factor applied against raw data; also known as *design weight*

3.1.3 Notations

6. Standard notations have been used in this and subsequent chapters of the handbook. In general, the upper case letters denote population values and lower case letters denote sample observations. For example, $Y_1, Y_2, Y_3, ----, Y_N$ denote population values. On the other hand $y_1, y_2, y_3, ----, y_n$ are usually used to denote sample values. It is apparent from the above that N stands for population size while n stands for the sample size. It is important to note that population parameters are denoted by either the upper cases of the English alphabet or by Greek letters. For example, $\bar{Y}$ and σ denote the population mean and standard deviation, respectively. The notations for estimators of population parameters are shown with the symbol $\wedge$, on top of the notation, or by lower case letters. The following are examples denoting the notations for sample mean, $\hat{\bar{Y}}$ or $\bar{y}$.

Table 3. 2 Selected notations used for population values and sample characteristics

Characteristic	Notation for		
	Population	Sample	
		Sample estimates with a symbol $\wedge$ on top of the notation.	Sample observations and estimates with lower cases.
Total units	N	n	n
Observations	$Y_1, Y_2, --, Y_i, --, Y_N$	$Y_1, Y_6, --, Y_j, --, Y_n$	$y_1, y_2, --, y_i, ---, y_n$
Sampling faction	-	f	f
Mean value	$\bar{Y}$	$\hat{\bar{Y}}$	$\bar{y}$
Proportion	P	$\hat{P}$	p
Parameter/ estimator	θ	$\hat{\theta}$	$\hat{\theta}$
Variance of y	$\sigma^2(y)$	$\hat{\sigma}^2$ (y)	$s^2(y)$
Standard deviation of y	$\sigma(y)$	$\hat{\sigma}(y)$	s (y)
Intra-class correlation	δ	$\hat{\delta}$	$\hat{\delta}$
Ratio	R	$\hat{R}$	r
Summation	$\sum_{i=1}^{N}$	$\sum_{i=1}^{n}$	$\sum_{i=1}^{n}$

3. 2 Probability sampling versus other sampling methods for household surveys

7. 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 recommended for household surveys.

3.2.1 Probability sampling

8. 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 (1) that each element must have a known mathematical chance of being selected, (2) that chance must be greater than zero and (3) it must be 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.

9. 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. None of these features are present when non-probability sampling methods are used. Because of these aspects of probability sampling it is strongly recommended that it always be used in household surveys. This is the recommended approach even when survey costs are greater than those of non-scientific, non-probability methods.

3.2.1.1 Probability sampling in stages

10. 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. The last stage involves selecting the specific households or persons to be interviewed. Those two stages and any intervening ones must utilize probability methods for proper sampling. For illustration, 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 the listing a systematic selection of 1 in every 5 is made for the survey interview, no matter how many households are listed 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.

11. Though not particularly efficient, the sample design of the example above 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 probabilities of selection at the data analysis stage of the survey operation (see discussion of survey weighting in chapter 6).

3.2.1.2 Calculating the probability

12. 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. If some of them, however, 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. See the next subsection for ways of handling the situation when areas are excluded from the survey.

13. 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. Thus 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).

14. 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 ith village, the probability would be equal to m_i/M_i. 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. It would be superfluous, therefore, 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.

3.2.1.3 When target population is ill-defined

15. 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 often occurs 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 important sub-groups of those special populations 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.

16. 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 more accurately. This is important 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 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.

17. 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.

3.2.2 Non-probability sampling methods

18. In contrast to probability sampling, there is no statistical theory to guide the use of non-probability samples. They 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 and to construct confidence intervals around the latter. Biased estimates can also be made with probability sampling under certain conditions, such as when it is desirable to make survey population distribution agreeable with other controls (see further discussion on this point in chapter 6).

19. In spite of their theoretical deficiencies, non-probability samples are frequently used in various settings and situations. The justification offered by practitioners is generally because of cost, convenience or 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 reasons they should not be used

3.2.2.1 Judgmental samples

20. 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 is based on opinions of which districts are typical or representative in some sense or context.

21. The main difficulty with this type of sample is that what constitutes a representative set of districts is subjective. Ironically, it is also highly dependent on the *choice* of experts themselves. With probability sampling, by contrast, the districts would first be stratified using, if necessary, whatever criteria the design team wanted to impose. Note that the stratification criteria may even be *subjective*, although there are guidelines for applying more objective criteria (see subsection on stratification). Then a probability 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. *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 (even when the strata themselves are subjectively defined). 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 calculating the probability of those that do happen to be selected.

3.2.2.2 Random walk and quota sampling

22. 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. The technique is often used 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. The latter is 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 random 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.

23. 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. They also claim 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.

24. Theoretically, given the conditions set forth in the last two sentences above, a probability sample is thus attainable. In practice, however, it is dubious whether it is ever actually achieved. It usually fails due 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.[2] 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 favour of a household next door which does not present such problems.

25. By substituting non-responding households 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 probability sample. In the latter case interviewers are generally required to make several callbacks to households where its members are temporarily unavailable. Moreover, interviewers are usually trained, for probability-based surveys, to make extra efforts to convince reluctant households to agree to be interviewed.

3.2.2.3 Convenience samples

26. 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.

3.3 Sample size determination for household surveys

[2] 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. Further, the sample should be selected 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.

27. 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 is 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.

3.3.1 Magnitudes of survey estimates

28. 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.

29. 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 estimate is the unemployment rate, that would be the estimate upon which to calculate the sample size.[3] When there are many key indicators a convention sometimes used is to calculate the sample size needed for each and then use the one that yields the largest sample. 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.

30. 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.

3.3.2 Target population

[3] 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.

31. 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 five years for anthropometric measurements of height and weight.

32. Calculation of the sample size must therefore take into consideration each of the target populations. As earlier mentioned household surveys frequently have 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 the 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.

3.3.3 Precision and statistical confidence

33. In the above paragraphs 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 bigger the sample size must be – and by orders of magnitude. Doubling the reliability requirement, for example, may necessitate *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. Simultaneously they must be careful not to use a relatively small sample size such that the main indicators would be too unreliable for informative analysis or meaningful planning.

34. 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, $|13-17|$.

35. 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 per cent confidence level.

36. 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 (though we do not recommend it) for important indicators. This is because the confidence interval surrounding estimates with greater error tolerances are too wide to achieve meaningful results for most analytical or policy needs. In general we recommend 5-10 percent relative errors for the main indicators, budget permitting. Otherwise, 12-15 percent relative error may be substituted

3.3.4 Analysis groups - domains

37. Another significant factor that has a large impact on the sample size is the number of domains. Domains are generally defined as the analytical sub-groups for which *equally* reliable data are wanted. The sample size is increased, approximately,[4] by a factor equal to the number of domains wanted. This is 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. This is an extremely important point which is often misunderstood by survey practitioners who think, erroneously, that the larger the population the larger the sample size must be.

38. 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.

3.3.4.1. Over-sampling for domain estimates

39. 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.

40. It is important to note two implications of deliberate over-sampling of sub-groups, whether for domains or strata. First, it necessitates the use of compensating survey weights to form the national-level estimates. Second and more importantly, the national estimates are somewhat less reliable than they would be if the sample were distributed proportionately among the sub-groups. Hence, the latter implication is a distinct limitation as well, because of the negative effect of over-sampling on the national-level estimates.

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

3.3.4.2 Choosing domains

41. 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. A 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, instead, in the analysis as major tabulation *categories* as opposed to domains. Then 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 given below 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 8,000 households to meet a specified precision requirement, then 16,000 would have to be sampled if urban and rural were separate domains – 8,000 in each sector. Instead, by treating them as tabulation groups, the *national* sample of 8,000 households would be selected, proportionately, by rural and urban, yielding 4,800 and 3,200 households, respectively, in each of the two sectors. Suppose, further, the anticipated standard error for a 10-percent characteristic, based on the sample of 8,000 households, is 0.7 percentage points. This is the standard error that applies to the national estimate (or to urban and rural separately if 8,000 households were sampled in each domain). For a national sample of 8,000 households selected proportionately by urban-rural, the corresponding standard error for rural would be approximately, 0.9 percentage points, calculated as the square root of the ratio of sample sizes times the standard error of the national estimate, or ($\left[\sqrt{8000/4800}\right] * 0.7$). For urban the standard error would be about 1.1 percentage points, or ($\left[\sqrt{8000/3200}\right] * 0.7$). Another way of evaluating the effect is that standard errors for all rural estimates would be about 29-percent higher ($\sqrt{8000/4800}$) than those for national estimates; for urban they would be about 58 percent higher, ($\sqrt{8000/3200}$).

42. Note that the last sentence of the example applies no matter what the standard error is at the national level. In other words it applies to every estimate tabulated in the survey. It is thus possible to analyze the impact on reliability, prior to sampling, for various sub-groups that might be considered as domains. In this way the survey team would have the information to help decide whether potential domains should be treated as tabulation groups,. As implied before, this means that proportionate allocation rather than equal allocation of the sample would be used. 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,[5] effectively making separate urban and rural domains.

43. 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.

3.3.5 Clustering effects

44. 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 can be measured numerically by the 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 *decrease* the sampling variance, but only to a small degree. By contrast, clustering increases the variance considerably. Therefore, *deff* indicates, primarily, how much clustering there is in the survey sample.

45. 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 useably reliable. Unfortunately, *deff* is not known before a survey is undertaken and can only be estimated afterwards, from the data themselves. Where previous surveys have been conducted or similar ones in other countries, the *deff* values from those surveys might be used as proxies in the calculation formula to estimate sample size. Otherwise, a default value of 1.5 to 2.0 for *deff* is typically used by sampling practitioners.

46. To keep the design effect as low as possible, the sample design should follow these general principles (see also summary guidelines at end of chapter):

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.

[5] That decision would be taken, for example, if the anticipated relative standard errors for (any of) the key urban indicators were greater than, say, 7.5 percent (the 95 percent confidence level would be 15 percent, suggested in this handbook as the maximum allowable).

d. Select a systematic sample of households at the last stage, geographically dispersed, rather than a segment of geographically contiguous households.

47. 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. The sampling design effect 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. 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.

3.3.6 Adjusting sample size for anticipated non-response

48. 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.

49. 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.

50. 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.

3.3.7 Sample size for master samples

51. 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.

52. 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 two considerations. The first is budgetary, as might be self-evident. The second is the anticipated sample sizes of the individual surveys that might be used over the time interval that the master sample will be used. The latter is 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 8,000 households in each of the three years of its operation. In year 5, however, there will be a replacement sample of 4,000 households for half of the 8,000 interviewed in year 1. Similarly, year 8 will replace the remaining 4,000 households from year 1 with 4,000 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. Accordingly, 12,000 households still remain that can be used for other surveys if necessary.

3.3.8 Estimating change or level

53. 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. This is especially true whenever 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. These are discussed in the subsection on sampling to estimate change or trend (section 3.9.2).

3.3.9 Survey budget

54. 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.

55. The sample size is initially calculated by the statistician, taking account of each of the parameters discussed in this chapter. It is often the result, however, that the size 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.

56. It is the responsibility of the sampling technician to help guide the discussion on cost versus precision. He/she should explain the trade-offs that occur from limiting the number of domains (less utility for users) to decreasing precision requirements (less reliability for key indicators), whenever the appropriate sample size has to be decreased because of budget considerations. The discussion should proceed along the lines of the examples given in the subsections above on precision and domains. The number of clusters is also a key component

in the survey costs as well as its precision, which the sampler must carefully weigh in guiding the survey team. See the section below on number of clusters, section 3.5.5, for more discussion.

3.3.10 Sample size calculation

57. In this subsection we provide the calculation formula for determining the sample size, taking into 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.

58. The estimation formula[6] for the sample size, n_h, is

$$n_h = (z^2)\,(r)\,(1\text{-}r)\,(f)\,(k)/\,(p)\,(\breve{n})\,(e^2), \text{ where} \qquad (3.1)$$

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;

$\breve{n}$ is the average household size (number of persons per household);

e is the margin of error to be attained.

Recommended values for some of the parameters are as follows:

59. 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 former is generally regarded as the standard for assigning the degree of confidence desired in assessing the margin of error in household surveys. 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-

[6] The formula for sample size also contains a factor, the so-called finite multiplier, which must be taken into account when the calculated sample size turns out to be a large percentage of the population size. That condition rarely pertains in large-scale, household surveys of the type being considered in this handbook. Accordingly, the finite multiplier is assumed to have a value of 1.0 and is thus ignored in formula (1).

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, $\breve{n}$, 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

$$n_h = (3.84)\ (1\text{-}r)\ (1.2)\ (1.1)/\ (r)\ (p)\ (6)\ (.01). \qquad (3.2)$$

Formula [2] reduces further to

$$n_h = (84.5)\ (1\text{-}r)/\ (r)\ (p). \qquad (3.3)$$

60. The reduced version may be used whenever *all* the recommended default values of the parameters are used in lieu of more precise values available from a country's own experience.

- ***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. It makes up about 65 per cent 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 1,170 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 per cent. 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 (3.2) and (3.3) 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.

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-per cent margin of relative error; then $e = 0.10r$ (or 0.005 standard error). The values for the

expected non-response rate, design effect and average household size are again the ones we have recommended. Then formula (3.3) gives nearly 10,800 households (84.5*0.95)/(0.05*0.15), a much larger sample size than the previous example. 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.

61. 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 per cent, 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 acute rate with 10-percent margin of relative error; then $e = 0.10r$ [7] (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 a little over 1,600 households (84.5*0.95)/(0.05).

62. 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.

63. For countries in which one or more of the assumptions discussed above do not hold, simple substitutions may easily be made in formula (3.1) 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.

64. 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, however, yield much higher sampling errors.

3.4 Stratification

65. 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-

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

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.

3.4.1 Stratification and sample allocation

66. 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 sth stratum. By contrast, a non-stratified sample would yield a sample size for the sth 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 5,000 households, independent selection of the sample by stratum will ensure that 3,000 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 3,000-2000 split, although that would be its expected distribution. The non-stratified sample could, by unlucky chance, produce a sample of, say, 3,200 rural households and 2,800 urban ones.

67. 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 that 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.

68. As implied by the preceding sentence stratification may also provide the means of allocating the sample implicitly, a simpler and more practical method than optimum allocation.[8] In other words, with proportionate sampling by stratum it is not necessary to calculate, in advance, the number of sample cases to allocate to each stratum.

- ***Example***

[8] Optimum allocation refers to allocating on the basis of cost functions. It is not discussed in this manual because it is rarely used in practice in developing countries, which may be due to the lack of firm cost figures for survey operations. The reader may find detailed information on optimum allocation in many of the references given for this chapter.

Suppose an objective of the sample design is to ensure, precisely, proportionate allocation of the total sample size to each of 10 provinces that make up the country. If, say, Province A contains 12 percent of the nation's population then 12 percent of the sample clusters should be selected in that province, provided the expected cluster size is constant. Suppose further that 400 is the total number of clusters to be selected nationwide. A method often used in many countries is to calculate (.12) times (400) = 48, and then *assign* 48 clusters to be selected in Province A. With proper stratification, however, that procedure is unnecessary. Instead, each province should be treated as a separate stratum in the sample selection process. Then, application of systematic *pps* sampling (see glossary in Table 3.1), with a single sampling interval, will automatically result in an expected 48 clusters in Province A. More about this type of stratification and its use in simplifying allocation schemes is discussed below in the section on implicit stratification.

3.4.2 Rules of stratification

69. There are two basic rules applied when stratifying a population. One of the rules is always required. The other should ordinarily 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.

70. 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.

71. 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.

72. An important point to note about *proportionate* selection 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[9] does not damage the survey estimates in terms of their reliability.

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

73. Another important point to note about stratification concerns sampling error estimation. While a single unit selected from each stratum suffices to meet the theoretical requirements of stratified sampling, a *minimum of two* must be chosen for the sample results to be used to calculate sampling errors of the survey estimates.

3.4.3 Implicit stratification

74. 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 multi-purpose 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 sub-populations, socio-economic groups, etc. Because of this geographic property, implicit stratification is also highly useful even when the subject-matter of the survey is focused on a single topic, whether labour force, household economic activity, poverty measurement, health or income and expenditures. The technique is highly recommended for these reasons and also for its simplicity of application.

75. To be applied correctly, implicit stratification requires using systematic selection at the first stage of sampling. The procedure is simple to implement and entails, first, arranging the file of *PSU*s in geographic sequence. In many countries the sequence would likely be urban by province and within province by district, followed by rural by province and within province by district. The next step is systematically selecting *PSU*s from the sorted file. The systematic selection is done either by equal probability (*epsem*) or, more likely, probability proportionate to size (*pps* sampling). *PPS* sampling is discussed further in section 3.6 below.

76. An important advantage of implicit stratification is that it eliminates the need to establish explicit geographic strata, as already mentioned. That, in turn, does away with the need to allocate the sample to those strata, especially when proportionate sampling is used. Another advantage is the simplicity described in the preceding paragraph, 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. An illustration of an implicit stratification scheme with systematic *pps* sampling is shown in Illustration 3.1. below.

Illustration 3.1. 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.*

3.5 Cluster sampling

77. The term, cluster sampling, was coined originally to refer to sample designs in which all members of a group were sampled. 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 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. Then, 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.

78. 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, 1,000 households in 50 locations (20 households per cluster) than 1,000 households 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 more or less similar characteristics. This so-called clustering effect has to be compensated in the sample design by increasing the sample size commensurately.

3.5.1 Characteristics of cluster sampling

79. Cluster sampling differs importantly from stratified sampling in two ways.[10] 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 the 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.

80. 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.

3.5.2 Cluster design effect

81. The clustering effect of a sample is partially measured by the design effect, *deff*, previously mentioned. However, *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.

[10] 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.

$$deff = f \approx 1 + \delta(\tilde{n} - 1), \text{ where} \tag{3.4}$$

f is the shortened symbol for *deff*,

δ is the intraclass (or intra-cluster) correlation, that is, the degree to which two units in a cluster are likely to have the same value compared to two units selected at random in the population,

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

82. Formula (3.4) 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.

83. In the expression above it can be seen that *deff* is a multiplicative function of two variables, the intraclass correlation, δ, and the size of the cluster, $\tilde{n}$. Thus, *deff*, increases as both δ and $\tilde{n}$ increase. While the sampler can exercise no control over the intraclass 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 intraclass 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 5,000 households. Suppose further that all households are the same size, 5 persons, just to simplify the illustration. The value of $\tilde{n}$ is then 50 for 10 households and 100 for 20 households. Simple substitution in expression (3.4) 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 between the two options. .First, is it 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, second, is it better 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 also be considered.

84. 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.

3.5.3 Cluster size

85. 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 intraclass 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, for keeping *deff* to a minimum is to urge that cluster sizes be as small as the budget can allow. Table3.2 displays *deffs* for varying values of the intraclass correlation and a constant cluster size.

Table 3.2. Comparison of Clustering Component of Design Effect for Varying Intraclass correlations, δ, and cluster sizes, $\tilde{n}$

	δ						
$\tilde{n}$	**.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

86. From Table 3.2 it is clearly seen that cluster sizes above 20 will give unacceptable *deff*s (greater than 3.0) unless the intraclass correlation is quite small. In evaluating the numbers in the table it is important to remember that $\tilde{n}$ refers to the number of units in the target population, not the number of households. In that respect the value of $\tilde{n}$ 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 $\tilde{n}$ and *b* are roughly equal for that target group and Table 3.2 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 $\tilde{n}$ in that case is 10 times the average household size; if the latter is 5.0 then $\tilde{n}$ is 50. Thus, 50 is the value of $\tilde{n}$ that must be viewed in Table 3.2 to assess its potential *deff*. Table 3.2 reveals that *deff* is very large except when δ 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 δ.

87. 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).

3.5.4 Calculating *deff*

88. 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 7) and then computing, for each variable, the ratio of its variance to that of a 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.

89. The square root of the ratio of variances gives the ratio of standard errors, or *deft* 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).

3.5.5 Number of clusters

90. 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.

3.6 Sampling in stages

91. On a theoretical level the perfect household survey sample plan is to select the sample, *n*, of households randomly from among appropriately identified strata comprising 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,[11] as we have previously noted in the discussion of the cost benefits that cluster sampling affords.

3.6.1 Benefits of sampling in stages

92. Selecting the sample in *stages* has practical benefits in the selection process itself. It permits the sampler to isolate, in successive steps, the geographic locations where the survey operations - notably, listing households and administering interviews - will take place. When

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

listing must be carried out because of an obsolete sampling frame, a stage of selection can be introduced to limit the size of the area to be listed.

93. With cluster sampling, in general, 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 size of 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. The *EA frame* (see more about frames in the next chapter) is complete because the entire country is divided into *EA*s. It is accurate because every household lives, by definition, in one and only one *EA*. And it is reasonably current in the sense that it is based on the most recent census, provided there have been no changes after the census in the definitions of the *EA*s. Suppose, further, the census is two years old. It is determined, therefore, that it will be necessary to compile a more 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 the listing workload 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.

3.6.2 Use of dummy stages

94. 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.[12] Dummy stages allow one to narrow the sub-universes to more manageable numbers by taking advantage of the hierarchical nature of administrative sub-divisions of a country.

95. 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

[12] One method of making a very large computer file more manageable for sampling, however, is to decompose it into separate sub-files for each stratum or administrative area (region or province, for example).

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.

96. 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. In order to illustrate the dummy character of the first two stages, it is necessary in this instance to show it mathematically. This is done by examining the probabilities at each selection stage and the overall probability.

First stage of selection: thanas

97. 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:

$$P_1 = \left(am_t\right) \Big/ \sum m_t \text{, where} \qquad (3.5)$$

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[13] in the tth thana according to the sampling frame used (for example, the most recent population census).

98. The factor, $\sum m_t$, 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 $\sum m_t \div a$. Thus, if the sampling interval is, say, 12,500 and the thana contains 13,800 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

99. 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

[13] 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.

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:

$$P_2 = (1)\left(m_u\right)/m_t \text{, where} \tag{3.6}$$

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 uth union according to the frame.

Third stage of selection: villages

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

$$P_3 = (1)\left(m_v\right)/m_u \text{, where} \tag{3.7}$$

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

(*1*) signifies that only one village is selected,

m_v is the number of households in the vth village according to the frame.

Fourth stage of selection: households

101. 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:

$$P_4 = (b)/m_v \text{, where} \tag{3. 8}$$

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

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

$$P = P_1\, P_2\, P_3\, P_4. \tag{3.9}$$

Substituting, we have:

$$P = \left[(am_t)/\sum m_t\right]\left[(1)(m_u)/m_t\right]\left[(1)(m_v)/m_u\right]\left[b/m_v\right]$$

$$= \left[(a)(b)\right]/\sum m_t \tag{3.10}$$

102. 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.

3.6.3 The two-stage design

103. 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).

104. Typically the two-stage design consists, simply, of a *pps* sample of several hundred geographical units, suitably stratified, 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.

105. 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.

3.7 Sampling with probability proportional to size (PPS)

106. In section 3.5 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.

3.7.1 *PPS* sampling

107. 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 be no advantage to using *pps* sampling. Suppose for example, every block in a particular city contained exactly 100 households and one wanted a sample of 1,000 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 1,000 altogether. The selection equation in this case is expressed as

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

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.

108. 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.

109. 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. In that case the result would be an overall sample size drastically different from the desired 1,000 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. That method is not generally recommended because it may reduce or complicate the use of other stratification factors in the sample design. *PPS* sampling is the preferred solution because it permits greater control over the ultimate sample size without the need for stratification by size.

110. To illustrate *pps* sampling we start with the selection equation, mentioned above, but expressed more formally for a two-stage design[14] as follows:

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

$$P(\alpha\beta) = P(\alpha)P(\beta|\alpha), \text{ where} \qquad (3.11)$$

$P(\alpha\beta)$ is the probability of selecting household β in cluster α,

$P(\alpha)$ is the probability of selecting cluster α,

$P(\beta|\alpha)$ is the conditional probability of selecting household β in the second stage given that cluster α was selected at the first stage.

111. 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 $P(\alpha\beta)$. 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 ith cluster, then, we need $P(\beta|\alpha)$ to be equal to b/m_i. Then,

$$P(\alpha\beta) = [P(\alpha)][b/m_i].$$

Since $n = ab$, we have

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

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

$$P(\alpha) = (a)(m_i)/N. \qquad (3.12)$$

112. Note that $N = \sum m_i$, 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

$$P(\alpha\beta) = [(a)(m_i)/\sum m_i][b/m_i] \qquad (3.13)$$
$$= [(ab)/\sum m_i]. \qquad (3.14)$$

113. 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, $\sum m_i$, is a constant equal to *N*. Illustration 3.2 below provides an example of how to select a sample of clusters using *pps*.

114. To physically select the sample, note that in illustration 3.2 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.

115. 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.

Illustration 3.2. 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 *PSU*s (clusters) from 32 in the universe using *pps*; Selection Interval (*I*) therefore equals 4541/8, OR 567.6, where 4,541 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.

3.7.2 *PPES* sampling (probability proportional to estimated size)

116. 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.

117. In household surveys the measure of size generally adopted for the first stage selection of primary sampling units (*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 each 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.

118. 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.

119. 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. It will 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 *ppes*, its probability, from the selection equation, is

$$P(\alpha\beta) = \left[(a)(m_i)/\sum m_i\right]\left[b/m_i^{'}\right], \text{ where} \qquad (3.15)$$

$m_i^{'}$ is the count of households according to the listing operation and the other terms are defined as previously.

120. 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.

121. 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,[15] be prudent to generate the survey estimates without weighting since 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 assess empirically if the differences are minor.

122. 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.

3.8 Options in sampling

123. 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 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.

3.8.1 *Epsem*, *PPS*, fixed-size, fixed-rate sampling

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

[15] 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.

Chart 3.1 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

125. 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 in the frame.

126. Selection of a fixed number of households in each sample cluster has two very important advantages. First, the sample size is controlled precisely. Second, the method provides the means for the survey manager to assign exact workloads to interviewers and to equalize those workloads if he/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. There is, however, a built-in quality control check since the number of households to be selected is known in advance. Still, the complications can create inefficiency from time lost through errors in selection and then have to correct them.

127. 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. It is preferable if the selection is done by someone other than the lister himself, in order to minimize the possibility of bias in the selection procedure.

128. 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 limitations.

129. One limitation 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 limitation 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.

130. 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 3.1 are self-weighting.

Plan 1 – *pps*, fixed cluster size

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

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

Limitations
- *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

131. 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

Limitations

- *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

132. 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*

Limitations
- 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
-

Limitations
- 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

3.8.2 Demographic and Health Survey (DHS)

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

134. The DHS, which has been widely applied in scores of developing countries since 1984, promotes the use of the *standard segment design*,[16] for its convenience and practicality, in its sampling manual. A standard segment is defined in terms of its size, usually 500 persons. Each geographical area unit of the country that comprises the sampling frame is assigned a measure

[16] The standard segment design was also used in the PAPCHILD survey programme of the 1980s-90s – see League of Arab States (1990).

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 area unit.

135. A *pps* sample of area units is selected using the number of standard segments as the measure of size, *MOS*. Since the area 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 area 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 area 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).

136. Each area 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.

137. 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 limitations noted above for plan 6 through its standard segmentation procedure; the overall sample size is controlled almost precisely as well as interviewer workloads.

138. An important advantage of the standard segment design is that the listing workload at the penultimate stage of selection is reduced substantially. For every area unit consisting of s segments, the listing workload is reduced to $1/s$ (when there is only one segment there is no reduction). For example, if a given area unit contains 4 segments the listing workload is only one-fourth what it would be if listing had to be carried out in the entire area. Sample preparation costs are thus reduced because of this feature.

139. While listing costs are lower, the reduction comes at an additional price. A limitation, therefore, of the standard segment design is that mapping operations must be conducted for segments with a *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 area 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.

3.8.3 Modified Cluster Design - Multiple Indicator Cluster Surveys (MICS)

140. 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, as mentioned, 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.

141. 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 (see the section on probability sampling versus other sampling methods for discussion of why probability sampling is the recommended approach in household surveys).

142. 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).

143. 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 area 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 area 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.

144. 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 limitation of having less control over the ultimate sample size. That is because a segment selected could be much bigger than the frame indicated due to growth.

145. Mapping, however, is required for MCS just as it is for the standard segment design of DHS, with all of the limitations 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 nonexistent for small areas. A final limitation is that the segment interviewed is a compact cluster – all the households are geographically contiguous – and this has a bigger design effect, because of the relatively high intra-class correlation, than the non-compact clusters of the standard segment design.

3.9 Special topics – two-phase samples and sampling for trends

146. 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.

3.9.1 Two-phase sampling

147. 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 present in only a small percentage of 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, area units where an ethnic 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.

148. 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.

149. The objective of the two-phase approach is to save costs by having a short, screening interview in the initial, large sample. It is followed by the more extensive interview at a later date, but only 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 1,000 orphans. From that, the orphan survey manager would then plan to sub-sample 80-per cent of those households for the detailed interview.

150. 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 (3.1)) the sampling technician and the survey manager would likely conclude that two-phase sampling is the most practical and cost-efficient design to use.

151. 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.

3.9.2 Sampling to estimate change or trend

152. 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.

153. To examine the reliability issue, again it is necessary to demonstrate it mathematically. We start by looking at the variance of the estimated change, $d = p_1 - p_2$. It is expressed by:

$$\begin{aligned}\sigma_d^2 &= \sigma_{p1}^2 + \sigma_{p2}^2 - 2\sigma_{p1,p2} \\ &= \sigma_{p1}^2 + \sigma_{p2}^2 - 2\rho\sigma_{p1}\sigma_{p2}, \text{ where}\end{aligned} \tag{3.16}$$

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

σ_d^2 is the variance of the difference,

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

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

ρ 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:

$$\sigma_{p1}^2 \approx \sigma_{p2}^2$$

Then, $\sigma_d^2 = 2\sigma_p^2 - 2\rho\sigma_p^2$ (We can drop the subscripts, 1 or 2).

$$\text{So, } \sigma_d^2 = 2\sigma_p^2 (1 - \rho). \tag{3.17}$$

154. To evaluate [16] we note that an estimate of σ_p^2 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, ρ, 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_d^2, of σ_d^2 is given as:

$$s_d^2 = 2[(pq)f/n](0.2), \quad or \quad 0.4\,(pq)f/n \tag{3. 18}$$

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

$$s_d^2 = 2[(pq)f/n](0.7), \quad or \quad 1.4\,(pq)f/n \tag{3.19}$$

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

$$s_d^2 = 2[(pq)f/n] \tag{3.20}$$

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

$$s_d^2 = 4[(pq)\ /n]\ . \tag{3.21}$$

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

$$\sigma_d^2 = 2\sigma_p^2\,(1 - F\rho). \tag{3.22}$$

158. 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.

159. 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. Both strategies produce estimates with smaller variance compared to the least attractive option of completely independent samples.

160. 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.

161. 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. This pattern has persisted over a number of years in the U.S. Various reasons have been postulated by experts to account for this bias and they include the following (Kasprzyk, 1989):

- 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.

162. 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.

163. From the above it can be seen that 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.

164. 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.

165. 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 examples of rotation sampling in master samples in chapter 4).

3.10 When implementation goes wrong

166. This section provides a summary of actions to take when implementation of the sample plan encounters obstacles. Most of the examples have already been discussed or alluded to in the preceding text. Many of the implementation obstacles can be forestalled, however, by very careful planning when the sample design is conceived. That is one of the important principles which this chapter and the next attempt to achieve. Still, despite the best planning there are unforeseen problems that can arise.

3.10.1 Target population definition and coverage

167. Problems often occur when the intended target population is not the actual population covered by the survey. This can happen in a variety of ways, as illustrated by the example.

- ***Example***

Consider a survey intended to cover the typical target population of a national, household survey – all people in the country. The actual population covered (that is, from which the sample is selected) is often less than the total for any of the following reasons:

- Persons living in institutional quarters such as hospitals, prisons and military barracks are not sampled.
- Persons residing in certain geographic areas may be purposely excluded from coverage. Those areas might include inaccessible terrain, ones affected by natural disasters, those declared off limits due to civil disorder or war, compounds or camps where refugees and other foreign workers reside, and so forth.
- Persons who do not have permanent living arrangements are ruled "out of scope" for the survey. These may include nomadic populations, boat people, gypsies, transient workers, etc.

168. The problem regarding such sub-populations with respect to the sample plan is that they are not usually identified in advance of the survey as groups that ought to be excluded. Implementation thus suffers when sample selection, by chance, chooses, say, (a) a cluster that turns out to be a work camp, prison or dormitory instead of a "traditional" residential area, or (b) a *PSU* that is in mountainous terrain and thought to be inaccessible. The "solution" often taken in such situations is to substitute another *PSU*. This solution is a biased procedure, however.

169. The acceptable solution is to avoid the problem at the design phase of the sample. It is done by, first, carefully defining the target population and specifying not only which sub-populations it includes but which ones are to be excluded from coverage. Second, the sample frame should then be modified to delete any geographic areas that are not to be covered by the survey. This applies as well to any special-purpose *EA*, for example, such as a work camp, that should be excluded. Third, the sample should be selected from the modified frame. See more about sampling frames in the next chapter.

170. It should also be born in mind that the solution suggested above also serves to define the target population more precisely. It is important that the exact target population be described in the survey reports so that the user is properly informed.

3.10.2 Sample size too large for survey budget

171. Another problem occurs when the calculated sample size is 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.

172. One way of reducing the precision (increasing the sampling error) that lowers the cost substantially is to select fewer *PSU*s yet retain the overall sample size. For example, instead of 600 *PSU*s of 15 households each (n = 9,000), the sample plan could be modified to select 400 *PSU*s of 22 or 23 households each ($n \approx$ 9,000). As for domains, a solution might be to settle upon 4 major regions of the country instead of, say, 10 provinces.

3.10.3 Cluster size larger or smaller than expected

173. A problem that frequently occurs is that a sample cluster may be much larger than its measure of size. This may happen from new housing construction, especially if the sample frame is old. For example, the survey team may expect 125 households in a given cluster but find 400 instead at the listing stage. A plausible solution in this case is to sub-divide the cluster into geographic sub-segments of approximately equal size in terms of population. The number of segments should be equal to the current count of households divided by the original measure of size, rounded to the nearest integer. In our example, this would be 400/125 or 3.2, rounded to 3 segments. The segments would be created through mapping and quick-counting of dwellings (as opposed to households). Then, one segment would be selected at random for listing.

174. The opposite problem may also occur. A cluster may be much smaller than expected, due to demolition, natural disaster or other reasons. There is often the temptation to substitute another cluster in such cases, but to do so is biased. Instead, the smaller cluster should be taken as it stands. While this may reduce the ultimate sample size from its target, the increase in sampling error would be minor unless a large number of such clusters are involved. By taking the smaller cluster without modification (or substitution) an unbiased estimate will nevertheless

be attained, because the cluster "represents" the current population change that has occurred since the frame was established.

3.10.4 Handling non-response cases

175. Though it pertains more to survey rather than sample implementation, non-response is a serious issue that can ruin household survey estimates (see chapters 6 and 8 for detailed discussion on non-response). If non-response is allowed to occur in more than 10-15 percent of the sample cases, the resulting bias in the estimates may make them highly questionable. Again, a tendency in many countries is to "solve" non-response by substituting households than do respond. The technique itself is biased because the substituted households still only represent responding households, not non-responding ones. The characteristics of the latter two groups are known to be different with respect to important survey variables, especially those related to socio-economic status. The preferred solution, which, unfortunately is never 100 percent successful, is to obtain responses from initially non-responding households. This must be done by planning, at the outset, to return to households that are non-response in a series of successive call-backs in the effort to gain their cooperation (for refusals) or to find them at home (for absentees or otherwise unavailable). As many as 5 call-backs may be necessary, but the minimum should be 3.

3.11 Summary guidelines

176. This section summarizes the main guidelines to be inferred from this chapter. While some of the guidelines would prevail under almost any circumstances (for example, "use probability sampling"), there are others where exceptions would be appropriate, depending upon a country's special circumstances, resources and requirements. For that reason the guidelines are presented more in the spirit of "rules of thumb" rather than fixed and unwavering recommendations. They are, in checklist format, as follows:

- Use probability sampling techniques at every stage of selection.
- Strive for as much simplicity in sample design as possible, as opposed to complexity.
- Seek selection techniques that yield self-weighting, or approximately self-weighting, samples within domains or overall if design does not include domains.
- Use two-stage sample design if possible.
- Calculate sample size using formula (1), adjusting the value of fixed parameters (such as expected non-response rate and average household size) as necessary to reflect country situation.
- Use design effect value of 2.0 as default in sample size formula unless better information is available for your country.
- Base sample size on the key estimate thought to comprise smallest percentage of population from among all the key estimates the survey will cover.
- Budget permitting, choose margin of error, or precision level, for key estimate (above) that is 10 percent of the estimate, that is, 10 percent relative error, *at 95 percent level of confidence*; otherwise, settle for 12-15 percent relative error.
- Define first-stage selection units, *PSU*s, as census enumeration areas, *EA*s, if convenient and appropriate.

- Utilize implicit stratification coupled with systematic *pps* sampling where possible, especially for multi-purpose designs.
- Limit number of estimation domains to as few as absolutely necessary (to control the sample size to a manageable level).
- Strive for large number – several hundred – of clusters (or *PSU*s if two stages); the more the better.
- Use small cluster sizes – 10-15 households; the smaller the better.
- Use a constant cluster size rather than a variable one, that is, fixed number of households instead of fixed rate.
- For domains aim toward a minimum of 50 *PSU*s each.
- Plan on a minimum of 3, but preferably 5, call-backs to convert non-response households.
- For rare populations consider the two-phase sampling approach of attaching a "rider" question onto an existing, large survey already planned to locate the target persons; and follow up with an intensive interview on a subsample.
- For surveys to measure change, interview the same households on both occasions if only two interviews are to be taken; if three or more interviews use a scheme of partial overlap by rotating new households into the sample on each occasion.

References and further reading

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

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

Hansen, M., Hurwitz, W. and Madow, W. (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.

Le, T. and Verma, V. (1997), *An Analysis of Sample Designs and Sampling Errors of the Demographic and Health Surveys*, DHS Analytical Reports No. 3, Macro International Inc., Calverton, Maryland.

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. 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, International Statistical Institute Session, Florence.

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

Turner, A., Magnani, R., and Shuaib, M. (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" and Appendix 7: "*Sampling Details*," UNICEF, New York.

United Nations Statistics Division (1984), *Handbook of Household Surveys*, revised edition ST/ESA/SER.F/31, United Nations, New York.

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

_____ (2005 forthcoming), *Household Sample Surveys in Developing and Transition Countries*, ST/ESA/STAT/SER.F/96, 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," *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.

Chapter 4

Sampling Frames and Master Samples

4.1 Sampling frames in household surveys

1. With the exception of sample frames, the preceding chapter covered the multi-faceted features of sample design and some of its options in household surveys. One of the most crucial aspects of sample design in household surveys, however, is its frame. For this reason a separate chapter is devoted to it.

2. The sampling frame has significant implications on the cost and the quality of any survey, household or otherwise. In household surveys faulty sampling frames are a common source of *nonsampling error*, particularly under-coverage of important population sub-groups. This chapter attempts to elaborate best practices in frame construction and usage taking into account various stages of sampling. It is divided into two sections. The first section covers general issues on frames and their development, with emphasis on multi-stage sample design in household surveys. The second section discusses the special issues that pertain when a *master* sample frame is to be used.

4.1.1 Definition[17] of sample frame

3. A simple, operational definition of a sampling frame is the *set of source materials from which the sample is selected.* The definition also encompasses the purpose of sampling frames. That is to provide a means for choosing the particular members of the target population that are to be interviewed in the survey. More than one set of materials may be necessary. That is generally the case in a household survey because of its multi-stage nature. The early stages of selection in household surveys are typically drawn from geographical *area* frames. The last stage may be selected either from an area or *list* frame (see subsections below on area and list frames).

4.1.1.1 Sample frame and target population

4. An important consideration in deciding upon the appropriate frame(s) to use for household surveys is the relationship between the survey target population and the unit of selection. The unit of selection determines the frame. It also determines the probability of selection at the last stage.

- ***Example***

To illustrate, in a survey whose target population is infant children the survey team might consider two potential frames. One might be medical facilities recording births within the past 12 months. The second would be households whose occupants include infants under 12 months

[17] The reader is referred to Chapter 3, Table 1.1 for the glossary of sampling-related terms used in both Chapters 3 and 4.

old. In the first instance the frame comprises two parts, one for each stage of selection. The first would be the list of hospitals and clinics where infants are born. The second is the list of all infants born in those facilities in the past 12 months. The units of selection are the medical facilities at the first stage and infants at the second stage. Thus, the unit of selection and the target population are synonymous terms at the *final* selection stage. In the second instance, however, the frame would likely be defined (in a latter stage of selection) as a list of households in small area units such as villages or city blocks. In applying the sample plan households would be selected and screened to ascertain the presence of children 0-12 months old. In this case, the household is the unit of selection upon which the probability of selection is based. Note however that members of the target population are not actually identified and surveyed until the households are screened for their presence. Thus, the unit of selection and the target population are different in the case of the household frame.

5. In household surveys – the subject of this handbook – the unit of selection and around which the sample design is based is the household. Yet the target population, even in a general-purpose survey, will differ depending upon the measurement objectives. Except for household income and expenditure surveys, the target population will usually be a population other than the household itself. Examples are employment surveys where the target population is generally persons 10 (or 14) years old or older, thus excluding young children altogether; surveys on reproductive health of women where the target population is women 14-49 years old (and often only ever-married women in that age group); etc.

4.1.2 Properties of sampling frames

6. As we discussed above the sampling frame must of course capture, in a statistical sense, the target population. Beyond that, a perfect sample frame is one that is *complete*, *accurate* and *up-to-date*. These are ideal properties that are unattainable in household surveys. Nevertheless it is essential to strive for them either in constructing a frame from scratch or using one that already exists. The quality of a frame may be assessed in terms of how well its idealized properties relate to the target population. Recall from chapter 3 that our definition of a probability sample – one in which every member of the target population has a known, non-zero chance of being selected – is a useful barometer for judging a frame's quality.

7. The degree to which there is failure to achieve each of the ideal properties produces survey results that are biased in various ways, but often in the direction of *under-estimating* the target population.

4.1.2.1 Completeness

8. The ideal frame would be deemed complete with respect to the target population if all its members (the *universe*) are covered by the frame. Coverage of the target population(s) is therefore an essential feature in judging whether the frame is suitable for a survey. If not suitable, then whether it can be repaired or further developed to make it more complete must be assessed by the survey team. In the previous example, infants born at home or other places outside medical facilities would not be covered in the survey if medical facilities were used as the sole frame for sampling. Hence, in this example there are significant numbers of the target

population that have a zero chance of inclusion in the sample, and the condition for a probability sample is violated. As a result, an estimate of the number of infants would be under-stated by the facility frame. Moreover, the *characteristics* of infants would likely be quite different from those born at home. The facility frame would therefore yield biased distributions for important indicators about the infants or their care.

9. Inadequate coverage is also a potential problem in household surveys. For example, a national survey plan may be intended to cover the entire population through a household survey. There are, however, various sub-groups such as persons living in institutions, nomadic households and boat people that do not reside in households. In such a case coverage of the total population is obviously not attainable through the household survey. Additional frames would have to be developed to cover non-household groups in order to give their members a non-zero probability of being included. Failing that, the actual target population would have to be modified to more carefully define what it includes. In that way, users are clearly informed of which segments of the population are excluded from coverage.

4.1.2.2 Accuracy

10. Accuracy is an important feature as well in sampling frames, although inaccuracies are more likely to occur in frames other than those used for household surveys. A frame can be said to be accurate if each member of the target population is included once and only once. A simple illustration is the case of a list of some kind that contains errors. An example would be a list of business establishments defined as those employing more than 50 workers. Errors could occur if (1) any establishment on the list had 49 or fewer workers, (2) any establishment with 50+ workers was missing from the list, or (3) an establishment was listed more than once (perhaps under different names).

11. In household surveys it is less likely such inaccuracies in the frame would be encountered. Some examples however would be the following: (1) a frame consisting of a computer file of enumeration areas, *EA*s, that is missing some of its elements, (2) a list frame of households in a village that is missing some of those living on the perimeter of the village, (3) a list frame of households in an area unit where some of the households are listed in more than one unit and (4) an old list frame of households that does not include newly constructed dwellings. The latter is also an example of a frame that is not current, discussed further below.

12. Missing *EA*s or listed households within an area unit mean of course that the affected households have no chance of being selected for the sample. Again, this would violate one of the conditions for a true probability sample. Duplicate listings also violate the probability criterion unless they are known about so that the true probabilities of selection can be adjusted. Unfortunately, omissions and erroneous duplicates of the types mentioned are often not known. The sampling technician may therefore be unaware of the need to correct the frame before sampling from it. On the other hand, a small percentage of cases that are omissions or duplicates in a frame will not usually cause any appreciable, or even noticeable, nonsampling bias in the survey estimates.

4.1.2.3. Current frame

13. Ideally of course a frame should be current in order for it to fulfill the other two properties of completeness and accuracy. An obsolete frame obviously contains inaccuracies and is likely to be incomplete, especially in household surveys. The quintessential example of a frame that is out of date is a population census that is several years old. The old census will not accurately reflect new construction or demolition of dwellings, in- or out-migrants in dwelling units, births or deaths. These deficiencies violate the criterion of a probability sample that each member of the target population must have a *known* chance of selection.

- ***Example***

Suppose the frame consists of *EA*s defined according to the most recent census, which is 4 years old and no up-dating of the frame has been done. Suppose further that numerous squatter areas have built up on the outskirts of the capital city in *EA*s that were, at the time of the census, either empty or virtually empty of population. The sample design would give no chance of inclusion to households living in the formerly empty *EA*s, thus violating the probability sample conditions. In *EA*s that were virtually empty another serious problem arises even though those *EA*s do not technically violate probability sampling requirements. The sample would undoubtedly be selected using *pps* with the *MOS* being the census population or household count. By virtue of its having very little population at the time of the census, any high-growth *EA* would have only a slight chance of being selected when *pps* sampling is used. As a result the sample could have an unacceptably high sampling variance.

4.1.3 Area frames

14. In this subsection and the next we discuss the two categories of frames that are used in sampling, whether for household surveys or other applications. It is important to note that in a multi-stage design the frame for each stage must be considered as a separate component. The specific frame is different at each stage. The sample design for a household survey will likely use both an area frame, discussed in this subsection, for the early stages and a list frame, next section, for the last stage.

15. In household surveys an area sampling frame comprises the geographical units of a country in a hierarchical arrangement. The units are variously labeled, administratively, from one country to another but typically include such terms, in descending order, as province or county; district; tract; ward and village (rural areas) or block (urban areas). For census purposes administrative sub-divisions are further classified into such entities as crew leader areas and enumeration areas or *EA*s. Often the census *EA* is the smallest geographical unit that is defined and delineated in a country.

16. For survey purposes there are four distinct characteristics of geographical units that are important for sample design, as follows:

(1) The geographical units cover, usually, the entire land area of a nation.
(2) Their boundaries are well-delineated.
(3) There are population figures available for them.

(4) They are mapped.

17. Coverage of the totality of the nation's geographic area is important, as we have noted, because it is one of the criteria for achieving a bona fide probability sample. Well-delineated boundaries that are mapped are invaluable in sample implementation because they pin down the locations where field work is conducted. Good boundary information also helps the interviewer locate the sample households that are ultimately selected for interview. Population figures are needed in sample design to assign measures of size and to calculate the probabilities of selection.

18. The usual starting point in development of an area frame for household surveys is a country's population census – for the four reasons cited above. In addition the *EA* is a conveniently-sized geographical unit to select in the latter stages of sampling (the penultimate stage in a two-stage design). In most countries *EA*s are purposely constructed to contain roughly equal numbers of households – often about 100 – in order to provide comparable workloads for census-takers.

19. An area frame is, paradoxically, also a *list* – because one must begin with a list of the administrative, geographic units of a population to select the early stages of a household survey sample. This leads to the discussion of list frames.

4.1.4 List frames

20. A list sampling frame is quite simply a frame made up of a list of the target population units. Theoretically, a list frame for household surveys exists for every country just after its census is taken. The fresh census provides, in principle, a geographically arranged listing of every household – or dwelling unit - in the country.

21. A newly completed census list is ideal as a household sampling frame because it is as current, complete and accurate as any household list could ever be. Because of its geographical arrangement it is fairly simple to stratify it for proper geographical distribution of the sample. When there is a need to conduct a census follow-up sample survey to obtain more detailed or supplemental information than the census can efficiently provide, the fresh census list is thus ideally suited to use as the list frame. It is important to recognize, however, that the new census list is only briefly available as a *current* frame. Obviously, the longer the interval between the census and a follow-on survey the less useful the census listings would be as the frame source.

22. There are other examples of lists that might be considered as appropriate sampling frames for household surveys depending upon their quality. One is a civil registry. Another is a register of utility connections. Civil registries would be candidates for frame use in countries where careful records are kept of its citizens and their addresses. In some instances they may be more useful than an area-based census frame, because the registry may likely be continuously up-dated. Utility – usually electricity – connections may be useful as a sampling frame whenever a country's census is seriously obsolete. It would of course have to be evaluated to assess potential problems and their impact. An obvious problem leading to under-coverage

would be households not having access to power. Another that would have to sorted out would be electrical hook-ups servicing multiple households.

23. Another list frame that is widely used in developed countries is a register of telephone subscribers. Sampling is done through *random digit dialing* (*RDD*) techniques to ensure that subscribers with unpublished telephone numbers have their proper chance of being selected. *RDD* sampling is not recommended, however, in countries that have low penetration rates for telephone ownership.

24. In a conventional household survey the last stage of selection is, invariably, based on a list frame concept. We have discussed previously how the penultimate design stage may yield a sample of clusters in which a current listing of households is compiled. From that list the sample households are selected. Thus we have an area frame defining the sample clusters but a list frame defining the sample households within the clusters.

4.1.5 Multiple frames

25. In the preceding chapter we discussed two-phase sampling in household surveys. It involves the use of screening techniques to identify a particular target group in the first phase, followed by a second-phase interview of a sub-sample of those identified. Another sampling technique that may accomplish much the same end result is to use more than one sampling frame. Usually this involves only two frames, in which we have a *dual-frame* design, but occasionally three of more frames may be used (*multi-frame* design)

4.1.5.1. Typical dual frame in household surveys

26. For simplicity of presentation we will discuss dual-frame designs though the principles are analogous for multi-frame designs. In general the methodology entails combining a general-population area frame with a list frame of persons known to be members of the particular target population under study. For example, consider a survey intended to study the characteristics of unemployed persons. The survey can be based on an area frame of households but supplemented with a list-frame sample of currently unemployed persons that are registered with the social services ministry. The objective of a dual-frame sample of this type is to build up the sample size with persons that have a very high probability of being in the target population. The approach can be a cheaper and more efficient alternative to two-phase sampling. It is necessary to use the general-purpose household frame to account for target population members who are not on the list. In the example mentioned they would comprise unemployed persons not registered with social services.

27. There are several limitations, however, with dual-frame designs. One is that the list-frame must be virtually current. If a large percentage of persons selected from the list have had a change in status that removes them from the target population, then use of the list frame is inefficient. In our example any unemployed person that has become employed by the time the survey is conducted would be ineligible, which illustrates why the list frame must be up to date.

28. Another limitation is that persons on the list frame will likely reside in dispersed locations throughout the community and it is costly to interview them due to travel. That of course is in stark contrast to the area-based household frame where the sample can be selected in clusters to reduce interviewing costs.

29. A serious issue with dual-frame designs is that of duplication. Generally, persons included on the list frame will also be included on the area frame. Again in our example, unemployed persons selected from a registry are members of households of course. They would have a duplicate chance of selection, therefore, when both frames are used. The duplication issue can be addressed in order to adjust for it properly. To do so, however, has implications for the content of the survey questionnaire. In our example each unemployed person interviewed in the household sample would have to be queried as to whether he/she is registered with the ministry on its list of unemployed. For those that respond in the affirmative, further work is necessary to match their names with the list-frame, a process that is error-prone and fraught with complications. When a successful match is found the survey weight of the person affected must be changed to $(1/ P_h + 1/ P_l)$ to reflect the fact that she has a probability, P_h, of being selected from the household frame and P_l of being selected from the list frame. It is important to note that matching must be done against the entire list frame and not just those on the frame that happened to have been selected in the sample. This is because the probability (and weight) is a function of the chance of selection irrespective of whether actual selection occurs.

4.1.5.2 Multiple frames for different types of living quarters

30. Another type of dual-frame sampling occurs when the target population resides in different kinds of living quarters that are non-overlapping. For example, a survey of orphans would most likely be designed to include orphans living in two types of housing arrangements. First, institutions such as orphanages would comprise one frame. Second, households would have to be sampled to cover orphans living with a surviving parent, other relatives or non-relatives. The dual-frame design would thus consist of a household frame plus an institutional frame and they are of course non-overlapping.

31. The objective of a design of this type is to cover the target population adequately (as close to 100-percent as possible). When significant numbers of the population live in each of the two types of living quarters, significant biases would occur if the sample were restricted to only one of the frames. A sample of orphans, for example, based only on those living in households would not only yield an underestimate of the population of orphans but a biased estimate in terms of their characteristics. Similar biases would occur for a survey exclusively based on orphans living in institutions.

32. The limitation discussed above regarding duplication does not pertain to designs from dual, non-overlapping frames. For that reason they are significantly less difficult to administer.

4.1.6 Typical frame(s) in two-stage designs

33. In the previous chapter we emphasized the practical value of two-stage sample designs. This subsection discusses the frame that is typically used in two-stage designs.

34. The geographical units – clusters – that comprise the first stage of selection are often defined as villages (or parts of villages) or census EAs in rural areas and city blocks in urban areas. The frame consists, then, of all the geographical units that make up the universe of study however defined – the nation as a whole, a province or set of provinces or the capital city. Sampling is done by compiling the list of units, checking it for completeness, stratifying the list in an appropriate fashion (often geographically) and then selecting a systematic sample of the units, the latter usually by pps.

35. If the file of clusters in the universe is very large there may have to be intermediate, dummy stages of selection, as discussed in the previous chapter. In that case the frame units are defined differently for each of the dummy stages. In the earlier example for Bangladesh the frame units for the two dummy stages were defined as thanas and unions.

36. The second-stage frame units in a two-stage design are simply the households in the first-stage sample clusters. When they are sampled from a list of households the frame is, by definition, a list frame. They may also be sampled as compact segments (see previous discussion in chapter 3, subsection 3.7.3) created by sub-dividing the clusters into geographical parts that are exhaustive and mutually exclusive. In that case the second-stage frame is an area frame.

4.1.7 Master sample frames

37. Here we just briefly mention the concept of a master sampling frame, which is discussed in considerable detail in section 4.2 below.

38. A master sample frame is one in which the frame is used to select samples either for multiple surveys, each with different content, or for use in different rounds of a continuing or periodic survey. With the exception of periodic up-dating as necessary, the sampling frame itself does not vary either from one survey to the other or from one round to another of the same survey. Instead – and this is its distinctive characteristic – the master sample frame is designed and constructed to be a stable, established framework for selecting the sub-samples that are needed for particular surveys or rounds of the same survey over an extended period of time.

4.1.8 Common problems of frames and suggested remedies

39. The problems that arise in household surveys from faulty frames are ones of both nonsampling bias and sampling variance. As implied in previous sub-sections common problems occur when the sampling frame is obsolete, inaccurate or incomplete. In the great majority of national, general-purpose surveys the basic frame is the most recent population census. That is the frame that is assumed in the remainder of this subsection. Problems of obsolescence, inaccuracy and incompleteness often occur together in census-based frames. They tend to increase in magnitude as the interval between the census and the survey increases.

40. We have mentioned that a frame must be current in order to reflect the current population. One based on, say, a five-year old census does not adequately account for population growth and migration. Even a current census frame can be incomplete and cause problems vis-à-vis household surveys if it does not cover military barracks, boat people, nomads and other important sub-populations that do not live in traditional household arrangements. Inaccuracies in both current and old census frames pose problems in a variety of ways. They include those arising from duplicate household listings, missing households or those enumerated or coded to the wrong EA.

41. Appropriate strategies for dealing with old, inaccurate or incomplete census frames depend in part on (1) the objectives of the survey and (2) the age of the frame. Regarding measurement objectives if a survey is purposely designed, for example, to cover only immobile households a census frame that excluded nomadic households would suffice. On the other hand, a procedure would have to be developed to create a frame of nomadic households if the survey is intended to cover them (in countries where they exist). In that respect, whether a census frame is complete or not depends on the definition of the target population, or sub-populations, to be covered by the survey.

42. Remedies to cope with problems of obsolescence and inaccuracies would differ depending on how old the census is. While it may not be prudent to offer a precise rule, owing to varying national conditions, a rule-of-thumb to guide an appropriate strategy for frame revamping or up-dating would be whether the census is more than two years old. As for inaccuracies of the type mentioned two paragraphs above, remedies given in subsection 3.8.2 are applicable.

4.1.8.1 Census frame more than two years old

43. The first situation applies to countries with old censuses - two years old or more. It is these old frames that present the biggest challenge in household survey sample design, especially in rapidly growing cities. Complete, countrywide up-dating of the old census frame is the ideal because, if successful, it assures that the resulting survey data are both as accurate, in terms of survey coverage, and as reliable as possible. Unfortunately, it is also the most expensive and time-consuming and, therefore, impractical. Still, there may be no alternative in those countries where the census is seriously obsolete.

44. Instead of complete up-dating a compromise would be to up-date the frame only in targeted areas, the latter identified by country experts familiar with growth patterns and demographic shifts. What is needed to up-date the census frame is fairly simple - a current MOS. For purposes of frame up-dating, the MOS would be defined as the number of dwelling units, as opposed to the number of households or persons.

45. It is important to recognize that the MOS does not need to be precise in order for the sample methodology to be valid. For example, if a given EA was thought to have 122 households on the basis of the last census, we would not be concerned if it currently has 115 or 132. For that reason it is not useful to attempt frame up-dating in old, established neighborhoods that are little changed over decades, even though individual inhabitants come

and go. Instead, what is of concern is when the current situation is drastically different from the last census – say, 250 households when 100 were expected. Such situations are likely in neighborhoods of heavy growth or demolition, such as squatter communities on the city fringes, high-rise development sites or demolition sites. Such areas would constitute the target areas for up-dating. We would rely upon country collaborators and experts to help identify the target areas and, of course, include only those where the changes are post-censal.

46. Up-dating generally entails several steps including (a) identification of the EAs that make up the targeted areas, (b) a quick-count canvass of the affected EAs to obtain a current MOS and (c) revision of the census file to show the up-dated MOS. As mentioned an approximate MOS is sufficient, which is why the quick-count canvass operation should be done to identify dwelling units rather than households. In quick-count canvassing, it is not necessary to knock on doors in order to count dwellings except, possibly, in multi-unit dwellings where the number of units is not apparent without entering the building.

47. Up-dating old census frames in the manner described above is necessary to stabilize the probabilities of selection for the penultimate stage units and, hence, the reliability of the survey estimates. Practically, the up-dating helps control not only the overall sample size but also the listing and interviewing workloads of the field staff. Moreover, it decreases the likelihood of encountering large clusters in the field that turn out to be much bigger than anticipated and having to sub-sample them or take some other appropriate action. Related to the last point is the fact that sub-sampling requires weighting adjustments – a complication in data processing. That potentiality would be diminished to the degree that no unexpectedly large clusters are encountered after sample selection at the penultimate stage.

48. The standard sample design for a household survey will likely entail compilation of a current listing of households in the sample clusters. In that case up-dating at the penultimate stage of selection would also occur (see reference to ppes sampling, which applies here as well, in following subsection). Hence the current listing for sample clusters that were not up-dated may resemble closely the census lists (though this is not guaranteed). It would be expected, however, that sample clusters from the up-dated portion of the census frame would yield current listings that would be significantly different from the census – both in the total number of households and in their specific identification.

49. One final point needs to be made about using an old census and this concerns sample validity rather than sample variance. As previously stated clusters are usually selected with probability proportionate to size. Failure to up-date the MOS for high-growth clusters in advance of sample selection would mean serious under-representation of areas that had small numbers of households in the census but have since grown significantly. Survey results would be biased and of course misleading because the characteristics of persons living in such high-growth areas are likely to be quite different from those in more stable neighborhoods.

4.1.8.2 Census frame two years old or less

50. This subsection applies to those countries that have comparatively recent censuses conducted within the past year or two and do not need general up-dating of the frame. In those

cases clusters would be selected using the original census *MOS*, since the latter would be expected to be quite accurate from the recent census. Updating per se would only take place at the penultimate stage of selection when field staff undertakes a current listing of households in the sample clusters. The sample households would be selected from the current listings and sampling weights would be adjusted, as necessary, in accordance with the procedures discussed in section 3.6.2 with respect to *ppes* sampling.

51. While a few clusters in the frame universe may have grown substantially since the census was completed, the number of such cases would not be expected to be so large as to significantly affect either field operations or survey precision. Any such clusters that happen to fall into the sample could be sub-segmented if necessary. Sub-segmentation, or "chunking", as it is known, is a field procedure intended to lessen the listing workload. The procedure involves (a) dividing the original cluster into sections, usually quadrants, (b) selecting one at random for listing and (c) selecting the households to be interviewed from that segment. Sub-segmentation does not improve sampling reliability because each sample chunk would carry an extra survey weighting factor equal to the number of chunks in the cluster – a factor of 4 if the cluster is divided into quadrants. Sub-segmentation does help, however, in containing field costs. The need for chunking can occur, even though the census is recent, again in high growth EAs that are drastically changed since the census. With a very recent census of course, it is expected there would be very few such areas.

52. It should be noted that the types of inaccuracies that were previously mentioned (duplicate or missing households, erroneous EA assignments) are partially corrected when up-dating that entails fresh listings of households in the penultimate stage is carried out. That of course is another strong reason to obtain current listings of households in surveys.

4.1.8.3 When a frame is used for another purpose

53. Survey managers sometimes question whether a household frame specifically constructed for one type of survey can be used for another. Can a sampling frame intended for a labour force survey, for example, be used in a sample design to measure health conditions, disability, poverty or agricultural holdings? Usually, however, it is not the frame itself that is problematic but the way it is stratified. Only when a frame has very unusual features – apart from its completeness, accuracy and currency, already discussed – might its use be constrained for household surveys other than the one originally planned. For example, if a survey focussing on cost-of-living is based only on urban communities (often the case in practice) the sampling frame would exclude rural areas. Clearly, such a frame would not be suitable to estimate poverty in countries where it is essentially a rural phenomenon.

54. Most household surveys are general-purpose, however, not only in terms of their content but in their sample designs. A labour force survey usually includes, for example, auxiliary information on demographics, educational attainment and other topics. In such cases an appropriate sample design is general-purpose as well, implying use of a customary sampling frame – one that covers all the nation's households. The frame may be stratified upon a variable specific to labour force measurement. For example, EAs might be classified according to the variable, percent unemployed in the latest census. Three strata of EAs might then be created –

low, medium and high unemployment. As mentioned above this is a stratification decision. The frame itself is unaffected. The solution is to "un-stratify" it if it is to be used for another survey such as, say, health.

55. A crucial task of the sampling statistician is to assess the sample frame in place when it is to be used for another type of survey. The assessment entails ensuring that the frame as constructed can meet the measurement objectives of the proposed survey. That would be done along the lines discussed throughout this chapter, notably, for completeness, accuracy and currency.

4.2 Master sampling frames

56. Master samples can be cost effective and efficient when a country has a sufficient number of independent surveys or periodic rounds of the same survey to sustain their use. It is perhaps self-evident that they be properly designed but it is also very important that they be properly maintained over time. The reader is urged to refer to *Sampling Frames and Sample Designs for Integrated Household Survey Programmes*, a document in the technical series published in 1986 by the UN Statistical Division under its National Household Survey Capability Programme (NHSCP). That report provides a much more comprehensive treatment of master sample frames and their uses.

4.2.1 Definition and use of a master sample

57. The sampling frame (or frames) for the first stage of selection in a household survey must cover the entire target population. When that frame is used for multiple surveys or multiple rounds of the same survey it is known as a master sample frame or, simply, a master sample.

58. Use of a master sample frame is the preferred design for any country that has a large-scale, continuing, intercensal household survey programme. Conversely, when there is not a continuing survey programme, master samples are not generally recommended. There are economies of scale in using the same frame units over time because much of the costs of sampling is absorbed in the developmental operations of the master frame rather than each time a survey is fielded. On the other hand countries that conduct only an occasional national survey between population censuses would not benefit appreciably from utilizing a master sample design.

59. The features of a master sample deal with the number, size and type of units at the first stage of selection. In general a master sample consists of an initial selection of primary sampling units (PSUs) that remain fixed for each subsample. Note that the latter stages are usually variable. For example, in the final stage of selection the particular households that are chosen for interview are usually different for independent surveys, while they may be the same or partially overlapping in repetitive surveys.

4.2.2 Ideal characteristics of PSUs for a master sample frame

60. The principles that govern the establishment of a master sample frame are little different from those for sampling frames in general. The master frame should be as complete, accurate and current as practicable. A master sample frame for household surveys is typically developed from the most recent census, just as a regular sample frame is. Because the master frame may be used during an entire intercensal period, however, it will usually require periodic and regular up-dating such as every 2-3 years. This is in contrast to a regular frame which is more likely to be up-dated on an ad hoc basis and only when a particular survey is being planned.

61. The features that are conducive to the development of a master frame are, likewise, similar to those for sampling frames generally. In defining units to use as the PSUs, for example, a constraining factor is that they should be area units that are already mapped. This is not a severe constraint, however, since the frame units will invariably be defined as administrative units already constructed for the census. An important feature that may differ from regular sampling frames, however, is that the size of the PSUs must be sufficiently large to accommodate multiple surveys without interviewing the same respondents repeatedly. Even this feature, however, can be relaxed in certain applications.

- ***Example***

A particular kind of master frame that has been used in some settings is based on a two-stage design. The first stage is a large sample of *EA*s (or similarly small and mapped area units). A sub-sample of the master *EA* sample is selected for each independent survey that utilizes the frame. Each sub-sample is listed or otherwise sub-segmented for the survey application at the time the latter is actively planned. To illustrate further, the master sample may be 10,000 *EA*s, of which 1,000 are sub-sampled for an employment survey. A household listing is undertaken in the 1000 *EA*s from which a second-stage sample of 15 households in each *EA* is chosen for the survey. The following year, another sub-sample of 800 *EA*s is selected from the master sample to be used in a health survey; and so on. In this way no *EA* is used more than once, so the size of the *PSU* is irrelevant.

62. The size of the *PSU* is important, however, when all the sub-samples generated by the master frame must come from the same set of *PSU*s; in the example above, *EA*s are the *PSU*s and a different subset of *EA*s is used in each sub-sample. When the former is the case, the design cannot be two stages but must be three or more. The method of sample selection of the *PSU*s in a master sample is not particularly an issue because they would be selected in the same way whether from a master sample or not. Generally the method would be by probability proportionate to size, *pps*, except in some rare cases where the *PSU*s are more or less equal in size; in the latter case, an equal probability sample of *PSU*s could be used.

4.2.3 Use of master samples to support surveys

63. In subsection 2.2.7 it was discussed how a large sample is needed for master samples in order to provide enough households to support multiple surveys over several years without having to interview the same respondents repeatedly. The anticipated sample sizes for all the proposed and potential surveys that may utilize the master sample frame are key parameters in designing its framework. For example if it is anticipated that 50,000 households would be interviewed in the various surveys to be served by the master sample, the sampling team would

have the basic information it needs to decide on the number and size of *PSU*s. Moreover, a plan of survey implementation can be developed in terms of use of the master sample, as shown in the following example (see also the illustration in subsection 2.2.7 for comparison):

- ***Example***

As in the example of 3.2.7 the master sample in Country A comprises 50,000 households. Three planned surveys will have sample and cluster sizes as follows: 16,000 households, 6 households per cluster for the income and expenditures survey; 12,000 households, 12 per cluster for the labour force survey; and 10,000 households, 20 per cluster for the health survey. The different cluster sizes are chosen to cope with the differential – by type of survey - effects of *deff*. In addition, there are 12,000 households to be held in reserve for other surveys if needed. The three planned surveys imply a total number of 4,167 *PSU*s (16000/6 + 12000/12 + 10000/20). Since the content of the surveys that might use the reserve sub-sample is unknown, it is decided to plan on a cluster size of 12, which adds another 1,000 *PSU*s for a grand total of 5,167. The master sample design team decides therefore to construct a master sample of 5,200 *PSU*s. The definition of the *PSU* must take account of the number of households to be interviewed. In this illustration each *PSU* must be large enough to yield 50 households for interview. With this information, the sampling team can then determine which geographical unit best serves to define the *PSU*s. If Country A has *EA*s that average 100 households with little variation around that average, then it would make sense to use *EA*s as the *PSU*s.

4.2.3.1 Advantages of multiple use of a master sample frame

There are distinct advantages to using a master sample. First and foremost the master sample plan serves as a coordinating tool for the line ministries and others who have a stake in any national statistical programme. This applies with respect to several aspects of survey-taking beyond sampling considerations per se, mainly in controlling costs and developing standardized procedures across sectors with respect to statistical definitions, wording of survey questions and data coding procedures.

64. A key advantage in a master sample programme is that of using the same PSUs. A field staff can be organized and maintained for the life cycle of the master sample. For example, interviewers can be hired, trained and available at the beginning of a master sample programme when it is known where the PSUs are located for all surveys that are to use the frame for, say, 10 years. To the extent necessary the interviewers can be hired locally from residents in or near the master sample PSUs. Survey materials such as PSU maps and household listing sheets can be generated at the start of the master sample programme, thus saving time as well as amortizing a significant portion of survey operating costs over all the anticipated surveys. Again it is important to emphasize, however, that such benefits only accrue when the master sample is to be heavily utilized.

65. In general, other advantages of master samples whether utilizing the same PSUs in a three-stage design or different PSUs in a two-stage design, include (1) the potential for integrating data, analytically, from two or more applications of the master sample using different content and (2) the potential to respond quickly to unforeseen data collection needs.

4.2.3.2 Limitations of multiple use of a master sample frame

66. There are some limitations to a master sample such as the possibility of exhausting the PSUs, that is, running out of households if the master frame is over-utilized. This, however, may be forestalled through adequate advance planning. Although it is not plausible to foresee all the uses that may be made of a master sample throughout its life cycle, reserve sub-subsamples can be designated for possible use so long as the master sample is big enough.

67. Another limitation is an ever-increasing amount of bias that accrues when appropriate up-dating is not carried through as the master sample ages. Finally, master samples are not well-suited to provide data on "special requirement" surveys such as those for particular provinces or rare sub-populations that may be of interest.

4.2.4 Allocation across domains (administrative regions, etc.)

68. National statistical offices have been under increasing pressure to tabulate and analyze their household survey data for important sub-national administrative areas such as major regions, provinces and large cities. Some countries such as Vietnam are even expected to routinely provide data at the district level. These requirements and expectations are driven by legitimate policy needs, generally on the grounds that socioeconomic programmes are focused on and developed for local areas as opposed to the nation as a whole.

69. In the parlance of statistical sampling these are domain estimates, previously discussed. They come at a heavy cost because the sample sizes necessary to achieve reliable results are enormous and often beyond the survey budgets that governments can generally muster. The need for domain data also affects the development of a master sample frame.

70. Considerations of how many domains to establish and of what type were discussed in section 3.2.4 on sample sizes and will not be repeated here. Once those decisions have been made the master sample frame can be constructed. As an example, one country may decide that surveys under its master sample programme are to provide data for only two domains – urban and rural. Another country with 12 provinces may want to provide estimates for each of them and decides that its survey resources can support the extra sample sizes needed if the provinces are treated as domains. A third country with 50 provinces may decide that it is too costly to produce estimates for each of them. Instead it may decide to define as domains the 8 major geographic regions into which the 50 provinces are divided. A fourth country may decide not to establish domains per se but instead tabulate its proportionately-allocated national sample by region, province, urban, rural and for selected large cities, intending to release to the public those sub-areas for which the sample size is deemed large enough to give reasonably reliable results.

71. For the last example in the preceding paragraph domain allocation is not relevant because the sample is proportionately distributed among the sub-national areas of interest. For the first three of the examples special steps must be taken to allocate the master sample PSUs appropriately. Since domain estimation implies equal reliability for each of the sub-population

groups or areas defined as a domain, the same number of sample PSUs should be selected in each domain – a requirement that pertains irrespective of whether the sample design is based on a master sample or not.

4.2.5 Maintenance and updating of master samples

72. In terms of its effect on population coverage proper maintenance of a master sample frame is a key element in its development and in planning for its use. The master sample of a given country is typically used for the decade between censuses, during which time far-reaching shifts in population movement are likely to occur. It is necessary to up-date the frame periodically to reflect population changes so that it will continue to be "representative."

73. Two types of up-dating are important. First, and this is the simplest to achieve, is to prepare new listings of households in the sample clusters selected at the penultimate stage. That procedure is generally recommended throughout this handbook, whether for master samples or single-use sample designs. In doing so, the sample clusters are automatically up-dated to reflect migration, births and deaths. This type of up-dating, confined to the sample clusters, helps to minimize coverage (nonsampling) error but the sampling variance increases over time unless the entire frame is up-dated.

74. There is also the need to periodically up-date the entire frame to properly account for postcensal growth on a large scale. As discussed previously this growth is of the type such as that which occurs in high-rise residential construction and expansion of squatter areas in cities. The EAs in which such high-growth takes place are invariably much smaller at the time the master sample frame is constructed. As a result, their measures of size become seriously understated as growth takes place. Thus their chances of selection in a pps design are minimized. The effect on the sampling variance can be drastic when such EAs do happen to be selected, because the current MOS may be larger than the original by orders of magnitude.

75. Problems of high-growth areas and their damaging effect on master samples can be reduced significantly by revising the frame regularly, say, every 2-3 years. Refer to subsections 3.8.1 and 3.8.2 for more discussion on methods of up-dating frames.

4.2.6 Rotation of PSUs in master samples

76. The reader is referred to subsection 3.8.2, "Sampling to estimate change or trend," in chapter 2 for a detailed discussion of the issue of sample overlap in connection with repetitive or continuing surveys intended to measure change or trends. Overlapping samples imply use of a sample scheme that utilizes replacement households when surveys are repeated. It is important to re-emphasize that overlapping samples are the preferred technique to estimate change in comparing, say, one year to the next. One method of replacement is sample rotation, which provides for partial overlap from survey to survey or occasion to occasion.

77. In the aforementioned subsection it was pointed out that both sampling reliability (desirable) and nonsampling error (undesirable) are greatest when the same households are used in each survey round. As a result a compromise is usually sought by using partial overlap in the

sample from one round to the next, especially when a survey is repeated three or more times; see subsection 3.8.2 for the rationale of partial overlap.

78. One method of introducing partial overlap is to replace or rotate the sample PSUs (as opposed to replacing the households within the sample PSUs). When there is a master sample of PSUs used not only for rounds of the same survey but for multiple surveys it is equally important to consider carefully the need to rotate them.

79. In order for a rotation plan to be feasibly implemented and to give meaningful results the degree of overlap between time periods should be the same and constant through time. For example, if the overlap between year 1 and year 2 is k percent, then it should also be k percent between year 2 and year 3, between year 3 and year 4, and so on. Accordingly, when entire PSUs are rotated this feature needs to be integrated into the rotation design.

4.2.7 Country examples of master samples

80. In this subsection we present descriptions of master samples in four developing countries – Cambodia, United Arab Emirates, Vietnam and Mozambique. Each illustrates some of the features and principles of master sampling such as one or two-stage sampling of master sample units and flexible application for particular surveys that are discussed in this chapter. In addition other features of sample design that are highlighted in the handbook are illustrated. These include, *interalia*, implicit stratification, optimum choice of cluster size to reduce the effects of *deff* and sample allocation for domains.

4.2.7.1 Cambodia 1998-99

81. The master sample of Cambodia illustrates the use of a two-stage design. A large sample of PSUs supplies a master list of second-stage area segments that are sub-sampled for use in particular surveys.

82. Cambodia's National Institute of Statistics developed a master sample in 1999 to use in the Government's intercensal household survey programme. The latter consists of a periodic Socio-Economic Survey and, potentially, surveys on health, labour force, income and expenditures, demography and ad hoc surveys. The 1997 population census served as the frame for design of the master sample which was carried out in two phases. The first phase was a pps selection of the PSUs, defined as villages with the measure of size being the census count of households. Selection of the PSUs was performed as a computer operation. The second phase entailed creation of area segments within the selected PSUs, which was a manual operation.

83. It was decided to use villages as the PSUs because they are large enough, on average (245 households in urban areas and 155 in rural), to have enough households to accommodate several surveys during the intercensal period. Thus the burden of repeatedly interviewing the same respondents is avoided. The alternative of using census EAs was considered but discarded because they are only half the size of villages on average. Special populations that are transitory or institutionalized were not included in the master sample, nor were military barracks.

84. A total of 600 PSUs was selected in the master sample because it was felt that number would give enough spread throughout the country to represent all the provinces adequately. Implicit stratification was used in selecting the sample, effectuated by sorting the village file in geographic sequence - urban-rural by province, district and commune. Thus the master sample was proportionately allocated, automatically, by urban-rural and by province.

85. An interesting feature of the master sample in Cambodia was the second phase of the sampling operation. As mentioned this entailed creating area segments within the selected PSUs. Note here that the second phase of master sample construction is not to be confused with the concept of second stage of sample selection, the latter of which pertains to selection of households for particular surveys. Within each selected master sample PSU, area segments of size 10 households (on average) were formed through a clerical task as an office operation. In that context it did not entail any field work except in unusual cases, because the 1997 census

listing books and existing sketch maps were used. The number of segments created within each master sample PSU was computed as the number of census households divided by 10 and rounded to the nearest integer. For example a village containing 187 households per the 1997 census was divided into 19 segments.

86. The segments, so created, constituted the building blocks to be sampled or sub-sampled in connection with the application for particular surveys. Selection of one or more segments from all, or a subset of, the PSUs is done for each survey or survey round that utilizes the master sample. An important feature is that creation of the master sample in the manner described affords the opportunity for each of the particular surveys which utilize it to be self-weighting, depending upon the details of their sample designs.

87. A key advantage of the master sample design is that it allows much flexibility in terms of how it is sub-sampled for use in particular surveys. Selection of the clusters (that is, segments) for each survey can yield a different set if desired. The typical PSU contains about 18-30 segments, providing an ample number of segments in each PSU to sustain all surveys. Moreover, repetition of the Socio-economic Survey on an annual basis is possible with a different set of segments every year. Alternatively, sample overlap is also possible by retention of some of the segments (say, 25 percent) from year-to-year in a pattern of rotation, where 75 percent of the segments is replaced each year.

88. A limitation of the master sample design is the use of compact clusters (all the households in the sample segment are adjacent to each other). This increases the design effect somewhat over non-compact segments, that is, a systematic sample of households within a larger cluster. It was thought the design effect would be reduced to some degree by limiting the cluster size, however, which is the reason segments of only size 10 households were settled upon.

89. It was anticipated that up-dating of the sample would take place every two or three years. Although it was recognized that it is better to up-date the entire master sample, it was decided to up-date only in the PSUs for the particular sample survey being planned at that time. Up-dating consisted of field visits to prepare a new listing of households in the affected segments. The same land area in those segments that contained the original set of households was re-listed, one reason why the segment boundaries are so important.

90. Another interesting note about the Cambodian master sample is the cooperation that was sought from village headmen in up-dating operations. They are known to maintain careful registers of all the households in their villages and, moreover, the registers are routinely kept current. Their listings in most instances are thought to be quite accurate. In addition the village headmen were also invaluable resources in terms of identifying and locating the land area appropriate to any particular segment.

4.2.7.2 United Arab Emirates (UAE) 1999

91. The master sample of UAE illustrates two important design features. First, the master sample design employs special stratification to cope with UAE's two diverse populations –

citizens and non-citizens. Second the design illustrates how the standard segment design (see subsection 3.7.2) can be exploited to deal with the issues of varying EA sizes and an old census.

92. UAE's master sample is described by the Ministry of Planning as a super sample of 500 PSUs based upon the 1995 population census as its sampling frame. It is intended to be used for particular household surveys until the next population census is undertaken. The PSUs are defined as census EAs, or parts thereof, such that, on average, a PSU contains about 60 households - both citizen and non-citizen.

93. Two strata were constructed prior to selection of the PSUs. The first consisted of enumeration areas in which one-third or more of the households were citizen households at the time of the census. The second stratum is all other enumeration areas. A total of 1686 enumeration areas was classified in stratum I and 2,986 in stratum II. Using systematic selection with probability proportional to size, a sample of 250 PSUs was selected in each of the two strata for a total of 500 PSUs nationwide. It was expected that this master sample would yield approximately equal numbers of citizen households and non-citizen households. First, large PSUs (those with 90 households or more) were segmented and one segment was randomly chosen within each such large PSU. A new, current listing of households was undertaken in the 500 PSUs to bring the frame up to date. The listing operation resulted in approximately 30,000 households in the sample PSUs to be utilized in various combinations for particular surveys. To facilitate flexible application, the households in each sample PSU were divided into 12 subsets, or panels, of an expected 5 households each.

94. A noteworthy feature of the master sample is that it is not self-weighting because the two strata are of unequal sizes. The first survey to utilize the master sample was the 1999 National Diabetes Survey, sponsored by the Ministry of Health. Others that were expected to be fielded include a labour force survey and an income and expenditure survey (or family budget survey).

95. A few more details of certain special circumstances in UAE that dictated how the master sample design was constructed are presented in the next few paragraphs. Two overriding considerations that were carefully taken into account were the target populations and the sampling frame.

96. There are two important target populations in the country - citizens and non-citizens. While the former constituted about 43 percent of the population, according to the 1995 population census, they comprise only about one-quarter of the nation's households. This is because non-citizen households are much smaller in terms of number of persons per household. The implication for sample design is that if a proportionate sample of the nation's households were selected nearly three-quarters of the responding households would be non-citizens. It further implies that the reliability of the resulting survey estimates for non-citizen households would be about 3 times better than that for citizen households. As the results were to be used to develop policy and plan programmes, such a disparity in the reliability of the estimates was not thought to be desirable or useful. The solution, in terms of sample design, was to treat the two disparate and unequal target populations as separate entities through application of proper stratification described three paragraphs above.

97. Another level of stratification was used as well. This was geographic stratification to ensure appropriate distribution of the sample by emirate and by urban-rural. The file of EAs was sorted in the following sequence prior to sample selection: first for the citizen stratum by urban, and within urban by emirate and within emirate by EA codes arranged in ascending order by the percent citizen, followed by rural, emirate and EA code; then the non-citizen stratum in the same sequence.

98. It was recognized that a key feature of the master sample frame must be a clear set of maps that delineate the area units to be designated as the sample areas, that is, the PSUs. The area units had to be small enough to be conveniently listed but, at the same time, large enough so that they could be clearly defined with respect to natural boundaries (for ease of location). Census enumeration areas were thought to be the only feasible area units that met these dual criteria. Unfortunately, maps were not used in the population census, and hence the existing enumeration areas were not clearly defined in terms of known boundaries. As a result it was necessary to ensure that good boundary information be developed for the master sample PSUs (EAs).

99. Preparations for the master sample of PSUs made use of the "standard segment design," already described in the preceding chapter. It is a methodology that has been used successfully in many countries through both the Demographic and Health Surveys programme and PAPCHILD (Pan Arab Survey of Maternal and Child Health).

100. It was decided to use the standard segment design because census enumeration areas in UAE are quite variable in size. Standard segments of approximately 60 households were created. The number of standard segments was calculated in each sample PSU as the total number of households (that is, citizen plus non-citizen) divided by 60 and rounded to the nearest integer.

101. For cases where the number of segments, that is, the measure of size was two or more, the PSU was divided into area segments. This required a field procedure in which a visit to the EA (PSU) was made and a sketch map was prepared using quick-counting and map-spotting of the dwellings (not the households). After segmentation one segment was chosen from each PSU at random. That segment became the actual geographic area for sampling in the master sample. Another visit to the field was made to obtain a current list of the households in each sample segment. The latter procedure was thought to be a vital component of the sampling operation in order to bring the three-year old master sample frame up to date.

102. The final operation for the master sample consisted of subdividing the newly listed households in each sample segment into 12 systematic subsets or panels. One or more of the panels was to be used for particular surveys. Since the average size of the segment was about 60 households, each panel contains 5 households on average.

103. The main reason for having 12 panels was because of the flexibility this number provided in forming combinations for use in surveys. The actual choice for a given survey depended on various factors including the objectives of the survey, the desired cluster size and

the overall sample size required for the survey. For example, two-fifths of the PSUs were to be used in the National Diabetes Survey. Hence, 4 of the 12 panels of households within those PSUs were included. This combination yielded an overall sample plan of 200 PSUs with clusters of size 20 (that is, 4 times 5 households) and a total sample size of approximately 4,000 households.

4.2.7.3 Vietnam 2001

104. The master sample of Vietnam has two distinguishing features. It demonstrates use of two stages in selecting the master sample and a third stage when applied to particular surveys. Secondly, it demonstrates how a master sample can be allocated to geographic domains.

105. The master sample is based on the 1999 census as the sampling frame and is a two-stage design. PSUs were defined as communes in rural areas and wards in urban areas. They were defined in this way because it was decided that a minimum of 300 households would be necessary in each PSU to serve the master sample. Alternatively, EAs were considered as PSUs but they were too small and would have had to be combined with adjacent EAs in order to qualify satisfactorily as PSUs. The latter task was thought to be much too tedious and time-consuming. The number of communes/wards, on the other hand, that had to be combined because of their small size was only 529 of the more than 10,000.

106. A total of 3,000 PSUs was selected by pps for the master sample. Each sample PSU contained, on average, 25 EAs in urban areas and 14 in rural areas. For the second stage of selection three EAs were selected in each sample PSU, using pps. The second-stage units, EAs, contain an average of approximately 100 households according to the 1999 census – 105 in urban areas and 99 in rural areas.

107. An objective of the master sample was to be able to provide fairly reliable data for each of Vietnam's 8 geographical regions. Sample selection was undertaken independently within each province. Thus, provinces served as strata for the master sample. It was desired to over-select the sample in certain small provinces that contain very small populations. Accordingly, the allocation of the sample among provinces was made by the method of pp√s, or probability proportionate to the square root of the size of the province. Proportional allocation between urban and rural areas was used.

108. In addition to the provincial-level stratification mentioned above, implicit geographical stratification within provinces was used. In applying the master sample to specific surveys, subsets of the EAs were to be used – for example, one-third of them for the Multi-purpose Household Survey. For survey applications a third stage of selection is administered in which a fixed number of households is selected from each sample EA. That number may vary by survey and by urban-rural. For example, 20 households per *EA* might be chosen for rural *EA*s and 10 per *EA* for urban ones.

4.2.7.4 Mozambique 1998-99

109. The master sample of Mozambique illustrates a case in which a single-stage selection of PSUs was selected to be used for all of the government-sponsored, national surveys in the

nation's intercensal household survey programme. It also illustrates how a flexible master sample can be adapted to meet the measurement objectives of a particular survey.

110. Master sample PSUs in Mozambique were defined and sampled in a straightforward way as described in various parts of this handbook. The PSUs were constructed from the 1997 Population Census as the sample frame. They consist of geographical groupings of, generally, 3-7 census EAs, the latter of which contain about 100 households on average. The master sample PSUs were selected using probability proportionate to size, or pps, sampling.

111. A total of 1,511 PSUs was selected to provide the framework for sampling to apply for Mozambique's integrated system of household surveys. The master sample PSUs were divided into panels, each comprising a systematic subset and therefore constituting a probability sample in its own right. There are 10 such panels of about 151 PSUs each. In the 5-Year (2000-2004) Plan, the Core Welfare Indicator's Questionnaire (CWIQ)18 that was conceived by the World Bank was the first survey to make use of the master sample.

112. The sample plan for QUIBB was designed with two measurement objectives in mind. The first was to obtain the relevant indicators necessary to profile the well-being of persons and households in Mozambique. The second was to provide reliable estimates of these indicators at the national level, for urban and rural areas separately and for each of the 11 provinces in the country. The sampling methodology for QUIBB utilized the Mozambique master sample to choose about 14,500 households in a stratified, clustered design. Accordingly, the first stage of selection was of course the master sample PSUs.

113. The second stage of selection for QUIBB was a sub-sample of the master sample PSUs. A total of 675 PSUs was sub-sampled from the 1511 in the master sample. They were selected systematically with equal probability and, moreover, equally allocated among the 11 provinces of Mozambique. At the third stage a sample of one enumeration area was selected in each of the QUIBB PSUs. Thus there are 675 clusters in the QUIBB sample – 475 rural and 200 urban. The enumeration areas were selected with equal probability because their sizes are roughly the same – as mentioned above, about 100 households on average, though there is variability. The final stage of selection occurred following field work in which the Instituto Nacional de Estatistica (INE) visited the clusters to compile a fresh list of households to bring the 1997 sample frame up to date. From the lists so compiled, a systematic sample of 20 households in rural areas and 25 in urban was selected for the QUIBB survey interviews. Sample selection for QUIBB was thus a four-stage selection process, although the master sample upon which it was based was a single stage.

114. Two design features for QUIBB illustrate the flexibility with which the master sample can be adapted to fit the particular requirements for a survey application.

115. First, in utilizing the master sample for QUIBB there was interest on the part of INE to use the aforementioned panels that had already been designated. With the desire to have about

[18] The Portuguese acronym is QUIBB (Questionario Indicadores Basicos Bem-Estar) and that term is used to refer to the survey in this document.

600 PSUs for QUIBB it was hoped that 4 of the panels could be used. The idea was discarded, however, when it was realized that the number of panels that would be necessary for QUIBB would be different for each province. This is because the measurement objective required more or less the same sample size by province. Instead, the 675 PSUs necessary for QUIBB were selected systematically from the entire master sample file without regard to panels. This departure from the original intent of the master sample arose because equal reliability was wanted at the provincial level for QUIBB. That is in contrast to the original design of the master sample, that is, to provide a proportionate sample by province. In the master sample plan, as originally conceived, national level estimates were expected to take precedence.

116. A second important issue regarding the QUIBB sample design was the cluster size. It was agreed that cluster sizes should be different for urban and rural households on the grounds that the sample design effect, or deff, is higher in rural areas where most of the economic livelihood is subsistence farming. In other words, each household in a rural area was likely to have very similar characteristics. The master sample afforded the possibility of selecting a different fixed number of households (25 and 20 respectively in urban and rural areas) in the final stage.

4.3 Summary guidelines

117. This section summarizes the main guidelines to be inferred from this chapter. As in the preceding chapter, they are presented more as "rules of thumb" rather than fixed recommendations. In checklist format, they are as follows:

- Use sample frames that are as complete, accurate and current as possible;
- Ensure that the sample frame covers the intended target population;
- Use most recent census as frame for household surveys if possible;
- Define *PSU*s in frame from area units such as census *EA*s with mapped, well-delineated boundaries and for which population figures are available;
- Use census list of households as frame at last stage only if very recent – usually no more than one-year old;
- Use dual or multiple frames with caution by ensuring procedures are in place to deal with duplications;
- Update census frame if more than two years old – nationwide or in specifically targeted areas known to have high growth;
    - Use quick-count canvassing to up-date old frame.
    - In sample clusters, update by making fresh list of households.
- Update only in sample clusters if census frame is no more than two years old;
    - Up-date by making fresh list of households;
- Use master sample or master sample frame only when large-scale, continuing survey programme is planned or underway;
- Define master sample *PSU*s that are large enough or numerous enough to sustain many surveys, or repeat survey rounds, during intercensal period;
- Update master sample frames along same guidelines as above for single-survey frames;
- Employ system of sample rotation – either households or *PSU*s – in repeat surveys that use master samples.

References and further reading

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.

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.

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

Pettersson, H. (2001), Mission Report: Recommendations Regarding Design of Master Sample for Household Surveys of Vietnam 25 November (unpublished), General Statistical Office, Hanoi.

_____ (2003), *The Design of Master Sampling Frames and Master Samples for Sample Surveys in Developing Countries*, United Nations Statistics Division, New York.

Turner, A. (1998), Mission Report to Kingdom of Cambodia, National Institute of Statistics 11-24 November (unpublished), National Institute of Statistics, Phnom Penh.

_____ (1999), Mission Report to United Arab Emirates, Ministry of Health and Central Department of Statistics 23 January – 3 February (unpublished), Central Department of Statistics, Abu Dhabi.

_____ (2000), Mission Report to Mozambique, Instituto Nacional de Estatistica 13-26 August (unpublished), Instituto Nacional de Estatistica, Maputo.

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.

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

Chapter 5

Documentation and Evaluation of Sample Designs

5.0 Introduction

1. This chapter, though comparatively short, nevertheless has a central role in the handbook. Documentation and evaluation of sample designs in particular and survey methodology in general are too often neglected in the rush to release survey findings. This is especially true iIn countries with little prior experience in conducting household surveys, metadata are often poorly documented in survey worksheets and reports. In some countries documentation of survey procedures including sample implementation is little valued. Thus there is no tradition in place to require it.

2. Evaluation of the survey results is often completely ignored. In this way errors creep into survey analysis. This is generally due to the fact that budgetary restrictions often preclude any formal studies or methods to assess the abundance of nonsampling errors that crop up in household surveys. Yet there are other barometers of data quality that should be readily available (such as rate of non-response) and these, too, are often not mentioned in survey reports.

3. It also emphasizes the importance of presenting relevant information to users on known limitations of the data, even when formal evaluation studies have not been conducted. In this regard it is important to note that it is beyond the scope of the chapter to discuss techniques for conducting formal evaluations of survey methodology, of which there are many.19 Instead it focuses on what information should be given to users to help them and evaluating gauge the quality of the survey – concentrating on sampling aspects.

5.1 Need for, and types of, sample documentation and evaluation

4. There are two types of documentation needed in household surveys. First, there is the need to keep careful records of the survey and sampling procedures as they are being carried out operationally in the survey process. Without such documentation errors creep into survey analysis. For example, probabilities of selection may not be fully known at the time of analysis without careful record-keeping.

5. The sampling technician should therefore take necessary steps to carefully document not only the sample plan for the particular survey undertaken but also its implementation. Sample designs often require adaptations at various stages of the field work because of unforeseen situations that arise in the conduct of the survey. It is important to record - step-by-step - all the procedures used in carrying out the sample plan to make sure the implementation is faithful to the design. When it is not it is even more important to document all the departures from the

[19] Special studies that are designed to evaluate specific types of nonsampling error in surveys include re-interview surveys (for evaluating response variability), post-enumeration surveys (for coverage evaluation), inter-penetrating samples (for evaluating interviewer variability), reverse record checks (for evaluating respondent recall errors) and so forth.

design, even minor ones. This information is necessary later at the analysis stage, in case any adjustments need to be made. Equally as important, documentation of this type is indispensable for planning future surveys.

6. The second type of documentation is report-writing. There ought to be two kinds of technical reports prepared for every household survey. One is a fairly brief, user-friendly description of the survey methodology including the sample plan and its implementation. This report would typically comprise the "technical" section (or an appendix) of the various substantive reports that are released to discuss and interpret the substantive findings of the survey.20 It should include a sub-section on what is known about the limitations of the data, discussed further below.

7. The other type of technical report should be a more detailed description of the survey methodology. That would be a report standing on its own (not part of the substantive series), intended more for professional researchers, social scientists and statisticians than policy-makers or the general public. An excellent example of the latter is (U.S. Bureau of the Census, 1978). Of course it is preferable if the detailed technical report and regular survey reports are produced concurrently, although the former is typically done, if at all, much later. It is useful, also, to consider having the technical report, or an abridged version, published in a statistical journal to ensure its longevity.

8. Documentation of both types discussed above is so important that it is recommended that National Statistical Offices assign a special office or officer to do it routinely for household surveys.

5.2 Labels for design variables

9. This section and the next few, sections 4.3 through 4.6, discuss documentation of the first type – record-keeping of survey processes related to sampling.

10. The identification of the units of selection at each stage must be clearly and uniquely labeled. In a multi-stage design this will mean establishing codes for the primary, secondary, tertiary and ultimate sampling units (depending upon how many stages are in the design). Normally a four-digit code will suffice for the first stage of selection and a three-digit code for the remaining stages. Geographic domains must also be properly labeled. In addition, the administrative codes identifying the geographic, administrative structure of the areas to which the sampling units belong should be part of the labelling process.

- ***Example***

Suppose a sample of 1200 *PSU*s, defined as census *EA*s, is selected for a two-stage design – 600 in each of two domains defined as urban and rural. A convenient way to code the *PSU*s is 0001 through 1, 200. Moreover, it is also useful to assign those codes in the same sequence that was used to select the *PSU*s. That feature may be needed for use in calculation of sampling variances. Thus if the rural *PSU*s were selected first they would be coded 0001 to 0600, while

[20] Guidelines on what to include in the report (on findings) from a sample survey can be found in the report of a United Nations Sub-commission on Statistical Sampling, 1964.

the urban ones would be coded 0601 to 1,200. Such a coding scheme has two advantages. First, each *PSU* is uniquely numbered and identified. and sSecondly, analysts can tell at a glance whether a *PSU* is urban or rural simply by its ID. In the second stage of the sample, each *PSU* is listed and 20 households are selected for interview. In this stage all listed households would be given a three-digit code (or four digits if some *EA*s were to contain more than 999 households), again in the sequence in which they are listed. The sample households would retain the code assigned in this manner, as opposed to assigning, say, codes 01-20 for the selected ones. Finally, administrative codes are assigned as necessary. Thus a sample household that might be coded as 09 003 008 0128 080 would identify it as the 80th household listed (and selected for interview) in *PSU* 0128, which belongs to civil division 008 in district 003 of province 09. Moreover, the *PSU* number instantly identifies the household as belonging to the rural domain. If the survey obtains information about the members of the households, each one of them would also carry a unique code of two-digits, 01 to 99.

11. It is perhaps apparent why proper labelling is essential. One clear reason is for quality control. As assignments are made to interviewers and questionnaires are returned from the field, they can be checked off against a master list to make sure that all sample households are accounted for. Secondly, the unique numbering systems is invaluable to the data processing staff because it allows tabulations to be made by geographic location.

12. For countries that have multi-survey programmes, it is highly desirable that design variables be labelled in a consistent, standardized fashion across all surveys. This has obvious advantages in data processing and in presentation of results by eliminating confusion among both the producers and users of the data.

13. In regard to the latter point, a multi-survey programme would benefit by assigning PSU codes to the entire universe of PSUs rather than just the sample ones, as was described in the preceding example. This is because different PSUs are often sampled for the different surveys, which is frequently the case when master samples are employed.

14. In general, master samples require design labels even more than one-time survey samples. A key use of master samples, as discussed previously, is in repeat rounds of the same survey. Proper labelling of the design variables that identify the stages of selection is crucial, in order to keep track of the cases that comprise overlapping portions of the sample from one survey to the next. Often, rotation panels (systematic subsets of the full sample) are designated for the purpose of facilitating the identification of units (households, clusters or PSUs) to be replaced in subsequent rounds of the survey. They of course require their own panel identification codes. Moreover, households that are added to the master sample during periodic up-dates must be properly coded. It should be done in such a way that the code scheme distinguishes new households from old ones.

5.3 Selection probabilities

15. One item of information that is often overlooked in sample documentation is calculation of the probabilities of selection at the various stages. Where information does exist it is often confined to the overall sample weight (from which the overall probability can be readily calculated) for each sample case.

16. A particularly important detail for proper documentation occurs when sub-sampling is done in the field during data collection. ItThis may happen when a sample segment/cluster is too large. For example, we discussed earlier that an unexpectedly large cluster may have to be segmented into, say, four parts of roughly equal size. One part is then selected at random for listing and interviewing. In that case the overall probability of the sampled segment (and the households/persons selected within) is one-fourth that of the original cluster; its weight is thus the inverse, or four. That weighting factor has to be reflected in the calculations when the data are analyzed.

17. Sub-sampling may also occur when there is more than one household in a dwelling (when the dwelling is the listing unit). One option, which is unbiased, is for the survey manager to interview all the households found; this is often the choice if there is only two. But if it were, say, five where one was expected, cost considerations may dictate that only one of them be interviewed – randomly picked of course. Again, careful recording of the sub-sample rate (1/5 in this example) is essential so that the probability of selection for the affected household can be accurately calculated by the sampling staff and the weight thereby properly adjusted (by a factor of five).

18. It is also useful to record the probabilities of selection at each stage, however, as mentioned in the opening paragraph of this section. For example, the probability of selecting each PSU is different whenever pps sampling is used. This is true even if the overall sample design is self-weighting. If the probabilities of selection of the PSUs are ignored or not recorded erreneouly it is may not be possible to properly figure out the overall sampling weights under certain conditions. First, maybe if those same PSUs ought to be used in another survey later, in which the last selection-stage differs from the original survey (say, fixed-rate sampling instead of fixed-size). Second, those PSUs might be sub-sampled for subsequent use in other surveys. It is highly useful to know what the original probabilities of selection are, in order to accurately determine the sub-sampling procedures.

5.4 Response rates and coverage rates at various stages of sample selection

19. As part of the evaluation process to examine the implementation of the sample survey, it is essential to provide information to users on response rates and coverage rates. It is useful to make as much detail as possible available. Thus it is important to provide not only the rate of response (or its complement, rate of non-response), but also a tabulation of the reasons for non-response. Categories of non-response would likely include the following:

- No one at home;
- Vacant dwelling unit;
- Demolished or uninhabitable dwelling uni;.

- Refusal.
- Away temporarily (holiday, etc.).

20. The definition of response rate may vary from country to country, in terms of which categories are included. Typically, however, completed response includes the first, fourth and fifth of the above categories. Those categories comprise cases in which a response should be obtained if at all possible. Vacant and demolished units are usually ignored (in calculating the response rate) on the grounds that it is not possible, by definition, to obtain a response in such units. Thus for example, a country may select 5,000 households with the following results: 4,772 completed interviews, 75 cases of "no one at home," 31 vacant dwellings, 17 demolished units, 12 refusals and 93 cases of "away temporarily." The response rate would be calculated, ordinarily excluding the vacant and demolished units, as 4772/(5000-31-17), or 96.4 percent.

21. When the survey target populations include both households – for variables such as household income or access to services – and individuals (say, health status of adult women), it is customary to calculate response rates at both the household and individual level. For example, 98 percent of the households may respond, but within the responding households a small percentage of the individuals may be nonrespondents.

22. Often, whole clusters are not interviewed for various reasons including issues of security such as civil strife or disorder and lack of accessibility due to the terrain or weather. Frequently when such problems occur substitute clusters are selected, a procedure that is seriously biased because the inhabitants of the substitute clusters are almost always likely to differ in very significant ways from those in the replaced clusters. Nevertheless, when such substitutions are made it is incumbent upon the survey team to record the number and location of such clusters. Moreover, it is also important to provide some information on under-coverage in such cases. This might be done by estimating, to the extent possible, the number of persons in the target population(s) thought to reside in the areas that the replaced clusters represent.

23. It is useful to note that problems of the type mentioned in the preceding paragraph can be reduced somewhat by identifying in advance of sample selection the areas of the country that are "out of scope" for survey interviewing due to security or accessibility concerns. Those identified should be excluded from the survey universe before sampling. An essential part of the survey documentation in such cases is careful discussion in, and the survey reports, should mentioning clearly that these areas are not "represented" by the sample.

5.5 Weighting: base weights, non-response and other adjustments

24. Calculation of survey weights is presented in the next chapter of the handbook. Here we emphasize the importance of documenting those calculations.

25. Weighting for household surveys generally involves up to three operations – calculation of the base or design weights, adjustments for non-response and adjustments for post-stratification. In many applications only the design weights are used, while in others the design weights may be adjusted by an additional factor to reflect non-response. In comparatively few applications the weighting may reflect another factor, either with or without non-response adjustments, intended to adjust the population distribution obtained from the sample to agree

with the distribution from an independent source of data such as a recent census. The latter is often referred to as post-stratified weighting. In some applications no weighting is done at all; this would occur only when two conditions are met: the sample is completely self-weighting and the data generated are restricted to percentage distributions, proportions and ratios, as opposed to estimated totals or absolutes.

26. When weighting is used it is necessary of course to carefully record the calculations. As mentioned previously the weights (or probabilities) at each stage of selection should be calculated and recorded. Also, separate weights at each phase of data operations should be recorded, that is, (1) design weights, (2) design weights after multiplication by the non-response adjustment factor(s) and (3) the latter after adjustment factors for post-stratification have been applied.

27. It is important to note that design weights will differ for each domain whenever the sample design includes domain estimation. In other words, even when the sample is self-weighting within domains, each domain will have its own distinct weight. Furthermore, each domain will have a or different set of weights if the design is not self-weighting within domains. In addition, it should be noted that non-response adjustments are often applied separately by important geographic sub-areas such as major regions, irrespective of whether domain estimation is present in the design. Finally, the design weight itself may be multiplied by an additional factor for particular clusters or households. This would be done whenever sub-sampling (discussed above in section 4.3) is used.

5.6 Information on sampling costs

28. While household surveys are usually budgeted very carefully it is equally important to keep records of actual expenditures of its their various operations. In keeping with the emphasis of this handbook this section covers only the costs of regarding the sampling operations. Record-keeping this on costs is especially important when useful for master sample designs are used. It is also indispensable for planning the sampling aspects of future surveys.

29. When master samples are utilized there is an initial, large start-up cost to effectuate its development. It generally includes aspects of (a) computer-manipulation of census files to establish the sample frame, (b) mapping or cartographic work to create PSUs and (c) computer-selection of sample PSUs. As mentioned in the preceding chapter the cost of the start-up operations is often shared by the Ministries that will make use of the master sample during its life cycle. Those costs should also be distributed over all the surveys for which the master sample is intended to be used, to the extent that they are all known about in advance. It is therefore essential that very careful record-keeping of the master sample development be achieved. as well as for planning the sampling aspects of future surveys.

30. Once the master sample is in place, records on cost need to be compiled with respect to maintaining it. As noted previously up-dating of master samples takes place periodically and that, of course, needs to be carefully monitored regarding its cost.

31. Sampling operations for which cost figures ought to be regularly obtained include those in the list that follows. The list applies to both one-time sample surveys and master samples.

(1) Salaries for sample design including fees for any outside consultant.
(2) Field costs for up-dating the sample frame including personnel and preparation of auxiliary materials such as maps.
(3) Computer costs to prepare the sample frame for selecting the sample of *PSU*s.
(4) Personnel costs to select the sample of *PSU*s (if not done by computer).
(5) Field costs to conduct the listing operation in the penultimate-stage sampling units including personnel and the preparation of materials such as cluster folders.
(6) Personnel costs to select the sample of households within the sample clusters.

5.7 Evaluation – limitations of survey data

32. Much of the documentation on proper record-keeping discussed in the preceding subsections is useful for evaluating aspects of the sample design and survey implementation, in addition to its importance in processing the survey results. Information on response rates, for example, helps to assess whether bias from non-response is serious or not. Sampling cost information may be used to evaluate the "economic" effectiveness of the sample design and its utility for future surveys.

33. As stated previously, formal evaluation of sample surveys covers multiple facets of nonsampling error – much beyond the scope of this handbook. See [United Nations, 1982], however, for a comprehensive treatment of the subject. On the other hand, sampling error can be estimated and that is discussed further below.

34. Despite the fact that formal evaluation studies are not often done for household surveys, it is nevertheless crucial that the survey documentation include information on the limitations of the data. A brief section of the substantive reports on findings should be devoted to this subject, often simply entitled, "Limitations of the survey data." In it, the reader must be informed of both sampling and nonsampling aspects of survey error.

35. A valuable publication that describes how to present information on survey errors is "Standards for Discussion and Presentation of Errors in Data," by the U.S. Bureau of the Census, 1974. Following are specimen paragraphs from it that are suggested as the kind of information that should be presented to users when survey findings are released:

> "The statistics in this report are estimates derived from a sample survey. There are two types of errors possible in an estimate based on a sample survey – sampling and nonsampling. Sampling errors occur because observations are made only on a sample, not on the entire population. Nonsampling errors can be attributed to many sources: inability to obtain information about all cases in the sample, definitional difficulties, differences in the interpretation of questions, inability or unwillingness to provide correct information on the part of respondents, mistakes in recording or coding the data obtained, and other errors of collection, response, processing, coverage, and estimation for missing

> data. Nonsampling errors also occur in complete censuses. The accuracy of a survey result is determined by the joint effects of sampling and nonsampling errors.
>
> The particular sample used in this survey is one of a large number of all possible samples of the same size that could have been selected using the same sample design. Estimates derived from the different samples would differ from each other. The deviation of a sample estimate from the average of all possible samples is called the sampling error. The standard error of a survey estimate is a measure of the variation among the estimates from the possible samples and thus is a measure of the precision with which an estimate from a particular sample approximates the average result of all possible samples. The relative standard error is defined as the standard error divided by the value being estimated.
>
> As calculated for this report, the standard error also partially measures the effect of nonsampling errors but does not measure any systematic biases in the data. Bias is the difference, averaged over all possible samples, between the estimate and the desired value. Obviously, the accuracy of a survey result depends upon both the sampling and nonsampling errors, measured by the standard error, and the bias and other types of nonsampling error, not measured by the standard error."[21]

36. As implied above an important component of sample evaluation is estimation of sampling errors, which should be undertaken for the key survey estimates. As mentioned discussed previously, one of the distinguishing characteristics of a probability sample is that the sample itself can be used to estimate standard errors. Methods of variance standard error estimation are discussed in detail in chapter 6. In addition there are efficient and reliable software packages available to estimate standard errors, which should be taken advantage of whenever possible.

37. Generally, estimates of standard errors are prepared for the key characteristics of interest in the survey, since it is neither practical nor necessary to calculate them for all the items. The standard errors of course provide the means for users to evaluate the reliability of the survey estimates and to construct confidence intervals around the point estimates.

38. The standard errors may also be used to evaluate the sample design itself. A particularly useful statistic for doing this is the sample design effect, deff, or more precisely, deft, the square root of deff. It is fairly straightforward to calculate deft for every data item for which the standard error is estimated. It only entails dividing the estimated standard error, for a given item, by the standard error from a simple random sample of the same sample size, namely, pq/n, where p is the estimated proportion; q is 1-p and n is the sample size. The exercise serves to confirm or refute the design effects that were assumed when the sample was being designed, since the actual deffs or defts cannot be known until after the survey has been conducted, the data processed and the standard errors estimated.

39. The sampling statistician can use the calculated design effects to evaluate whether the cluster sizes are of reasonable size for key data items and take corrective action if necessary. For

[21] Source: United States Bureau of the Census (1974), *Standards for Discussion and Presentation of Errors in Data, Technical Paper 32*, Washington, Appendix I, page I-1.

example, if *deft* is much larger than anticipated for certain key items, the sample for a future survey may be designed to use smaller cluster sizes.

5.8 Summary guidelines

40. This section summarizes the main guidelines from this chapter. As in preceding chapters, they are presented more as "rules of thumb" rather than fixed recommendations. In checklist format, they are as follows:

- Document sampling aspects of surveys in two ways – proper record-keeping and provision of technical information to users;
- Keep detailed records of sampling processes including costs;
- Develop codes for sample design variables – administrative areas, *PSU*s, clusters, households, persons, etc;
- Strive for standardized coding of design variables that are consistent over all surveys;
- Record all deviations or departures from original sample plan that occur in implementation of survey;
- Calculate and record probabilities of selection at each stage of sampling;
- Record, especially, information on sub-sampling that occurs in field work;
- Record information on the number and types of non-response;
- Record design weights, adjustments for non-response and post-stratified adjustments – separately for domains when latter are used;
- Keep detailed cost records of each operation in sample design and implementation;
- In master samples keep cost records of both its development (start-up) and maintenance;
- Prepare technical reports for users on sampling and survey methodology.
- Prepare brief report on limitations of data for all substantive publications that report on survey results;
- Prepare more intensive technical report on all aspects of sampling methodology.
- Calculate sampling errors for key variables and present these in technical reports.
- Calculate design effects (*deff* or *deft*) for key variables.
- Assign an officer to be in charge of documentation.

References and further reading

Casley, D.J. and Lury, D.A. (1981), *Data Collection in Developing Countries*, Clarendon Press, Oxford.

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

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.

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

World Bank (1999), *Core Welfare Indicators Questionnaire (CWIQ) Handbook,*" World Bank, Washington.

Chapter 6

Construction and Use of Sample Weights

6.1 Introduction

1. This chapter describes the various stages in the development of sample weights and their use in computing estimates of characteristics of interest from household survey data. In particular, the adjustment of sample weights to compensate for various imperfections in the selected sample is described. Attention is restricted to descriptive estimates that are widely produced in most survey reports. The important ideas presented are illustrated using real examples of current surveys conducted in developing countries, or ones that mimic real survey situations.

6.2 Need for sampling weights

2. Household surveys are, in general, based on complex sample designs, primarily to control cost. The resulting samples are likely to have imperfections that might lead to bias and other departures between the sample and the reference population. Such imperfections include the selection of units with unequal probabilities, non-coverage of the population and non-response. Sample weights are needed to correct these imperfections and thereby derive appropriate estimates of characteristics of interest. In summary, the purposes of weighting are to:

(i) Compensate for unequal probabilities of selection;

(ii) Compensate for (unit) non-response; and

(iii) Adjust the weighted sample distribution for key variables of interest (for example, age, race, and sex) to make it conform to a known population distribution.

3. We shall discuss in detail the procedures used for each of these scenarios in the sections that follow. Once the imperfections in the sample are compensated for, weights can then be used in the estimation of population characteristics of interest and also in the estimation of the sampling errors of the survey estimates generated.

4. When weights are not used to compensate for differential selection rates within strata (whenever the sample is so-designed) and for the sample imperfections mentioned above, the resulting estimates of population parameters will, in general, be biased. See sections 6.3, 6.4 and 6.5 for examples of the weighting procedures employed under each scenario, including a comparison of the weighted and unweighted estimates in each case.

6.2.1 Overview

5. The rest of the chapter is organized as follows. Section 6.2 deals with the development of sample weights in the context of a multi-stage sample design, including the adjustment of sample weights to account for duplicates in the sample and for units whose eligibility for the survey is not known at the time of sample selection. Section 6.3 discusses weighting for unequal probabilities of selection. Several numerical examples are provided, including a case study of weight development for a national household survey. The section ends with a discussion of self-weighting samples. The issues of non-response and non-coverage in household surveys are addressed in sections 6.4 and 6.5 respectively. Sources and consequences of non-response and non-coverage are discussed. Methods for compensating for non-response and non-coverage are also presented, including numerical examples illustrating the adjustment of sample weights for non-response and non-coverage. Section 6.6 discusses the issue of inflation in the variance of survey estimates as a result of the use of sample weights in the analysis of household survey data. A numerical example is also provided to illustrate the calculation of the increase in variance due to weighting. Section 6.7 discusses the issue of weight trimming and presents an example of a trimming procedure by which the trimmed weights are re-scaled in such a way as to add up to the sum of the original weights. Finally, some concluding remarks are presented in section 6.8.

6.3 Development of sampling weights

6. Once the probabilities of selection of sampled units have been determined, the construction of sampling weights can begin. The probability of selection of a sampled unit depends on the sample design used to select the unit. Chapter 3 provides detailed descriptions of the most commonly used sampling designs and the probabilities of selection corresponding to these designs. It is assumed throughout that the probabilities of selection have been determined.

7. The development of sampling weights is sometimes considered to be the first step in the analysis of the survey data. It usually starts with the construction of the *base* or *design weight* for each sampled unit, to reflect their unequal probabilities of selection. The base weight of a sampled unit is the reciprocal of its probability of selection into the sample. In mathematical notation, if a unit is included in the sample with probability p_i, then its base weight, denoted by w_i, is given by

$$w_i = 1/p_i. \tag{6.1}$$

8. For example, a sampled unit selected with probability 1/50 represents 50 units in the population from which the sample was drawn. Thus sample weights act as inflation factors to represent the number of units in the survey population that are represented by the sample unit to which the weight is assigned. The sum of the sample weights provides an unbiased estimate of the total number of units in the target population.

9. For multi-stage designs, the base weights must reflect the probabilities of selection at each stage. For instance, in the case of a two-stage design in which the ith PSU is selected with probability p_i at the first stage, and the jth household is selected within a sampled PSU with

probability $p_{j(i)}$ at the second stage, then the overall probability of selection (p_{ij}) of each household in the sample is given by the product of these two probabilities, or

$$p_{ij} = p_i * p_{j(i)} \tag{6.2}$$

and the overall base weight of the household is obtained as before, by taking the reciprocal of its overall probability of selection. Correspondingly, if the base weight for the jth household is $w_{ij,b}$, the weight attributable to compensation for non-response is $w_{ij,nr}$, and the weight attributable to the compensation for non-coverage is $w_{ij,nc}$, then the overall weight of the household is given by:

$$w_{ij} = w_{ij,b} * w_{ij,nr} * w_{ij,nc}. \tag{6.3}$$

6.3.1 Adjustments of sample weights for unknown eligibility

10. During data collection in household surveys, there are sometimes instances when the eligibility of a household is in question. For example, the interviewer may not find anyone home at a sampled dwelling unit at the time of data collection or after repeated visits. In such a case, it is not known whether the dwelling unit is occupied or not. If it is actually occupied, then it should be classified as a non-responding dwelling unit (under the category of "not-at-home"). Otherwise, it is out of scope for the survey and therefore ineligible to be counted as a sample unit. Sometimes, interviewers assume that if no one is found in a dwelling unit during repeated visits, then that dwelling unit is unoccupied and hence ineligible. This is, in general, an incorrect assumption; one that often leads to erroneously inflated response rates.

11. When the eligibility of some sampled dwelling units is unknown, their weights must be adjusted to account for this fact. The idea is to make some assumptions that would permit the estimation of the proportion of dwelling units with unknown eligibility that are actually eligible. The simplest approach is to take the proportion of sampled dwelling units known to be either eligible or ineligible, and apply that to those of unknown eligibility. For instance, suppose that a sample of 300 dwelling units have the following response dispositions:

Table 6.1 Illustration of Response Categories in a Survey

Response Category	Number of Dwelling Units
Complete interviews	215
Eligible non-respondents	25
Ineligibles	10
Unknown eligibility	50

12. Note that the proportion of dwelling units of known eligibility that are actually eligible is (215+25)/(215+25+10) = 0.96. We can therefore assume that the same proportion (0.96) of the dwelling units with unknown eligibility can be considered eligible. In other words, 96% of the 50 dwelling units with unknown eligibility (or 48 dwelling units) are actually eligible. We then adjust the weights of the eligible dwelling units (completed interviews and eligible non-respondents) using an adjustment factor defined as follows:

$$F_{ue} = \frac{\sum_{c} w_{ij,b} + \sum_{nr} w_{ij,b} + \varepsilon \times \sum_{ue} w_{ij,b}}{\sum_{c} w_{ij,b} + \sum_{nr} w_{ij,b}}, \qquad (6.4)$$

where ε denotes the proportion of the unknown eligibility cases that are estimated to be eligible (ε=0.96 in this example). The summations over *c*, *nr*, and *ue* in the above formula denote, respectively, the sum of the base weights of dwellings with complete interviews, with eligible non-respondents, and of unknown eligibility. The adjusted base weights of dwellings with complete interviews and eligible non-respondents are then obtained by multiplying their initial base weights $w_{ij,b}$ by the factor F_{ue}.

6.3.2 Adjustments of sample weights for duplicates

13. If it is known a priori that some units have duplicates on the frame, then increased probability of selection of such units can be compensated for by assigning to them weighting factors that are reciprocals of the number of duplicate listings on the frame if such units end up in the sample. Often however, duplicates are discovered only after the sample is selected, and the probabilities of selection of such sampled units need to be adjusted to account for the duplication. This adjustment is implemented as follows: Suppose the ith sampled unit has a probability of selection, denoted by p_{i1} and suppose there are *k-1* additional records on the sampling frame that are identified by this sampled unit as duplicates, each with selection probabilities given by $p_{i2}, \ldots, p_{ik}$. Then, the adjusted probability of selection of the sampled unit in question is given by

$$p_i = 1 - (1 - p_{i1})(1 - p_{i2}) \ldots (1 - p_{ik}) \qquad (6.5)$$

The sampled unit is then weighted accordingly, that is, by $1/p_i$.

14. We now illustrate the procedures for constructing sample weights under scenarios outlined above, with specific examples.

6.4 Weighting for unequal probabilities of selection

15. For ease of exposition, let us assume a two-stage design with census enumeration areas as PSUs and households as second-stage units. Suppose an *epsem* sample of n PSUs is selected from a total of N at the first stage, and then m households are selected from each sampled PSU. The probability of selection of a household obviously will depend on the total number of households in the PSU in which it is located. Let M_i denote the number of households in PSU i. Then, the probability of selection of a PSU is n/N and the conditional probability of selection of a household in the ith sampled PSU is m/M_i. Therefore, the overall probability of selection of a household is given by:

$$p_{ij} = p_i \times p_{j(i)} = \frac{n}{N} \times \frac{m}{M_i} = \frac{nm}{N} \times \frac{1}{M_i} \tag{6.6}$$

Also, the weight of a sampled household under this design is given by:

$$w_i = \frac{1}{p_{ij}} = \frac{N}{nm} \times M_i . \tag{6.7}$$

Example 1:

An *epsem* sample of 5 households is selected from 250. One adult is selected at random in each sampled household. The monthly income (y_{ij}) and the level of education (z_{ij} = 1, if secondary or higher; 0 otherwise) of the jth sampled adult in the ith household are recorded. Let M_i denote the number of adults in household i. Then, the overall probability of selection of a sampled adult is given by:

$$p_{ij} = p_i \times p_{j(i)} = \frac{5}{250} \times \frac{1}{M_i} = \frac{1}{50} \times \frac{1}{M_i}$$

Therefore, the weight of a sampled adult is given by:

$$w_i = \frac{1}{p_{ij}} = 50 \times M_i .$$

16. We now illustrate the computation of basic estimates under the above design. Let us assume that the data obtained from the single sampled adult for each household in the first-stage sample of 5 households are as given in the table below. Note that the number of adults in each household and the corresponding overall weight of the adult sampled from each household are given in the second and third columns respectively.

Table 6.2 Illustration of Weights under Unequal Selection Probabilities

Sampled Household	M_i	w_i	y_{ij}	z_{ij}	$w_i y_{ij}$	$w_i z_{ij}$	$w_i z_{ij} y_{ij}$
1	3	150	70	1	10,500	150	10,500
2	1	50	30	0	1,500	0	0
3	3	150	90	1	13,500	150	13,500
4	5	250	50	1	12,500	250	12,500
5	4	200	60	0	12,000	0	0
TOTAL	**16**	**800**	**300**	**3**	**50,000**	**550**	**36,500**

Estimates of various characteristics can then be obtained from the above table as follows:

(i) The estimate of average monthly income is

$$\bar{y}_w = \frac{\sum w_i y_{ij}}{\sum w_i} = \frac{50{,}000}{800} = 62.5 .$$

If weights were not used, this estimate would be 60 (or 300/5).

(ii) The estimate of the proportion of people with secondary or higher education is

$$\bar{y}_w = \frac{\sum w_i z_{ij}}{\sum w_i} = \frac{550}{800} = 0.6875 \text{ or } 68.75\%.$$

If weights are not used, this estimate would be 3/5 or 0.60 or 60%.

(iii) The estimate of the total number of people with secondary or higher education is

$$\hat{t} = \sum w_i z_{ij} = 550 .$$

(iv) The estimate of the mean monthly income of adults with secondary or higher education is

$$\bar{y}_w = \frac{\sum w_i z_{ij} y_{ij}}{\sum w_i z_{ij}} = \frac{36{,}500}{550} = 66.36 .$$

Sometimes, the sampling weights are "normed", that is, the weights are multiplied by the ratio:

$$\frac{\text{number of respondents}}{\text{sum of weights of all respondents}} \qquad (6.8)$$

17. Thus the sum of the normed weights is the realized sample size for analysis (number of respondents). Note that normed weights cannot be used for estimating totals, such as total number of adults with secondary or higher education. In this case, sampled units need to be weighted by the reciprocal of their selection probabilities, that is, the regular sampling weights must be used. However, for estimating means and proportions, the weights need only be proportional to the reciprocals of the selection probabilities. In other words, it does not matter whether the regular weights or normed weights (which are proportional to the regular weights) are used to obtain estimates of averages of population parameters such as the mean number or proportion of women of childbearing age with access to primary health care. Both types of weights will yield the same result.

18. For instance, in the preceding example, the weights w_i's are proportional to M_i (w_i=50* M_i). If we use M_i as the weights, then the estimate of the proportion with secondary or higher education is

$$\hat{p} = \frac{\sum M_i z_{ij}}{\sum M_i} \frac{3\times1+1\times0+3\times1+5\times1+4\times0}{3+1+3+5+4} = \frac{11}{16} = 0.6875 \text{ or } 68.75\%,$$

exactly the same as before. However, for the estimate of the total number of adults with secondary or higher education, the regular sampling weights (w_i= 50* M_i) must be used to obtain the correct result. That is,

$$\hat{t}_s = \sum (50\times M_i) z_{ij} = 50\sum M_i z_{ij} = 50\times 11 = 550$$

Example 2:

19. A two-stage sample of households is selected in the rural areas of a country. At the first stage, 50 villages are sampled with probability proportional to their numbers of households at the last Census. The total number of households in the rural areas at the last census was 300,000. The first-stage sample selection was followed by a listing operation to compile lists of dwelling units for each of the selected villages. Sometimes, a single dwelling unit was found to consist of more than one household.

20. We now consider various sub-sampling design options (for selecting households from selected dwelling units) and specify the selection equation for the overall probability of selection of a household into the sample. Throughout, we will use the following notation: Let D_i denote the number of dwelling units in village i, and let H_{ij} denote the number of households in dwelling j of village i. Then, the total number of households in a village, denoted by H_i, is given by $H_i = \sum_j H_{ij}$. Note that $\sum_i H_i = \sum_i \sum_j H_{ij} = 300{,}000$. The selection probabilities calculated here are based on the formulas introduced in Chapter 3.

Design Option 1:

21. Fifteen dwelling units are selected by simple random sample without replacement (SRSWOR) from the list for each selected village. All households at selected dwelling units are included in the sample. Since all households at selected dwelling units are included in the sample, there are only two stages of sample selection: villages and dwelling units. Under this design, the selection equation for the overall probability of selection of a household into the sample is given by:

p_{ij} = Pr (village i selected)*Pr(dwelling unit j selected given that village i is selected)

$$p_{ij} = 50 \times \frac{H_i}{\sum_i H_i} \times \frac{15}{D_i} = \frac{750}{\sum_i M_i} \times \frac{H_i}{D_i},$$

and the base weight is given by:

$$w_{ij} = \frac{\sum_i M_i}{750} \times \frac{D_i}{H_i}.$$

22. Note that the overall probability of selection will vary from village to village depending on the ratio H_i/D_i, of the number of households to the number of dwelling units. Therefore, we conclude that this design is not self-weighting (see more about self-weighting designs in section 6.3.3 below) It would be self-weighting if every dwelling unit contained only one household, that is, the ratio H_i/D_i is equal for all sampled villages.

Design Option 2:

23. Dwellings are sampled systematically within selected villages with a sampling rate in a village inversely proportional to its number of households at the last Census. All households at selected dwellings are included in the sample. As before, there are only two stages of selection: villages and dwellings. The conditional probability of selecting a dwelling in a selected village *i* can be expressed as k/H_i, where k is the constant of proportionality. Therefore, the selection equation for the overall probability of selection of a household into the sample under this design is given by:

$$p_{ij} = 50 \times \frac{H_i}{\sum_i H_i} \times \frac{k}{H_i} = \frac{50 \times k}{\sum_i M_i},$$

and the base weight is given by:

$$w_{ij} = \frac{\sum_i M_i}{50 \times k},$$

which is a constant. Therefore, we conclude that Design Option 2 is a self-weighting design.

Design Option 3:

24. Dwellings are sampled systematically within selected villages with a sampling rate in a village inversely proportional to its number of households at the last Census. One household is selected at random in each selected dwelling. In this case, there are three stages of selection: villages, dwellings, and households. Therefore, the selection equation for the overall probability of selection of a household into the sample under this design is given by:

$$p_{ij} = 50 \times \frac{H_i}{\sum_i H_i} \times \frac{k}{H_i} \times \frac{1}{H_{ij}},$$

and the base weight is given by:

$$w_{ij} = \frac{\sum_i M_i}{50} \times \frac{H_{ij}}{k},$$

which will vary from dwelling unit to dwelling unit depending on the number of households in the dwelling unit. Therefore, we conclude that Design Option 3 is not self-weighting.

6.4.1 Case study in construction of weights: Vietnam National Health Survey 2001

25. We now proceed to illustrate the construction of the sampling weights for an actual survey, the National Health Survey (VNHS) conducted in Viet Nam in 2001. The survey was based on a stratified three-stage sample design. There were 122 strata in all, defined by urban or rural domains within 61 provinces. Sample selection was then carried out independently within each stratum. At the first stage, communes or wards were selected with probability proportional to size (number of households at the 1999 population and housing census). At the second stage, two enumeration areas (EAs) were selected in each sampled commune or ward, with systematic sampling at a sampling rate inversely proportional to the number of enumeration areas in the commune or ward. At the third and final stage, fifteen households were selected in each sampled EA again by systematic sampling.

26. The basic sample weights for sampled households under the VNHS design can be developed as follows: Let H_i and E_i denote, respectively, the number of households and the number of EAs (at the 1999 census) in commune i, and let H_{ij} denote the number of households in EA j of commune i. Then, the overall probability of selection of household k in EA j in commune i is given by:

$$p_{ijk} = n_c \times \frac{H_i}{\sum_i H_i} \times \frac{2}{E_i} \times \frac{15}{H_{ij}}$$

where n_c is the number of communes selected in a given stratum and $\sum_i H_i$ is the total number of households in the stratum. The household sampling weight (w_{ijk}) is the reciprocal of the selection probability. That is,

$$w_{ijk} = \frac{E_i \times H_{ij} \times \sum_i M_i}{30 \times n_c \times H_i}. \tag{6.9}$$

6.4.2 Self-weighting samples

27. When the weights of all sampled units are the same, the sample is referred to as *self-weighting*. Even though higher stage units are often selected with varying probabilities for reasons of sampling efficiency, such varying probabilities can be cancelled out by the probabilities of selection at subsequent stages. Design Option 2 of Example 2 above provides an example of this situation.

28. In practice, however, household survey samples are rarely self-weighting at the national level for several reasons. First, sampling units are often selected, by design, with unequal probabilities of selection. Indeed, even though the PSUs are often selected with probability proportional to size, and households selected at an appropriate rate within PSUs to yield a self-weighting design, this may be nullified by the selection of one person for interview in each sampled household. Second, the selected sample often has deficiencies including non-response (see Section 6.4) and non-coverage (see Section 6.5). Third, the need for precise estimates for domains and special subpopulations often requires over-sampling these domains to obtain sample sizes large enough to meet pre-specified precision requirements. Fourth, when the sample design entails preparing a current listing of households in the selected clusters (PSUs or SSUs) and a pre-determined fixed number of households is to be selected in each cluster, the actual probability of selection of the household is somewhat different than its design probability, the latter of which was based on frame counts rather than current counts of households; consequently, unequal probabilities of selection arise even though a self-weighting design may have been targeted.

29. In spite of the impediments outlined in the preceding paragraph, obtaining self-weighting samples should be the goal of every sample design exercise because of the advantages they offer both in the implementation of design and the analysis of the data generated by the design. With self-weighting samples, survey estimates can be derived from un-weighted data and the results then inflated, if necessary, by a constant factor to get appropriate estimates of population parameters. Furthermore, analyses based on self-weighting samples are more straightforward, and the results are more readily understood and accepted by non-statisticians and the general public.

6.5 Adjustment of sample weights for non-response

30. It is rarely the case that all desired information is obtained from all sampled units in surveys. For instance, some households may provide no data at all while other households may provide only partial data, that is, data on some but not all questions in the survey. The former type of non-response is called *unit* or *total non-response*, while the latter is called *item non-*

response. If there are any systematic differences between the respondents and non-respondents, then estimates naively based solely on the respondents will be biased. A key point of good survey practice that is emphasized throughout this handbook is that it is important to keep survey non-response as low as possible. This is necessary in order to reduce the possibility that the survey estimates could be biased in some way by failing to include (or including a disproportionately small percentage of) a particular portion of the population. For example, persons who live in urban areas and have relatively high incomes might be less likely to participate in a multi-purpose survey that includes income modules. Failure to obtain responses from a large segment of this portion of the population could affect national estimates of average household income, educational attainment, literacy, etc.

6.5.1 Reducing non-response bias in household surveys

31. The size of the non-response bias for a sample mean, for instance, is a function of two factors:

- the proportion of the population that does not respond; and
- the size of the difference in the population mean of the characteristic of interest between respondent and non-respondent groups.

32. Reducing the bias due to non-response therefore requires that either the non-response rate be small, or that there are small differences between responding and non-responding households and persons. With proper record keeping of every sampled unit that is selected for the survey it is possible to estimate, directly from the survey data, the non-response rate for the entire sample and for sub-domains of interest. Furthermore, special, carefully designed studies can be carried out to evaluate the differences between respondents and non-respondents (Groves and Couper, 1998).

33. For panel surveys (in which data are collected from the same panel of sampled units repeatedly over time) the survey designer has access to more data for studying and adjusting for the effects of potential non-response bias than in one-time or cross-sectional surveys. Here, non-response may arise from units being lost over the course of the survey or for refusing to participate in later rounds of the survey due to respondent fatigue or other reasons, and so on. Data collected on previous panel waves can then be used to learn more about differences between respondents and non-respondents and to serve as the basis for the kind of adjustments described below. More details on various techniques used for compensating for non-response in survey research are provided in Brick and Kalton (1996), Lepkowski (1988), and references cited therein.

6.5.2 Compensating for non-response

34. A number of techniques can be employed to increase response rates and hence reduce the bias associated with non-response in household surveys. One is refusal conversion through "callbacks", in which interviewers make not one, but several attempts, to complete an interview with a sampled household. Higher response rates can also be improved with better interviewer training. However, no matter how much effort is devoted to boosting response rates, non-response will always be an inevitable feature of every household survey. Consequently, survey

designers often make adjustments to compensate for non-response. Three basic approaches for making such adjustments for unit non-response are as follows:

(i) Adjusting the sample size, by drawing a larger initial sample than needed, in order to account for expected non-response;
(iii) Adjustment of the sample weights to account for non-response.

35. It is advisable that unit non-response in household surveys always be handled by adjusting the sample weights to account for non-responding households. Section 5.4.3 outlines the steps for carrying out non-response adjustment of sample weights, followed by a numerical example.

36. There are several problems associated with substitution, which is equivalent to imputation of the non-responding unit's entire record (Kalton, 1983). First, it increases the probabilities of selection for the potential substitutes, because non-sampled households close to non-responding sampled households have a higher probability of selection than those close to responding sampled households. Second, attempts to substitute for non-responding households are time-consuming, prone to errors and bias, and very difficult to check or monitor. For example, a substitution may be made using a convenient household rather than the household specifically designated to serve as the substitute or replacement for a non-responding household, thereby introducing another source of bias. Because of all these problems, substitution should not be used to compensate for non-response in household surveys, unless there is good justification for a particular application.

37. For partial or item non-response, the standard method of compensation is *imputation* which is not covered in this handbook.

6.5.3 Non-response adjustment of sample weights

38. The procedure of adjusting sample weights is frequently used to compensate for non-response in large household surveys. Essentially, the adjustment transfers the base weights of all eligible non-responding sampled units to the responding units, and is implemented in the following steps:

Step 1: Apply the initial design weights (for unequal selection probabilities and other adjustments discussed in the preceding sections, if applicable);

Step 2: Partition the sample into subgroups and compute weighted response rates for each subgroup;

Step 3: Use the reciprocal of the subgroup response rates for non-response adjustments; and

Step 4: Calculate the non-response adjusted weight for the ith sample unit as:

$$w_i = w_{1i} * w_{2i}, \tag{6.10}$$

where w_{1i} is the initial weight and w_{2i} is the non-response adjustment weight. Note that the weighted non-response rate can be defined as the ratio of the weighted number of interviews completed with eligible sampled cases to the weighted number of eligible sampled cases.

We now illustrate the ideas presented in this section with an example.

Example:
A stratified multi-stage sample of 1,000 households is selected from two regions (North and South) of a country. Households in the North are sampled at a rate of 1/100 and those in the South at a rate of 1/200. Response rates in urban areas are lower than those in rural areas. Let n_h denote the number of households sampled in stratum h, let r_h denote the number of eligible households that responded to the survey, and let t_h denote the number of responding households with access to primary health care. Then, the non-response adjusted weight for the households in stratum h is given by:

$$w_h = w_{1h} * w_{2h}, \tag{6.11}$$

where $w_{2h} = n_h / r_h$. Assume that the stratum-level data are as given in the following table:

Table 6.3. Illustration of Non-response Adjustment in Weighting

Stratum	n_h	r_h	t_h	w_{1h}	w_{2h}	w_h	$w_h\, r_h$	$w_h\, t_h$
North-Urban	100	80	70	100	1.25	125	10,000	8,750
North-Rural	300	120	100	100	2.50	250	30,000	25,000
South-Urban	200	170	150	200	1.18	236	40,120	35,400
South-Rural	400	360	180	200	1.11	222	79,920	39,960
Total	**1,000**	**730**	**500**				**160,040**	**109,110**

Therefore, the estimated proportion of households with access to primary health care is:

$$\hat{p} = \frac{\sum w_h t_h}{\sum w_h r_h} = \frac{109{,}110}{160{,}040} = 0.682 \quad \text{or } 68.2\%$$

The estimated number of households with access to primary health care is

$$\hat{t} = \sum w_h t_h = 109{,}110 = 68.2\% \text{ of } 160{,}040$$

Note that the unweighted estimated proportion of households with access to primary health care, using only the respondent data is

$$\hat{p}_{uw} = \frac{\sum t_h}{\sum r_h} = \frac{500}{730} = 0.685 \text{ or } 68.5\%,$$

and the estimated proportion using the initial weights without non-response adjustment is

$$\hat{p}_1 = \frac{\sum w_{1h} t_h}{\sum w_{1h} r_h} = \frac{83{,}000}{126{,}000} = 0.659 \text{ or } 65.9\%.$$

39. Note also this example is provided for the purpose of illustrating how initial weights are adjusted to compensate for non-response. The results show considerable disparity between the estimated proportion using only the initial weights compared to that using non-response adjusted weights, but the difference between the unweighted proportion and the non-response-adjusted proportion appears to be negligible.

40. After non-response adjustments of the weights, further adjustments can be made to the weights as appropriate. In the next section, we consider adjustment of the weights to account for non-coverage.

6.6 Adjustment of sample weights for non-coverage

41. Non-coverage refers to the failure of the sampling frame to cover all of the target population and thus some population units have no probability of selection into the sample

selected for the household survey. This is just one of many possible deficiencies of sampling frames used to select samples for surveys. Refer to chapter 4 for detailed discussion of sapling frames.

42. Non-coverage is a major concern for household surveys, especially those conducted in developing countries. Evidence of the impact of non-coverage can be seen from the fact that sample estimates of population counts based on some developing-country surveys fall well short of population estimates from other sources. Therefore, the identification, evaluation, and control of non-coverage in household surveys should be a key area for methodological work and training in national statistical offices.

43. In this section, we discuss some sources of non-coverage in household surveys and one procedure used to compensate for non-coverage, namely statistical adjustment of the weights via post-stratification.

6.6.1 Sources of non-coverage in household surveys

44. Most household surveys in developing countries are based on stratified multi-stage area probability designs. The first-stage units, or primary sampling units, are usually geographic area units. At the second stage, a list of households or dwelling units is created, from which the sample of households is selected. At the last stage, a list of house members or residents is created, from which the sample of persons is selected. Thus non-coverage may occur at three levels: the PSU level, the household level, and the person level.

45. Since PSUs are generally based on enumeration areas identified and used in a preceding population and housing census, they are expected to cover the entire geographic extent of the target population. Thus, the size of PSU non-coverage is generally small. For household surveys in developing countries, PSU non-coverage is not as serious as non-coverage at subsequent stages of the design. However, non-coverage of PSUs does occur in most surveys. For instance, a survey may be designed to provide estimates for the entire population in a country, or a region of a country, but some PSUs may be excluded on purpose at the design stage, because some regions of a country are inaccessible due to civil war or unrest, a natural disaster, or other reasons. Also, remote areas with very few households or persons are sometimes removed from the sampling frames for household surveys because they are too costly to cover, and they represent a small proportion of the population and so have very little effect on the population figures. (See Chapter 4 for more discussion including numerous examples of non-coverage of PSUs in household surveys.) In reporting results for such a survey, the exclusion of these areas must be explicitly stated. The impression should not be created that survey results apply to the entire country or region, when in fact a portion of the population is not covered. The non-coverage properties of the survey must be fully reported in the survey report.

46. Non-coverage becomes a more serious problem at the household level. Most surveys consider households to be the collection of persons who are usually related in some way, and who usually reside in a dwelling or housing unit. There are important definitional issues to resolve, such as who is a usual resident; and what is a dwelling unit? How are multi-unit

structures (such as apartment buildings) and dwelling units with multiple households handled? It may be easy to identify the dwelling unit, but complex social structures may make it difficult to identify the households within the dwelling unit. There is thus a lot of potential for misinterpretation or inconsistent interpretation of these concepts by different interviewers, or in different countries or cultures. In any event, strict operational instructions are needed to guide interviewers on whom to consider a household member or what to consider a dwelling unit.

47. Other factors that contribute to non-coverage include the inadvertent omission of dwelling units from listings prepared during field operations, or sub-populations of interest (for example, young children or the elderly), and omissions due to errors in measurement, non-inclusion of absent household members, and omissions due to misunderstanding of survey concepts. There is also a temporal dimension to the problem, that is, dwelling units may be unoccupied or under construction at the time of listing, but become occupied at the time of data collection. For household surveys in developing countries, the non-coverage problem is exacerbated by the fact that most censuses in developing countries, the unique basis for constructing sampling frames, do not provide detailed addresses of sampling units at the household and person levels. Frequently, out of date or inaccurate administrative household listings are used, and individuals within a household are deliberately or accidentally omitted from a household listing of residents. More details on sources of non-coverage are provided in Lepkowski (2003) and references cited therein.

6.6.2 Compensating for non-coverage in household surveys

48. There are several procedures for handling the problem of non-coverage in household surveys (Lepkowski, 2003). These include:

(i) improved field procedures such as the use of multiple frames and improved listing procedures; and
(ii) compensating for the non-coverage through a statistical adjustment of the weights.

49. The example below illustrates the second procedure. If reliable control totals are available for the entire population and for specified subgroups of the population, one could attempt to adjust the weights of the sample units in such a way as to make the sum of weights match the control totals within the specified subgroups. The subgroups are called *post-strata*, and the statistical adjustment procedure is called *post-stratification*. This procedure compensates for non-coverage by adjusting the weighted sampling distribution for certain variables so as to conform to a known population distribution. See Lehtonen and Pahkinen (1995) for some practical examples of how to analyze survey data with post-stratification. A simple example is provided below, just to illustrate the procedure.

Example:
In the preceding example, suppose that the number of households is known, from an independent source such as a current civil register, to be 45, 025 in the North and 115, 800 in the South. Suppose further that the weighted sample totals are respectively 40,000 and 120,040.

Step 1: Compute the post-stratification factors.

For the North region, we have: $w_{3h} = \frac{45{,}025}{40{,}000} = 1.126$; and

For the South region, we have: $w_{3h} = \frac{115{,}800}{120{,}040} = 0.965$.

Step 2: Compute final, adjusted weight: $w_f = w_h \times w_{3h}$.

The numerical results are summarized in the following table:

Table 6.4. Illustration of Post-stratified Weighting for Coverage Adjustment

Stratum	r_h	t_h	w_h	w_f	w_f*r_h	w_f*t_h
North-Urban	80	70	125	140.75	11,256	9,849
North-Rural	120	100	250	281.40	33,768	28,140
South-Urban	170	150	236	227.77	38,709	34,155
South-Rural	360	180	222	214.20	77,112	38,556
Total	**730**	**500**			**160,845**	**110,700**

Therefore, the estimated proportion of households with access to primary health care is:

$$\hat{p}_f = \frac{\sum w_f t_h}{\sum w_f r_h} = \frac{110{,}700}{160{,}845} = 0.688 \text{ or } 68.8\%.$$

50. Note that with the weights adjusted by post-stratification, the weighted sample counts for the North and South regions are respectively 45,024 (11,256+33,768) and 115,821 (38,709+77,112), which closely match the independent control totals given above.

6.7 Increase in sampling variance due to weighting

51. Even though the use of weights in the analysis of survey data tends to reduce the bias in the estimates, it could also inflate the variances of such estimates. To simplify the discussion, we consider a stratified single-stage design with equal-probability samples within strata. If the stratum variances (that is, variances among units in the strata) are not the same in every stratum, then having unequal stratum weights across strata (for instance, weights inversely proportional to the stratum variances) might produce more precise survey estimates. However, if the stratum variances are the same in every stratum, then having unequal weights will lead to higher variances in the survey estimates than having equal weights.

52. The effect of using weights is to increase the variance in an estimated population mean by the factor:

$$L = n \times \frac{\sum_h n_h w_h^2}{(\sum_h n_h w_h)^2} \tag{6.12}$$

where $n = \sum_h n_h$ is the total realized sample size, w_h is the final weight, and n_h is the realized sample size for stratum h. The above formula can also be written in terms of the coefficient of variation of the weights as:

$$L = n \times \frac{\sum_j w_j^2}{(\sum_j w_j)^2} = 1 + CV^2(w_j) \tag{6.13}$$

where $CV^2(w_j) = \frac{n}{(\sum_j w_j)^2}\left\{\sum_j w_j^2 - \frac{1}{n}(\sum_j w_j)^2\right\} = \frac{Variance\ of\ weights}{(mean\ of\ weights)^2}$.

Example:
We now illustrate the calculation of the variance inflation factor using the data in the example in section 6.4.3, with final weights w_f and realized stratum sample sizes n_h.

Table 6.5. Stratum Parameters for Variance Illustration

Stratum	r_h	w_f	$w_f r_h$	$w_j^2 r_h$
North-Urban	80	140.75	11,256	1,583,719
North-Rural	120	281.40	33,768	9,502,315
South-Urban	170	227.77	38,709	8,814,039
South-Rural	360	214.20	77,112	16,517,390
Total	**730**		**160,845**	**36,417,463**

Therefore, $L = 730 \times \frac{36{,}417{,}463}{(160{,}845)^2} = 1.03$

In other words, there is an increase in variance in the survey estimates of about 3 percent due to the use of weights.

6.8 Trimming of Weights

53. Once the weights have been calculated and adjusted to compensate for the imperfections discussed above, it is advisable to examine the distribution of the adjusted weights. Extremely large weights, even if affecting only a small portion of sampled cases, can result in a substantial increase in the variance of survey estimates. Therefore, it is a common practice to trim extreme weights to some maximum value, in order to limit the associated variation in the weights

(thereby reducing the variance of survey estimates), and at the same time prevent a small number of sampled units from dominating the overall estimate. Weight trimming is most frequently used after the adjustment of weights for non-response.

54. While the trimming of weights tends to reduce the variance of estimates, it also introduces bias in the estimators. In some instances, the reduction in variance due to the trimming of extremely large weights may more than offset the increase in the bias incurred, thereby reducing the mean-squared error of the survey estimators. In practice, weight trimming should be done only when it is justified, that is, when it can be verified that the bias introduced due to the use of trimmed (as opposed to the original) weights has less impact on the total mean-squared error than the corresponding reduction in variance achieved by trimming.

55. For any stratified design, the weight trimming process should ideally be done within each stratum. The process starts with specifying an upper bound for the original weights, and then adjusting the entire set of weights so that the sum of the trimmed weights is the same as that of the original weights. Let w_{hi} denote the final weight for the ith unit in stratum h, and let w_{hB} denote the upper bound for the weights specified for stratum h. Then, the trimmed weight for the ith sampled unit in stratum h can be defined as:

$$w_{hi(T)} = \begin{cases} w_{hi} & \text{if } w_{hi} < w_{hB} \\ w_{hB} & \text{if } w_{hi} \geq w_{hB} \end{cases} \tag{6.14}$$

56. Now, the trimmed weights for the entire sample can be further adjusted so that their sum is exactly the same as the sum of the original weights. For ease of exposition, we shall assume constant weights within strata, and drop the subscript i in the rest of this discussion. Let F_T denote the ratio of the sum of the original weights to the sum of the trimmed weights. That is,

$$F_T = \frac{\sum_h n_h w_h}{\sum_h n_h w_{h(T)}} \tag{6.15}$$

where the sums in the ratio are taken across all strata and, hence, over all units in the sample. If we define the adjusted trimmed weight for the hth stratum as

$$w^*_{h(T)} = F_T \times w_{h(T)} \tag{6.16}$$

then clearly, $\sum_h n_h w^*_{h(T)} = \sum_h n_h w_h$, as desired. We now illustrate the trimming procedure with an example. This example is designed for illustrative purposes only, in order to aid understanding of the procedure.

57. The first two columns of the table below give the total number of units and the final weight in each of 7 strata. A maximum weight of 250 is chosen and so the original weights are truncated at 250, as shown in the third column of the table.

Table 6.6. Illustration of Weight Trimming

n_h	w_h	$w_{h(T)}$	$n_h w_h$	$n_h w_{h(T)}$	$n_h w^*_{h(T)}$
80	140.75	140.75	11260	11260	11823.00
100	150.25	150.25	15025	15025	15776.25
125	175.00	175.00	21875	21875	22968.75
150	200.00	200.00	30000	30000	31500.00
120	250.00	250.00	30000	30000	31500.00
120	275.13	250.00	33015	30000	31500.00
170	285.40	250.00	48518	42500	44625.00
865			**189693**	**180660**	**189693**

Note that in this case,

$$F_T = \frac{\sum_i n_h w_{hi}}{\sum_i n_h w_{hi(T)}} = \frac{189693}{180660} = 1.05$$

The trimmed weights have been re-scaled so that they sum up to the original total $\sum_h n_h w_h = 189693$, by multiplying each weight by $F_T = 1.05$.

6.9 Concluding Remarks

58. Sample weights have now come to be regarded as an integral part of the analysis of household survey data in developing countries, as in the rest of the world. Most survey programmes now advocate the use of weights even in the rare situations involving self-weighting samples (in which case the weights would be 1). In the past, tremendous efforts were expended by survey designers for the virtually unattainable goal of achieving self-weighting samples and hence making weights unnecessary in the analysis of survey data. The conventional wisdom was that the use of weights made the analyses too complicated, and that there was very little, if any, computing infrastructure for weighted analysis. However, advances in computer technology in the past decade have invalidated this argument. Computer hardware and software are now affordable and available in many developing countries. In addition, many specialized computer software packages are now available specifically for the analysis of survey data. These are reviewed and compared in Chapter 7.

59. As discussed, the use of weights reduces biases due to imperfections in the sample related to non-coverage and non-response. Non-response and non-coverage are different types of error due to the failure of a designed survey to obtain information from some units in the target

population. For household surveys in developing countries, non-coverage is a more serious problem than non-response. The chapter provides examples of procedures for developing and statistically adjusting the basic weights to compensate for some of these problems of household surveys and for using the adjusted weights in the estimation of parameters of interest. The advent of fast-speed computers and affordable or free statistical software should make the use of weights a routine aspect of the analysis of household survey data even in developing countries. However, as the chapter has demonstrated, the development of sampling weights increases the complexity of survey operations in various ways. For example, weights need to be calculated for each stage of sample selection; they then need to be adjusted to account for various imperfections in the sample; and finally, the weights need to be stored and used appropriately in all subsequent analyses. Accordingly, careful attention must be devoted to the development of weighting operations and the actual calculation of the weights to be used in the survey analysis.

References and further reading

Brick, J.M. and Kalton, G. (1996), "Handling Missing Data in Survey Research," *Statistical Methods in Medical Research*, Vol. 5, 215-238.

Cochran, W.G. (1977), *Sampling Techniques*, 3rd edition, John Wiley & Sons, New York.

Groves, R.M., Dillman, D.A., Eltinge, J.L. and Little, R.J.A. (2002), *Survey Non-response*, John Wiley & Sons, New York.

Groves, R.M. and Couper, M.P. (1998), *Non-response in Household Interview Surveys*, John Wiley & Sons, New York.

Kalton, G. and Kasprzyk, D. (1986), "The Treatment of Missing Survey Data," *Survey Methodology*, Vol. 12, pp. 1-16.

Kalton, G. (1983), *Compensating for Missing Survey Data*, Survey Research Center, University of Michigan, Ann Arbor.

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

Kish, L. and Hess, I. (1950), "On Non-coverage of Sample Dwellings," *Journal of the American Statistical Association*, Vol. 53, pp. 509-524.

Lehtonen, R. and Pahkinen, E.J. (1995). *Practical Methods for Design and Analysis of Complex Surveys*, Wiley, New York.

Lepkowski, J. M. (2003), *Non-observation Error in Household Surveys in Developing Countries*, Technical Report on Surveys in Developing and Transition Countries, United Nations, New York.

Lessler, J. and Kalsbeek, W. (1992), *Nonsampling Error in Surveys*, John Wiley & Sons, New York.

Levy, P. S. and Lemeshow, S. (1999), *Sampling of Populations: Methods and Applications*, Third edition, John Wiley & Sons, New York.

Lohr, S. (1999), *Sampling: Design and Analysis,* Duxbury Press, Pacific Grove.

Yansaneh, I. S. (2004), *Overview of Sample Design Issues for Household Surveys in Developing Countries*, in Household Surveys in Developing and Transition Countries: Technical Report, United Nations, New York.

Chapter 7

Estimation of Sampling Errors for Survey Data

7.1 Introduction

1. The present chapter provides a brief overview of the various methods used for estimating sampling errors for household survey data generated by various sample designs, from standard designs that can be found in any introductory textbook on sampling theory [for example, Cochran (1977)], to more complex designs used for large-scale household surveys. For the standard sample designs, formulas are provided along with numerical examples to illustrate the estimation of sampling errors, the construction of confidence intervals, and the calculation of design effects and effective sample sizes. Sampling error estimation methods for more complex designs are then presented. The merits and demerits of each method are discussed and numerical examples are provided to illustrate the implementation of the procedures. An example is provided, based on data from a real survey, to illustrate the fact that standard statistical software packages underestimate the sampling errors of survey estimates, leading to wrong conclusions about the parameters of interest to the survey. To avoid this problem, the chapter strongly recommends the use of special statistical software packages that take full account of the complex nature of the designs commonly used for household surveys. Several of these software packages are described and compared.

7.1.1 Sampling error estimation for complex survey data

2. The analytical objectives of well-designed household surveys have in recent times moved beyond basic summary tables of counts or totals of parameters of interest. Analysts are now also interested in hypothesis generation and testing or model building. For instance, instead of simply estimating the proportion of a population in poverty or with secondary or higher education, analysts now want to evaluate the impact of policies, or explore the way in which a key response variable, e.g. academic performance of a school-going child, or the poverty level of a household, is affected by factors such as region, socio-economic status, gender, and age.

3. Answering these types of questions requires detailed analyses of data at the household or person level. The publication of the results of such analyses must, of necessity, include appropriate measures of the precision or accuracy of the estimates derived from the survey data. Information on the precision of survey estimates is required for proper use and interpretation of survey results and also for the evaluation and improvement of sample designs and procedures. Such monitoring and evaluation of sample designs are particularly important in the case of large national survey programmes, which are frequently designed to be the only source of detailed information on a great variety of topics.

4. One of the key measures of precision in sample surveys is the sampling variance, an indicator of the variability introduced by choosing a sample instead of enumerating the whole population, assuming that the information collected in the survey is correct. The sampling variance is a measure of the variability of the sampling distribution of an estimator. The standard error, or square root of the variance, is used to measure the sampling error. For any given survey, an estimator of this sampling error can be evaluated and used to indicate the accuracy of the estimates.

5. The form of the variance estimator, and how it is evaluated, depends on the underlying sample design. For standard designs, these estimators are often evaluated by the use of simple formulas. However, for complex sample designs used for household surveys, which often involve stratification, clustering, and unequal probability sampling, the forms of these estimators are often complex and difficult to evaluate. The calculation of sampling errors in this instance requires procedures that take into account the complexity of the sample design that generated the data, which in turn often requires the use of appropriate computer software.

6. In many developing countries, the analysis of household survey data is frequently restricted to basic tabular analysis, with estimates of means and totals, with no indication of the precision or accuracy of these estimates. Even in national statistical offices with an extensive infrastructure for statistical data collection and processing, one often finds a lack of expertise on detailed analysis of micro-level data. Some survey designers or analysts are often surprised to learn, for instance, that the clustering of elements introduces correlations among the elements that reduce the precision of the estimates relative to the simple random samples they are accustomed to analyzing; or that the use of weights in analysis generally inflates the sampling errors; or that the standard software packages they routinely use in their work do not appropriately account for these losses in precision.

7. The present chapter attempts to remedy this situation by providing a brief overview of methods of computing estimates of sampling error for the kinds of complex designs usually employed for household surveys in developing countries, as well as statistical software packages used in the analysis of such surveys. Several numerical examples are provided to illustrate the variance procedures discussed.

7.1.2 Overview of the chapter

8. Section 7.2 provides a first-principle definition of sampling variance under simple random sampling, including numerical examples illustrating the calculation of sampling variance and the construction of confidence intervals. Definitions of other measures of sampling error are provided in Section 7.3. Section 7.4 provides formulas for the calculation of sampling variance under various standard designs, such as stratified sampling and cluster sampling. Several numerical examples are introduced to facilitate understanding of the concepts. Section 7.5 discusses common features of household survey designs, as well as the contents and structure of survey data required for appropriate estimation of sampling error. The general form of the estimates of interest in household surveys is also presented. Section 7.6 provides brief guidelines on the presentation of information on sampling errors. Section 7.7 describes practical methods

of calculating sampling errors under more complex designs. These methods frequently require special procedures and the use of specialized computer software packages. The pitfalls of using standard statistical software for the analysis of survey data are discussed in Section 7.8, using an example based on data from an immunization coverage survey conducted in Burundi in 1989. Some publicly available software packages for sampling error estimation for household survey data are reviewed and compared in sections 7.9 and 7.10. The chapter ends with some concluding remarks in Section 7.11.

7.2 Sampling variance under simple random sampling

9. The sampling variance of an estimate can be defined as the average squared deviation about the average value of the estimate, where the average is taken across all possible samples. As indicated in Chapter 3, simple random sampling is the most elementary of sampling techniques, but it is rarely used in large-scale surveys because its implementation is very inefficient and prohibitively expensive.

10. To facilitate understanding of the concept of sampling variance, we consider a small population of five households (N=5), from which a small sample of two households (size n=2) is selected by simple random sampling without replacement (SRSWOR). Suppose that the variable of interest is the monthly household expenditure on food, and that the expenditures of each of the four households are as given in Table 7.1 below:

Table 7.2 Expenditure on Food per Household

Household (i)	Expenditure on Food in dollars (Y_i)
1	10
2	20
3	30
4	40
5	50

11. First, note that since we know the value of the variable of interest for all the households in our population, we can calculate the value of the parameter corresponding to the average monthly household expenditure on food:

$$\overline{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i = \frac{10+20+30+40+50}{5} = \frac{150}{5} = 30$$

The SRSWOR estimator for the average monthly expenditure on food is

$$\hat{\overline{Y}} = \frac{1}{2}\sum_{i \in S} Y_i ,$$

Where the summation is taken over the units selected into the sample. Clearly, the estimate obtained depends on the sample selected. Table 7.2 below shows all possible samples, the estimate based on each sample, the deviations of each sample estimate from the population mean, and the squared deviations. Note that $\hat{\bar{Y}}_{ave}$ denotes the average of all the sample-based estimates.

Table 7.2: Calculating the True Sampling Variance of $\hat{\bar{Y}}$

Sample *I*	**Sample units**	**Sample Estimate** ($\hat{\bar{Y}}_i$)	$\hat{\bar{Y}}_i - \hat{\bar{Y}}_{ave}$	$(\hat{\bar{Y}}_i - \hat{\bar{Y}}_{ave})^2$
1	(1, 2)	15	-15	225
2	(1, 3)	20	-10	100
3	(1, 4)	25	-5	25
4	(1, 5)	30	0	0
5	(2, 3)	25	-5	25
6	(2, 4)	30	0	0
7	(2, 5)	35	5	25
8	(3, 4)	35	5	25
9	(3, 5)	40	10	100
10	(4, 5)	45	15	225
	Average	**30**	**0**	**750**

Note that the average of the estimates based on all possible samples is:

$$\hat{\bar{Y}}_{ave} = \frac{1}{10}\sum_{i=1}^{10}\hat{\bar{Y}}_i = \frac{15+20+25+30+25+30+35+35+40+45}{10} = \frac{300}{10} = 30 = \bar{Y}$$

12. In other words, the average value of the estimate across all possible samples is equal to the population average. An estimate with such a property is known as *unbiased* for the parameter it is estimating.

13. The true sampling variance of the estimated average monthly expenditures on food from an SRSWOR of size *n*=2 from this population is

$$Var(\hat{\bar{Y}}) = \frac{1}{10}\sum_{i=1}^{10}(\hat{\bar{Y}}_i - \hat{\bar{Y}}_{ave})^2 = \frac{750}{10} = 75$$

14. The problem with the above approach is that it is not practical to select all possible samples from the population. In practice, only one sample is selected, and the population values are not known. A more practical approach is to use formulas for calculating variance. Such formulas exist for all standard sample designs.

Under simple random sampling without replacement, the sampling variance of an estimated mean ($\hat{\bar{Y}}$), based on a sample of size *n*, is given by the expression:

$$Var(\hat{\bar{Y}}) = (1 - \frac{n}{N})\frac{S^2}{n} \qquad (7.1)$$

Where $S^2 = \frac{\sum_{i=1}^{N}(Y_i - \hat{\bar{Y}})^2}{N-1}$ is a measure of the variability of the characteristic of interest (population variance of *Y*)? Usually, S^2 is unknown and must be estimated from the sample (see equation 7-2 below). It can be clearly seen from the above formula that the sampling variance depends on the following factors:

(i) the population variance of the characteristic of interest;
(ii) the size of the population;
(iii) the sample size; and
(iv) The sample design and method of estimation.

The proportion of the population that is in the sample, *n/N*, is called the sampling fraction (denoted by *f*) and the factor $[1-(n/N)]$ or *1-f*, which is the proportion of the population not included in the sample), is called the finite population correction factor (*fpc*). The *fpc* represents the adjustment made to the standard error of the estimate to account for the fact that the sample is selected without replacement from a finite population. Note, however, that when the sampling fraction is small, the *fpc* can be ignored. In practice, the *fpc* can be ignored if it does not exceed 5% (Cochran 1977).

15. The above formula indicates that the sampling variance is inversely proportional to the sample size. As the sample size increases, the sampling variance decreases, and for a census or complete enumeration (where *n*=*N*), there is no sampling variance. Note that non-response effectively decreases the sample size and so increases sampling variability.

16. It can be shown that an unbiased estimate of the sampling variance of the estimated mean is given by:

$$v(\hat{\bar{Y}}) = (1-\frac{n}{N})\frac{s^2}{n} \tag{7.2}$$

where $s^2 = \frac{\sum_{i=1}^{n}(Y_i - \hat{\bar{Y}})^2}{n-1}$ is an estimate of the population variance, S^2, based on the sample. This is referred to as the sample variance. The 95% confidence interval for the population mean is given by:

$$\hat{\bar{Y}} \pm 1.96\sqrt{v(\hat{\bar{Y}})} \tag{7.3}$$

For a population proportion, the sample based estimate and estimated variance are given by:

$$\hat{P} = \frac{\textit{number of units with characteristic}}{n} \tag{7.4}$$

$$v(\hat{P}) = (1 - \frac{n}{N}) \frac{\hat{P}(1 - \hat{P})}{n - 1} \tag{7.5}$$

Table 7.3 below summarizes the estimates of various population quantities and the variances of the estimates under simple random sampling without replacement.

Table 7.3 Estimates and their variances for selected population characteristics

Parameter	Estimate	Variance of Estimate
Population Mean ($\hat{\bar{Y}}$)	$\hat{\bar{Y}} = \frac{1}{n} \sum_{i \varepsilon Sample} Y_i$	$v(\hat{\bar{Y}}) = (1 - \frac{n}{N}) \frac{s^2}{n}$
Population Total	$\hat{T} = N\hat{\bar{Y}}$	$v(\hat{T}) = N^2 v(\hat{\bar{Y}})$
Population Proportion for a Category	$\hat{P} = \frac{number\ of\ sampled\ units\ in\ category}{n}$	$v(\hat{P}) = (1 - \frac{n}{N}) \frac{\hat{P}(1 - \hat{P})}{n - 1}$

In general, the (1-α)% confidence interval for the population mean is given by:

$$Estimate \pm z_{1-\alpha/2} \sqrt{Estimated\ Variance\ of\ Estimate} \tag{7.6}$$

where $z_{1-\alpha/2}$ is the (1-α/2)-th percentile of the standard normal distribution.

The following example illustrates the estimation of sampling variance based on a selected sample.

Example 1

Consider a simple random sample of n = 20 households drawn from a large population of N=20,000 households. The data collected are presented in Table 7.4 below, where the variable Y denotes weekly household expenditure on food, and the variable Z indicates whether or not a household possesses a TV (=1 if Yes, and 0 otherwise).

Table 7.4: Weekly household expenditure on food TV ownership for sampled households

Household (*i*)	Y_i	Z_i	i	Y_i	Z_i
1	5	0	11	7	1
2	10	1	12	8	1
3	5	0	13	9	1
4	9	1	14	10	1
5	5	1	15	8	1
6	6	1	16	8	0
7	7	0	17	5	0
8	15	1	18	7	0
9	12	1	19	12	1
10	8	0	20	4	0
			Total	**160**	**12**

The estimate of the population mean monthly household expenditure on food is

$$\hat{\bar{Y}} = \frac{1}{20}\sum_{i=1}^{20} Y_i = \frac{5+10+.....+12+4}{20} = \frac{160}{20} = 8$$

The estimated variance of the estimated mean is

$$v(\hat{\bar{Y}}) = (1-\frac{20}{20{,}000})\left\{\frac{(5-8)^2+(10-8)^2+.....+(12-8)^2+(4-8)^2}{19}\right\} = 7.87$$

The 95% confidence interval for the population mean is

$$8 \pm 1.96 \times \sqrt{7.87} = (2.50,\ 13.50)$$

The estimate of the population total monthly household expenditure on food is

$$\hat{Y} = N\hat{\bar{Y}} = 20{,}000 \times 8 = 160{,}000$$

The estimated variance of the estimated total is

$$v(\hat{Y}) = 20{,}000^2 \times 7.87 = 3{,}148{,}000{,}000$$

The 95% confidence interval for the population mean is

$$160{,}000 \pm 1.96 \times \sqrt{3{,}148{,}000{,}000} = (50030,\ 269970)$$

The estimate of the population proportion of households that possess a TV is

$$\hat{P} = \frac{1}{20}\sum_{i=1}^{20} Z_i = \frac{12}{20} = 0.6$$

The estimated variance of the estimated proportion of the households with TV is

$$v(\hat{P}) = (1 - \frac{20}{20{,}000})\frac{0.6(1-0.6)}{19} = 0.0126$$

The 95% confidence interval for the population mean is

$$0.6 \pm 1.96 \times \sqrt{0.0126} = (0.38, 0.82)$$

7.3 Other measures of sampling error

17. In addition to sampling variance, there are other measures of sampling error. These include the standard error, coefficient of variation, and design effect. These measures are algebraically related in the sense that it is possible to derive the expression of any one of the measures from the others using simple algebraic operations.

7.3.1 Standard error

18. The standard error of an estimator is the square root of its sampling variance. This measure is easier to interpret since it provides an indication of sampling error using the same scale as the estimate whereas the variance is based on squared differences.

19. A question frequently encountered in the design of surveys is how large a standard error is considered acceptable. The answer to this question depends on the magnitude of the estimate. For example, a standard error of 100 would be considered small when estimating annual income but large when estimating the average weight of people. Also, the standard error of $\sqrt{3{,}148{,}000{,}000}$ = 56,107 for the estimated total of 160,000 obtained in Example 1 above can be considered too large.

7.3.2 Coefficient of variation

20. The coefficient of variation (CV) of an estimate is the ratio of its standard error to the average value of the estimate itself. Thus the CV provides a measure of the sampling error relative to the characteristic being measured. It is usually expressed as a percentage.

21. The CV is useful in comparing the precision of survey estimates whose sizes or scales differ. However, it is not useful for estimators of characteristics whose true value can be zero or negative, including estimates of change, for example, change in average income over two years.

7.3.3 Design effect

22. The design effect (denoted by *deff*) is defined as the ratio of the sampling variance of an estimator under a given design to the sampling variance of the estimator based on a simple random sample of the same size. It can be thought of as the factor by which the variance of an estimate based on a simple random sample of the same size must be multiplied to take account of the complexities of the actual sample design due to such factors as stratification, clustering and weighting.

23. In other words, an estimator based on data from a complex sample of size n has the same variance as the estimator computed from data from a simple random sample of size *n/deff*. For this reason, the ratio *n/deff* is sometimes called the "effective sample size" for estimation based on data from a complex design. For a general discussion of "effective sample size" calculations, see Kish (1995), Potthoff et al. (1992) and references cited therein. Also, see various sections of Chapter 3 for a more detailed discussion of design effects and their use in sample design.

7.4 Calculating sampling variance for other standard designs

24. For simple designs and simple linear estimates such as means, proportions, and totals, it is usually possible to derive formulae that can be used to calculate variances of estimates. However, for the kinds of complex designs and estimates associated with household surveys, this is often difficult or impossible. In this section, we provide examples to illustrate the calculation of sampling variance for stratified and single-stage cluster sample designs. Formulas and examples of variance calculations for other standard sample designs are provided in textbooks such as Cochran (1977) and Kish (1965).

7.4.1 Stratified sampling

25. A detailed description of stratified sampling is provided in Chapter 4. In this section we shall concern ourselves only with the estimation of variance under the design. Consider a stratified design with H strata, with sample estimates of population means for the strata given by $\overline{Y}_1$, $\overline{Y}_2$, …, $\overline{Y}_H$, and sample estimates of the population variances for the strata given by S_1^2, S_2^2, …, S_H^2. Recall from Chapter 4 that an estimator of the population mean under this design is:

$$\hat{\overline{Y}}_{st} = \sum_{h=1}^{H} \hat{\overline{Y}}_h \tag{7.7}$$

where $\hat{\overline{Y}}_h$ is the sample-based estimate of $\overline{Y}_h$, $h = 1, \ldots, H$. The variance of the estimator is given by

$$v(\hat{\overline{Y}}_{st}) = \sum_{h=1}^{H} v(\hat{\overline{Y}}_h) \tag{7.8}$$

With stratified random sampling, the estimator and its estimated variance are given by

$$\hat{\bar{Y}}_{st} = \sum_{h=1}^{H} \frac{N_h}{N} \bar{y}_h = \sum_{h=1}^{H} W_h \bar{y}_h \tag{7.9}$$

where $\bar{y}_h$ is the sample mean for stratum h, N_h is the population size in stratum h, and $W_h = \frac{N_h}{N}$, h=1, …, H. The estimated variance of this estimate under stratified random sampling is given by:

$$v(\hat{\bar{Y}}_{st}) = \sum_{h=1}^{H} W_h^2 v(\bar{y}_h) = \sum_{h=1}^{H} \left(\frac{N_h}{N} \right)^2 (1 - \frac{n_h}{N_h}) \frac{s_h^2}{n_h} \tag{7.10}$$

where n_h is the sample size in stratum h, and s_h^2 is the sample variance, a sample-based estimate of S_h^2, h=1, …, H.

Example 2

We now apply these results to an example of a stratified design involving three strata with parameters as given in Table 7.5 below. Suppose we are interested in estimating the population mean, based on an overall sample of size 1,500.

Table 7.5: Example data for a Stratified Sample Design

Parameter	Population	Stratum 1 (Capital City)	Stratum 2 (Province-Urban)	Stratum 3 (Province-Urban)
Size	N =1,000,000	N_1 = 300,000	N_2 = 500,000	N_3 = 200,000
Variance	S^2=75,000	$S_1^2 = ?$	$S_2^2 = ?$	$S_3^2 = ?$
Mean	$\bar{Y}$ =?	$\bar{Y}_1$ =?	$\bar{Y}_2$ =?	$\bar{Y}_3$ =?
Cost per unit	N/A	C_1=1	C_2=4	C_3=16
Sample size under optimal allocation[22]	n=1,500	n_1=857	n_2=595	n_3=48
Sample mean	N/A	$\bar{y}_1$=4,000	$\bar{y}_2$ =2,500	$\bar{y}_3$ =1,000
Sample variance	N/A	$s_1^2 = 90{,}000$	$s_2^2 = 62{,}500$	$s_3^2 = 10{,}000$

The estimate of the population mean is

$$\hat{\bar{Y}}_{st} = \frac{300{,}000}{1{,}000{,}000} \times 4{,}000 + \frac{500{,}000}{1{,}000{,}000} \times 2{,}500 + \frac{200{,}000}{1{,}000{,}000} \times 1{,}000 = 2{,}650$$

The estimated variance of the above estimate is

$$v(\hat{\bar{Y}}_{st}) = \left(\frac{300{,}000}{1{,}000{,}000}\right)^2 \left(1 - \frac{857}{300{,}000}\right)\left(\frac{90{,}000}{857}\right) + \left(\frac{500{,}000}{1{,}000{,}000}\right)^2 \left(1 - \frac{595}{500{,}000}\right)\left(\frac{62{,}500}{595}\right) +$$
$$\left(\frac{200{,}000}{1{,}000{,}000}\right)^2 \left(1 - \frac{48}{200{,}000}\right)\left(\frac{10{,}000}{48}\right) = 43.98516$$

The 95% confidence interval for the population mean is

$$2650 \pm 1.96 \times \sqrt{43.98516} = (2637, 2663)$$

Note that the estimated variance of the estimated mean under simple random sampling is given by:

$$v(\hat{\bar{Y}}_{SRS}) = (1 - \frac{1{,}500}{1{,}000{,}000}) \times \frac{75{,}000}{1{,}500} = 49.925$$

[22] See Chapter 4

Therefore, the design effect of this stratified design is $\frac{43.98516}{49.925} = 0.88$ and the effective sample size is $\frac{1{,}500}{0.88} = 1{,}705$. This means that the estimate based on stratified random sample of size 1,500 has the same variance as that based on a simple random sample of size 1,705.

7.4.2 Single-stage cluster sampling

26. A detailed description of the cluster sampling technique is provided in Chapter 4. In this section, we provide a simple example to illustrate the calculation of sampling errors for the special case of single-stage cluster sampling.

Example 3

Suppose we are interested in estimating the proportion of school-age children in a province that have been immunized against polio. Assume, for simplicity, that there are a total of 500 equal-sized enumeration areas (EAs) in the province, each with 25 school-age children. The EAs serve as the clusters in this example. Suppose we select 10 EAs by simple random sampling without replacement out of the 500 EAs in the province, and the proportion of immunized school-aged children is measured for each sampled EA, with the results as shown in Table 7.6 below:

Table 7.6: Proportion of immunized school-aged children in 10 EAs

Sampled EA (i)	1	2	3	4	5	6	7	8	9	10
Sample proportion ($\hat{P}_i$)	$\frac{8}{25}$	$\frac{10}{25}$	$\frac{12}{25}$	$\frac{14}{25}$	$\frac{15}{25}$	$\frac{17}{25}$	$\frac{20}{25}$	$\frac{20}{25}$	$\frac{21}{25}$	$\frac{23}{25}$

For this example, the estimate of the proportion of immunized school-age children in the province is:

$$\hat{P} = \frac{160}{250} = 0.64 \quad \text{or } 64\%$$

Furthermore, the sample variance is

$$s_P^2 = \frac{1}{10-1}\sum_{i=1}^{10}(\hat{P}_i - \hat{P})^2 = 0.040533$$

Therefore, the variance of the estimated proportion is

$$v(\hat{P}) = (1 - \frac{10}{500}) \times \frac{0.040533}{10} = 0.003972$$

Note that under simple random sampling, the estimated variance of the estimated proportion is

$$v(\hat{P}_{SRS}) = (1 - \frac{250}{12{,}500}) \times \frac{0.64(1-0.64)}{250-1} = 0.0009078$$

Therefore, the design effect for this cluster sample design is $\frac{0.003972}{0.0009078} = 4.38$ and the effective sample size is $\frac{250}{4.38} = 57$. This means that the estimate based on the cluster sample of size 250 has the same variance as that based on an SRS of size 57.

7.5 Common features of household survey sample designs and data

7.5.1 Deviations of household survey designs from simple random sampling

27. As earlier stated, simple random sampling is rarely used in practice for large-scale household surveys because they are too expensive to implement. However, a thorough understanding of this design is important because it forms the theoretical basis for more complex sample designs. Most sample designs for household surveys deviate from simple random sampling for one or more of three reasons:

(i) Stratification at one or more stages of sampling;
(ii) Clustering of units in one or more stages of sampling, which reduces costs but inflates the variance of estimates due to the correlations among the units in the same cluster;
(iii) Weighting to compensate for such sample imperfections as unequal probabilities of selection, non-response, and non-coverage (see Chapter 6 for details).

28. A sample design is referred to as *complex* if it has one or more of the above features. Most household survey designs are complex and thus violate the assumptions of simple random sampling. Therefore, analyzing household survey data as if they were generated by a simple random sample design would lead to errors in the analysis and in the inferences based on such data. Furthermore, as already mentioned, the estimates of interest in most household surveys cannot be expressed as linear functions of the observations and so there may not be any closed-form formulas for the variances. The next sections address the issue of variance estimation methods for household survey designs that take into account the complexities outlined above.

7.5.2 Preparation of data files for analysis

29. Survey data collected in developing countries are sometimes not amenable to analysis beyond basic frequencies and tabulations. There are several reasons for this. First, there may be very limited or no technical documentation of the sample design for the survey. Second the data files may not have the format, structure, and the requisite information that would allow any sophisticated analysis. Third, there may be a lack of the appropriate computer software and technical expertise.

30. In order for sample survey data to be analyzed appropriately, the associated database must contain all the information reflecting the sample selection process. In particular, the database should include appropriate labels for the sample design strata, primary sampling units, secondary sampling units (SSUs), etc. Sometimes, the actual strata and PSUs used in selecting the sample for a survey need to be modified for purposes of variance estimation. Such modifications are necessary to make the actual sample design fit into one of the sample design options available in at least one of the statistical analysis software packages (see Section 7.9). The strata and PSUs created for variance estimation are sometimes called pseudo or variance strata and pseudo or variance PSUs. The relevant sample design variables, as well as the variables created for variance estimation purposes, should be included in the data file, along with corresponding documentation on how these variables are defined and used. A minimum set of three variables is required for variance estimation: the sample weight, the stratum (or pseudo-stratum) and the PSU (or pseudo-PSU). These three variables summarize the sample design, and their inclusion in the survey data set allows the appropriate analysis of the data, accounting for the complexities in the sample design.

31. Furthermore, sample weights should be developed for each sampling unit in the data file. These weights should reflect the probability of selection of each sampling unit as well as compensate for survey non-response and other deficiencies in the sample. The sample weights and the labels for the design variables are required for the appropriate estimation of the variability of the survey estimates. As mentioned in Chapter 6 and in the preceding sections of this chapter, sample weights are important not only for generating appropriate survey estimates, but also for the estimation of the sampling errors of those estimates. Therefore, it is essential that all information on weights be incorporated into the data files. In particular, whenever non-response, post-stratification, or other types of adjustments are made, the survey documentation must contain a description of these adjustment procedures.

7.5.3 Types of Survey Estimates

32. For most household surveys, the most common survey estimates of interest are in the form of totals and ratios. Assume a stratified three-stage design with PSUs at the first stage, SSUs at the second stage, and households at the third stage. The survey estimate of a total can be expressed as:

$$\hat{Y} = \sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{j=1}^{m_i} W_{hijk} Y_{hijk} \tag{7.11}$$

where

W_{hijk} = the final weight for the kth household selected in the jth SSU in the ith PSU in h^{th} stratum;

Y_{hijk} = the value for the variable Y for the kth household selected in the jth SSU in the ith PSU in h^{th} stratum.

33. At the most basic level, the weights associated with the sample units are inversely proportional to the probabilities of selection of the units into the sample. However, more sophisticated methods are often used to compute the weights to be applied in the analysis. Some of these methods are described in Chapter 6, and references cited therein.

The survey estimate of a ratio is defined as:

$$\hat{R} = \frac{\hat{Y}}{\hat{X}} \tag{7.12}$$

where $\hat{Y}$ and $\hat{X}$ are estimates of totals for variables Y and X, respectively, calculated as specified in equation (7-4) above.

34. In the case of multi-stage sampling, means and proportions are just special cases of the ratio estimator. In the case of the mean, the variable X, in the denominator of the ratio, is a count variable, defined to equal 1 for each element so that the denominator is the sum of the weights. In the case of a proportion, the variable X in the denominator is also defined to equal 1 for all elements; and the variable Y in the numerator is a binomial variable, defined to equal either 0 or 1, depending on whether or not the unit observed possesses the characteristic whose proportion is being estimated. In most household surveys, the denominator in the ratio estimator is variously defined as total population, total females, total males, total rural population, total population in a given province or district, etc.

7.6 Guidelines for presentation of information on sampling errors

7.6.1 Determining what to report

35. For large-scale national surveys with numerous variables and domains of interest and several, often competing objectives, it is not practical to present each and every estimate along with its associated sampling error. Not only will this drastically increase the volume of the publication, it is also likely to clutter the presentation of substantive results. In light of the expected variability in the sampling error estimates themselves presenting the results for too many individual variables may lead to confusion and a perception of inconsistency regarding the overall quality of the survey data collected. It is much more useful to present sampling error information for a few of the most important characteristics of interest upfront while relegating the rest to an appendix.

36. In presenting information on sampling errors, it is important to keep in mind its potential impact on the interpretation of the results of the survey and policy decisions that may derive from such interpretation. Sampling error information should always be viewed as just one, and not always the most significant component of total survey error. In some survey situations, non-sampling errors (see Chapter 8) might have a more significant impact than sampling errors on the overall quality of the survey data. For this reason, it is recommended that the information on sampling errors include a discussion of the main sources of non-sampling errors and some

qualitative assessments of thei impact on the overall quality of the survey data. Since sampling errors become more critically important at lower levels of disaggregation, it is also recommended that some cautionary remarks be included on the degree to which the survey data may be disaggregated.

37. In general, information on sampling errors should include enough detail to facilitate correct interpretation of the survey results, and to satisfy the requirements of the entire spectrum of data users, from the general data user or policy maker (whose interest is to use the survey results is to formulate policy), to the substantive analyst (engaged in further analysis and reporting of the results), to the sampling statistician (who is concerned with statistical efficiency of the design compared to other alternatives, and with features of the design that might be used in designing future surveys).

7.6.2 How to report sampling error information

Sampling errors may be presented in three different forms:

(1) As absolute values of standard errors;
(2) As relative standard errors (squared roots of relative variances; and
(3) As confidence intervals.

38. The choice among the above three forms of presentation depends on the nature of the estimate. In situations where the estimates vary in size and units of measurement, the same value of standard errors may be applicable to the estimates when expressed in relative terms and, consequently, it is more efficient to present relative standard errors. However, in general, absolute standard errors are much easier to understand and to relate to the estimate, especially in the case of percentages, proportions, and rates. Using confidence intervals requires a choice of confidence level (say 90, or 95, or 99 per cent). Since this varies according to survey objectives and the precision requirements on the estimates, it is important to specify the confidence level being used in the presentation of sampling error information, and then to retain this confidence level throughout in determining the significance of the results. The interval most frequently used in practice is the 95 per cent confidence interval, that is:

$$\text{Estimate} \pm 1.96 \times \text{Standard Error} \tag{7.13}$$

39. For more details on the subject of presentation of information on sampling errors, including specific guidelines for various categories of users and a number of illustrations, see United Nations (1993) and references cited therein.

7.6.3 Rule of thumb in reporting standard errors

40. A frequently used rule of thumb for reporting standard errors is to report the standard error to two significant digits and then to report the corresponding point estimate to the same number of decimal places as the standard error. For example:

(1) If the point estimate is 73456 with standard error of 2345, then we report the point estimate as 73500 and the standard error as 2300.

(2) If the point estimate is 1.54328 with standard error of 0.01356, then we report the point estimate as 1.543 and the standard error as 0.014.

41. The general reasoning behind this rule of thumb appears to be related to t-statistics. Two significant digits in the standard error and the corresponding number of digits in the point estimate ensures that there is not too much rounding error effect in the resulting t-statistics and, at the same time avoids the implication of an excessive level of precision given by point estimates reported to a large number of irrelevant digits. Note, however, that this rule of thumb does not necessarily work in settings where t-statistics are not of primary interest.

7.7 Methods of variance estimation for household surveys

42. In this section, we briefly describe some conventional methods for estimating variances or sampling errors for estimates based on survey data. Methods for the estimation of sampling errors for household surveys can be classified into four broad categories:

(i) Exact Methods;
(ii) The Ultimate Cluster Method;
(iii) Linearization Approximations; and
(iv) Replication Techniques.

We shall now briefly discuss each of these in turn. Interested readers can obtain more details from such references as Kish and Frankel (1974), Wolter (1985), or Lehtonen and Pahkinen (1995).

7.7.1 Exact methods

43. Sections 7.2 and 7.4 contain several examples of exact methods of variance estimation for standard sample designs. These methods constitute the best approach to variance estimation where they are applicable. However, their application to the calculation of sampling variances of estimates based on household survey data is complicated by several factors. First, sample designs used for most household surveys are more complex than simple random sampling (see Section 7.5.1 above). Second, the estimates of interest might not be in the form of simple linear functions of the observed values, and so the sampling variance can frequently not be expressed in a closed-form formula, such as that of the sample mean under simple random sampling, given in equation (7.1) or stratified sampling, given in equation (7.3). Furthermore, exact methods depend on the sample design under consideration in a particular application, the estimate of interest, and the weighting procedures used.

44. In the following sections, we discuss methods of variance estimation for sample designs usually employed for household surveys. These methods are designed to overcome the shortcomings of the exact methods.

7.7.2 Ultimate cluster method

45. The ultimate cluster method of variance estimation (Hansen, Hurwitz, and Madow, 1953, pp. 257-259) can be used to estimate the variances of survey estimates based on a sample generated by a complex sample design. Under this method, the ultimate cluster consists of the entire sample from a PSU, regardless of sampling at subsequent stages of the multi-stage design. Variance estimates are computed using only between PSU totals without having to compute variance components at each stage of selection.

46. Suppose a sample of n_h PSUs are selected from stratum h (with any number of stages within PSUs). Then the estimate of the total for stratum h is given by

$$\hat{Y}_h = \sum_{i=1}^{n_h} \hat{Y}_{hi} \tag{7.14}$$

where

$$\hat{Y}_{hi} = \sum_{j=1}^{m_i} W_{hijk} Y_{hijk}$$

Note that PSU-level estimate $\hat{Y}_{hi}$ is an estimate of $\frac{\hat{Y}_h}{n_h}$. Thus the variance of the individual PSU level estimates is given by

$$v(\hat{Y}_{hi}) = \frac{1}{n_h - 1} \sum_{i=1}^{n_h} \left(\hat{Y}_{hi} - \frac{\hat{Y}_h}{n_h} \right)^2 \tag{7.15}$$

and the variance of their total, $\hat{Y}_h$, the stratum-level total, estimated from a random sample of size n_h estimator of the population total for stratum h is given by

$$v(\hat{Y}_h) = \frac{n_h}{n_h - 1} \sum_{i=1}^{n_h} \left(\hat{Y}_{hi} - \frac{\hat{Y}_h}{n_h} \right)^2 \tag{7.16}$$

Note that simple algebraic manipulation yields the following equivalent expression for the variance estimator of the population total for stratum h

$$v(\hat{Y}_h) = \frac{n_h}{n_h - 1} \left\{ \sum_{i=1}^{n_h} \hat{Y}_{hi}^2 - \frac{\left(\sum_{i=1}^{n_h} \hat{Y}_{hi} \right)^2}{n_h} \right\} \tag{7.17}$$

Finally, with independent sampling across strata, the variance estimator for the overall population total is obtained by taking the sum of the variances of the stratum-level totals. That is,

$$v(\hat{Y}) = \sum_{h=1}^{H} v(\hat{Y}_h) \tag{7.18}$$

Note that sometimes an ad hoc finite population correction factor ($1-n_h/N_h$) is used in the above formulas.

47. The expression (7-5) is remarkable in the sense that the variance of the estimated total is a function of the appropriately weighted PSU totals $\hat{Y}_{hi}$ only, without any reference to the structure and manner of sampling within PSUs. This considerably simplifies the variance estimation formula because it does not require the computation of variance components attributable to the other stages of sampling within PSUs. This feature affords the ultimate cluster method great flexibility in handling different sample designs, and is indeed one of its major strengths and reasons for its widespread use in survey work.

Now, the variance estimator of a ratio, $\hat{R} = \frac{\hat{Y}}{\hat{X}}$, is given by

$$v(\hat{R}) = \frac{1}{\hat{X}^2}\left\{v(\hat{Y}) + \hat{R}^2 v(\hat{X}) - 2\operatorname{cov}(\hat{Y},\hat{X})\right\} \tag{7.18}$$

where $v(\hat{Y})$ and $v(\hat{X})$ are calculated according to the formula for the variance of an estimated total, and

$$\operatorname{cov}(\hat{Y},\hat{X}) = \sum_{h=1}^{H}\left\{\frac{n_h}{n_h-1}\sum_{i=1}^{n_h}\left(\hat{X}_{hi} - \frac{\hat{X}_h}{n_h}\right)\left(\hat{Y}_{hi} - \frac{\hat{Y}_h}{n_h}\right)\right\}, \tag{7.19}$$

or, equivalently,

$$\operatorname{cov}(\hat{Y},\hat{X}) = \frac{n_h}{n_h-1}\left\{\sum_{i=1}^{n_h}\hat{X}_{hi}\hat{Y}_{hi} - \frac{\left(\sum_{i=1}^{n_h}\hat{X}_{hi}\right)\left(\sum_{i=1}^{n_h}\hat{Y}_{hi}\right)}{n_h}\right\}. \tag{7.20}$$

48. Note that the above formula for the variance of a ratio can be simplified by using the fact that the relative variance of a ratio is approximately equal to the difference between the relative variances of the numerator and the denominator. Recall that the relative variance of an estimator is the ratio of its variance to its square. Thus, for an estimated ratio, $\hat{R}$, the relative variance, denoted by $relvar(\hat{R})$, is given by:

$$relvar(\hat{R}) = \frac{v(\hat{R})}{\hat{R}^2} \tag{7.21}$$

Therefore an estimate of the variance of the ratio is given by:

$$v(\hat{R}) = \hat{R}^2 relvar(\hat{R}) = \hat{R}^2 \left\{ relvar(\hat{Y}) - relvar(\hat{X}) \right\} \tag{7.22}$$

49. The ultimate cluster method of calculating sampling errors for estimated totals and ratios could be schematised in the following steps:

Step 1. For each stratum separately, compute the weighted estimate $\hat{Y}_{hi}$ for the characteristic of interest, *Y*, for each PSU (in accordance with the weighting procedures specified in Chapter 6)

Step 2. Calculate the squared value of each estimated PSU value from Step 1

Step 3: Calculate the sum of the squares of the values from Step 2 over all PSUs in the stratum

Step 4: Calculate the sum of the estimated PSU totals from Step 1 over all PSUs

Step 5: Square the result of Step 4 and divide by n_h, the number of PSUs in the stratum

Step 6: Subtract the result of Step 5 from that of Step 3, and multiply this difference by the factor $n_h/(n_h - 1)$. This is the estimated variance for the characteristic at the stratum level

Step 7: Sum the result of Step 6 over all strata to obtain the overall estimated variance for the characteristic of interest

Step 8: Calculate the square root of the result of Step 7 to obtain the estimated sampling error for the characteristic of interest

50. To calculate the estimated sampling error for ratios, such as estimated proportions, we proceed as follows:

Step 9: Calculate the relative variance of the numerator, $\hat{Y}$, by dividing the result of Step 7 by the square of the estimate of the numerator.

Step 10: Repeat Step 9 to obtain the relative variance of the denominator, $\hat{X}$.

Step 11: Subtract the result of Step 10 from that of Step 9.

Step 12: Multiply the result of Step 11 by the square of the estimated ratio ($\hat{R}$). This is the estimated variance of $\hat{R}$.

Step 13: Calculate the square root of the result of Step 12 to obtain the estimated sampling error for $\hat{R}$.

Example 4

We now consider a hypothetical example to illustrate the ultimate cluster method of variance estimation. Suppose are interested in estimating the total weekly expenditure on food for households in City A. We design a survey employing a stratified three-stage clustered design involving three strata with two PSUs selected from each stratum and two households selected from each sampled PSU. The weekly expenditure on food is then recorded for each sampled household in the survey. Table 7.7 below shows the data obtained from such a survey, including the weights (W_{hij}) and weekly expenditure on food, in dollars (Y_{hij}), for each sampled household.

Table 7.7: Weekly Household Expenditure on Food

Stratum	**PSU**	**Household**	**Weight** W_{hij}	**Expenditure** Y_{hij}	$W_{hij}*Y_{hij}$
1	1	1	1	30	30
		2	1	28	28
	2	1	3	12	36
		2	3	15	45
2	1	1	5	6	30
		2	5	7	35
	2	1	2	16	32
		2	2	18	36
3	1	1	6	7	42
		2	6	8	48
	2	1	4	13	52
		2	4	15	60
Total			**42**		**474**

51. Recall from Chapter 6 that an estimate of the total household weekly expenditure on food for the city is given by:

$$\hat{Y} = \sum_{h=1}^{3}\sum_{i=1}^{2}\sum_{j=1}^{2} W_{hij} Y_{hij} = 474.$$

Also, an estimate of the average household weekly expenditure on food is given by

$$\hat{\bar{Y}} = \frac{\sum_{h=1}^{3}\sum_{i=1}^{2}\sum_{j=1}^{2} W_{hij} Y_{hij}}{\sum_{h=1}^{3}\sum_{i=1}^{2}\sum_{j=1}^{2} W_{hij}} = \frac{474}{42} = 11 \text{ (to the nearest dollar)}$$

52. We now implement the steps of the ultimate cluster method of variance estimation in the columns of Table 7.8 below. The column numbers correspond to the steps outlined above.
Table 7.8: Steps in the ultimate cluster method of variance estimation

Stratum	PSU	Step 1	Step 2	Step 3	Step 4	Step 5	Step 6	Step 7
1	1	58	3364	9925	139	9660.5	529	
	2	81	6561	_	-		-	
2	1	65	4225	8849	133	8844.5	9	
	2	68	4624	_	-		-	
3	1	90	8100	20644	202	20402	484	
	2	112	12544	-	-		-	
Total								**1,022**

53. The stratum-level variance estimates are 529 for stratum 1, 9 for stratum 2, and 484 for stratum 3. The overall estimate of variance for the estimated total weekly household income (Step 7 in our scheme) is obtained by adding together the stratum-level estimates, which is 1,022.

7.7.3 Linearization approximations

54. Most estimates of interest in household surveys are non-linear. Some examples are the average body mass index of school-age children in a country, the proportion of income spent on housing costs in a given city, the odds ratio of a subset of the population having a characteristic compared to another subpopulation, etc. In the Linearization method, such non-linear estimates are "linearized" using a Taylor Series expansion. This involves expressing the estimate in terms of a Taylor's series expansion, and then approximating the variance of the estimate by the variance of the first-order or linear part of the Taylor series expansion using the exact methods discussed in earlier sections.

Suppose we wish to estimate the variance of an estimate z of a parameter Z and suppose z is a non-linear function of simple estimates $y_1, y_2, \ldots, y_m$ of parameters $Y_1, Y_2, \ldots, Y_m$. That is:

$$z = f(y_1, y_2, \ldots, y_m) \tag{7.23}$$

Assuming that z is close to Z, the Taylor series expansion of z to terms of the first-degree in (z-Z) is:

$$z = Z + \sum_{i=1}^{m} d_i (y_i - Y_i) \tag{7.24}$$

where the d_i's are the partial derivates of z with respect to the y_i's, that is, $d_i = \frac{\partial z}{\partial y_i}$, which are functions of the basic estimates (y_i). This means that the variance of z can be approximated by the variance of the linear function in equation (7-6) above, which we know how to calculate (from the exact methods presented in preceding sections). That is

$$v(z) = v(\sum d_i y_i) = \sum_{i=1}^{m} d_i^2 v(y_i) + \sum_{i \neq j} d_i d_j \operatorname{cov}(y_i, y_j) \tag{7.25}$$

55. Equation (7-7) involves an mxm covariance matrix of m basic estimates $y_1, y_2, \ldots, y_m$, with m variance terms and $m(m-1)/2$ identical covariance terms, which can be evaluated from the exact methods for linear statistics discussed in earlier sections.

Example 5 (Variance of a Ratio)

56. To illustrate the Linearization approach, we consider the estimation of variance for a ratio, that is,

$$z = r = \frac{y}{x} \tag{7.26}$$

Note that for this case, $\frac{\partial r}{\partial y} = \frac{1}{x}$ and $\frac{\partial r}{\partial x} = -\frac{y}{x^2} = -\frac{r}{x}$. Therefore,

$$v(r) = \frac{1}{x^2}\left\{v(y) + r^2 v(x) - 2r \operatorname{cov}(y, x)\right\}, \tag{7.27}$$

which is the familiar expression for the variance of a ratio found in most sampling textbooks.

57. Linearization is widely used in practice because it can be applied to almost all sample designs and to any statistic that can be linearized, that is, expressed as a linear function of familiar statistics such as means or totals, with coefficients coming from partial derivatives required in the Taylor series expansion. Once linearized, the variance of the nonlinear estimate can be approximated using the exact methods described above. See Cochran (1977) and Lohr (2001) for technical details about the linearization process including examples.

Advantages

58. Because the linearization method of variance estimation has been in use for a long time, its theory is well developed, and it is applicable to a wider class of sampling designs than replication methods (described below). If the partial derivatives are known, and quadratic and higher order terms in the Taylor series expansion are of negligible size, then linearization produces an approximate variance estimate for almost all linear estimators of interest, such as ratios and regression coefficients.

Limitations

59. Linearization works well only if the above assumptions about partial derivatives and higher order terms are correct. Otherwise, serious biases in the estimates may result. Also, it is generally difficult to apply the method to complex functions involving weights. A separate formula must be developed for each type of estimator, and this can require special programming. The method cannot be applied to statistics such as the median and other percentiles that are not smooth functions of population totals or means.

60. Furthermore, it is difficult to apply non-response and non-coverage adjustments with the linearization approach. The approach depends on the sample design, the estimate of interest, and the weighting procedures. It also requires that the sample design information (strata, PSUs, weights) be included on the data file.

7.7.4 Replication

61. The Replication approach refers to a class of methods that involve the taking of repeated subsamples, or *replicates*, from the data, re-computing the weighted survey estimate for each replicate, and the full sample, and then computing the variance as a function of the deviations of these replicate estimates from the full-sample estimate. The approach can be summarized in the following steps:

Step 1: Delete different subsamples from the full sample to form replicate samples;
Step 2: Produce replicate weights by repeating the estimation process for each replicate sample;
Step 3: Produce an estimate from the full sample and from each set of replicate weights; and
Step 4: Compute the variance of the estimate as of the squared deviations of the replicate estimates from the full sample estimate.

For instance, suppose k replicates are created from a sample, each with estimates $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_k$ of a parameter θ, and suppose the estimate based on the full sample is $\hat{\theta}_0$, then, the replication-based estimate of the variance is given by

$$Var(\hat{\theta}) = \frac{1}{c}\sum_{r=1}^{k}(\hat{\theta}_r - \hat{\theta}_0)^2 \tag{7.28}$$

where c is a constant, which depends on the estimation method. Replication methods defer in the value of the constant and in manner in which the replicates are formed. See the next section for a brief review of the most commonly used replication techniques.

7.7.4.1 Data file structure

62. Whatever the replication technique, the data file structure is the same and is shown in Table 7.9 below:

Table 7.9: Data File Structure for Replication

Record	Data	Full Sample Weight	Replicate Weights				
			1	2	3	….	*k*
1	Data 1	W_1	W_{11}	0	W_{13}	…	W_{1k}
2	Data 2	W_2	0	W_{22}	W_{23}	…	W_{2k}
3	Data 3	W_3	W_{31}	W_{32}	0	…	W_{3k}
	•						
	•						
	•						
N	Data *n*	W_n	W_{n1}	W_{n2}	W_{n3}	…	0

Advantages

63. The main advantage of the Replication approach relative to the Linearization approach is that replication uses the same basic estimation method regardless of the statistic being estimated (because the variance estimate is a function of the sample, not of the estimate), whereas a linearization approximation must be developed analytically for each statistic, a potentially laborious exercise in large household surveys with large numbers of characteristics of interest. Furthermore, replication techniques are convenient to use and are applicable to almost all statistics, linear or non-linear. With replication, estimates can be easily computed for subpopulations and the effects of nonresponse and other adjustments can be reflected in the replicate weights.

Limitations

64. Replication techniques are computer-intensive, mainly because they require the computation of a set of replicate weights, which are the analysis weights, re-calculated for each of the replicates selected so that each replicate appropriately represents the same population as the full sample. Also, the formation of replicates may be complicated by restrictions in the sample design (see Section 7.7.5 below), and these restrictions can sometimes lead to the overestimation of sampling errors.

65. We conclude our general comparison of linearization and replication approaches to sampling error estimation by noting that the two approaches do not produce identical estimates of sampling error. However, empirical investigations (Kish and Frankel, 1974) have shown that

for large samples and many statistics, the differences between the results produced by these two methods are negligible.

7.7.5 Some replication techniques

The most commonly used replication techniques are:

(1) Random Groups;
(2) Balanced Repeated Replication (BRR);
(3) Jackknife Replication (JK1, JK2, and JKn); and
(4) Bootstrap

We now briefly discuss each of these techniques in turn.

7.7.5.1 Random groups

66. The random group technique entails dividing the full sample into *k* groups in such a way as to preserve the survey design, that is, each group represents a miniature version of the survey, mirroring the sample design. For instance, if the full sample is an SRS of size *n*, then the random groups can be formed by randomly apportioning the *n* observations into *k* groups, each of size *n/k*. If it is a cluster sample, then the PSUs are randomly divided among the *k* groups, in such a way that each PSU retains all its observations, and so each random group is still a cluster sample. If it is a stratified multi-stage sample, then random groups can be formed by selecting a sample of PSUs from each stratum. Note that the total number of random groups to be formed in this instance cannot exceed the number of sampled PSUs in the smallest stratum.

67. The random group method can be easily used to estimate sampling errors for both linear statistics such as means and totals and smooth functions thereof, and nonlinear ones such as ratios and percentiles. No special software is necessary to estimate the sampling error, which is simple the standard deviation of the estimates based on the random groups from that based on the full sample. However, creating the random groups can be difficult in complex sample designs, since each random group must preserve the design structure of the complete survey. Furthermore, the number of random groups may be limited by the survey design itself. For instance, for a design with two PSUs per stratum, only two random groups can be formed and, in general, a small number of random groups leads to imprecise estimates of sampling error. A general rule of thumb is to have at least ten random groups in order to obtain a more stable estimate of sampling error.

7.7.5.2 Balanced repeated replication

68. Balanced repeated replication (BRR) assumes a design with two PSUs per stratum. Forming a replicate involves dividing each sampling stratum into two primary sampling units (PSUs), and selecting one of the two PSUs in each stratum, in a prescribed pattern, to represent the entire stratum. The technique can be adapted to other designs by grouping the PSUs into "pseudo-strata", each with two PSUs.

7.7.5.3 Jackknife

69. Like the BRR, the Jackknife is a generalization of the random group method that allows the replicate groups to overlap. There are three types of Jackknife: the JK1, JK2, and JKn techniques.

70. The JK1 is the typical drop-one-unit Jackknife for SRS designs. However, it can be used for other designs if the sampled units are grouped into random subsets each resembling the full sample.

71. The JK2 is similar to the BRR in the sense that it assumes a two-PSU per stratum design. In case of self-representing PSUs (see Chapter 4), pairs of secondary sampling units (SSUs) can be created. Like the BRR, the JK2 can be adapted to other designs by grouping PSUs into pseudo-strata, each with two PSUs. One PSU is then dropped at random from each stratum in turn to form the replicates.

72. The JKn is the typical drop-one-unit Jackknife for stratified designs. To create replicates, each PSU is dropped in turn from each stratum. The remaining PSUs in the stratum are re-weighted to estimate the stratum total. The number of replicates is equal to the number of PSUs (or pseudo-PSUs).

7.7.5.4 Bootstrap

73. The Bootstrap technique starts with the selection of the full sample that reproduces all the important features of the whole population. The full sample is then treated as if it were the whole population, and subsamples are then drawn from it. As before, the estimate of sampling error is obtained by taking the standard deviation of the estimates based on the resamples from that based on the full sample.

74. The Bootstrap works well for general sample designs and for non-smooth functions such as percentiles. However, it requires more computations than the other replication techniques.

We end this section by specifying in Table 7.10 below, the value of the constant c in the variance formula (7-8) that corresponds to the various replication methods.

Table 7.10: Values of the constant factor for in the Variance formula for various replication techniques

Replication Technique	**Value of constant *c* in Formula (7-6)**
Random Group	*k(k-1)*
BRR	*k*
JK1	*1*
JK2	*2*
JKn	*k/(k-1)*
Bootstrap	*k-1*

Example 6 (Jackknife)

75. We now give a simple numerical example to illustrate the Jackknife method of variance estimation. Suppose we have a sample of size 3. We can create 3 subsamples each of size 2 by dropping one unit at a time from the full sample. Table 7.11 below gives the values of a variable (*Y*). For the three subsamples, an "X" indicates which units of the sample are included in the subsample.

Table 7.11: Values for a small sample and its subsamples

		Subsample (*g*)		
Sample Unit	***Y***	**1**	**2**	**3**
1	5	X	X	
2	7	X		X
3	9		X	X
Sample Total	21			
Sample Mean	**7**	**6**	**7**	**8**

Sample variance: $s^2 = \dfrac{(5-7)^2 + (7-7)^2 + (9-7)^2}{3-1} = 4$

Estimate of variance of sample mean (ignoring *fpc*): $\dfrac{s^2}{n} = \dfrac{4}{3}$

Mean of subsample means: $\dfrac{\bar{y}_1 + \bar{y}_2 + \bar{y}_3}{3} = \dfrac{6+7+8}{3} = 7$

The Jackknife variance estimate of the sample mean is given by:

$$V_J(\bar{y}) = \frac{n-1}{n}\sum_{g=1}^{3}(\bar{y}_g - \bar{y})^2 = \frac{3-1}{3}\left[(6-7)^2 + (7-7)^2 + (8-7)^2\right] = \frac{4}{3},$$

which is exactly the same as the estimated variance of the sample mean, calculated above.

Example 7 (Formation of Replicates)

76. We continue with the data in Example 4 (Section 7.7.2) to illustrate the formation of replicate samples for various replication methods and also and the calculation of variances by the Jackknife method.

Table 7.12: **FULL SAMPLE**

Stratum	**PSU**	**Household**	**Weight** W_{hij}	**Expenditure** Y_{hij}	$W_{hij}*Y_{hij}$
1	1	1	1	30	30
		2	1	28	28
	2	1	3	12	36
		2	3	15	45
2	1	1	5	6	30
		2	5	7	35
	2	1	2	16	32
		2	2	18	36
3	1	1	6	7	42
		2	6	8	48
	2	1	4	13	52
		2	4	15	60
Total			**42**		**474**

Estimated mean based on full sample = $\hat{\bar{Y}}_0 = \dfrac{\sum_{h=1}^{3}\sum_{i=1}^{2}\sum_{j=1}^{2} W_{hij} Y_{hij}}{\sum_{h=1}^{3}\sum_{i=1}^{2}\sum_{j=1}^{2} W_{hij}} = \dfrac{474}{42} = 11$

Table 7.13: **THE JACKKNIFE (JKn) METHOD**
DROP PSU 2 FROM STRATUM 1

Stratum	**PSU**	**Household**	**Weight** W_{hij}	**Expenditure** Y_{hij}	$W_{hij}*Y_{hij}$
1	1	1	1	30	30
		2	1	28	28
2	1	1	5	6	30
		2	5	7	35
	2	1	2	16	32
		2	2	18	36
3	1	1	6	7	42
		2	6	8	48
	2	1	4	13	52
		2	4	15	60
Totals			**42**		**474**

Estimated mean based on the above replicate sample = $\hat{\bar{Y}}_1 = \frac{393}{36} = 11$

We can continue this process, dropping one PSU at a time from each stratum. A total of 6 replicate samples can be formed in this way. Table 7.14 below shows the estimates of mean weekly household income based on each of the 6 replicate samples.

Table 7.14: Replicate-based Estimates

Replicate ***j***	**PSU Deleted**	**Estimate** $\hat{\bar{Y}}_j$	$\hat{\bar{Y}}_j - \hat{\bar{Y}}_0$	$(\hat{\bar{Y}}_j - \hat{\bar{Y}}_0)^2$
1	PSU 2, Stratum1	11	0	0
2	PSU 1, Stratum1	10	-1	1
3	PSU 2, Stratum2	10	-1	1
4	PSU 1, Stratum2	12	1	1
5	PSU 2, Stratum3	12	1	1
6	PSU 1, Stratum3	13	2	4
Total				**8**

The Jackknife estimate of the variance of the estimated mean is given by:

$$Var_{JK}(\hat{\bar{Y}}) = \sum_{h=1}^{H}\left\{\frac{n_h - 1}{n_h}\sum_{j=1}^{n_h}(\hat{\bar{Y}}_j - \hat{\bar{Y}}_0)^2\right\} = \frac{1}{2}\times 8 = 4$$

(Note that for this example, H=3 and n_h=2 for all h).

77. We conclude this section by giving another example of the formation of replicate samples using the BRR method. The results shown in Table 7.15 below are for the pattern of deletion of PSUs specified in the table title.

Table 7.15: **THE BRR METHOD**
DROP PSU 2 FROM STRATA 1 AND 3; PSU 1 FROM STRATUM 2

Stratum	**PSU**	**Household**	**Weight** W_{hij}	**Expenditure** Y_{hij}	$W_{hij}*Y_{hij}$
1	1	1	1	30	30
		2	1	28	28
2	2	1	2	16	32
		2	2	18	36
3	1	1	6	7	42
		2	6	8	48
Totals			**18**		**216**

Estimated mean based on the above BRR sample = $\hat{\bar{Y}}_{1,BRR} = \frac{216}{18} = 12$.

7.8 Pitfalls of using standard statistical software packages to analyze household survey data

78. Appropriate analyses of household survey data require that sampling errors of estimates be computed in a manner that takes into account the complexity of the design that generated the data. This includes stratification, clustering, unequal-probability sampling, non-response, and other adjustments of sample weights (see Chapter 6 for details on the development and adjustment of weights). Standard statistical software packages do not account for these complexities because they typically assume that the sample elements were selected from the population by simple random sampling. As demonstrated in Chapter 6, point estimates of population parameters are impacted by the sample weights associated with each observation. These weights depend upon the selection probabilities and other survey design features such as stratification and clustering. When they ignore the sample weights, standard packages yield biased point estimates. Doing a weighted analysis with these packages reduces the bias in the point estimates somewhat, but even then, the sampling errors of point estimates are often grossly underestimated because the variance estimation procedure typically does not take into account such other design features as stratification and clustering. This means that inferences drawn from such analyses would be misleading. For instance, differences between groups might be erroneously declared significant or hypotheses might be erroneously rejected. Wrong inferences from the analyses of household data could, for instance, have significant implication for resource allocation and policy formulation at the national and regional levels.

79. We now use an example from Brogan (2004) to illustrate the fact that the use of standard statistical software packages can lead to biased point estimates, inappropriate standard errors and confidence intervals, and misleading tests of significance. The example is based on a dataset from a tetanus toxoid (TT) immunization coverage sample survey conducted in Burundi in 1989. One of the objectives of the survey was to compare seropositivity (defined as a tetanus antitoxin titer of at least 0.01 IU/ml) with history of tetanus toxoid vaccinations. For more information on this survey's methodology and its published results, see Brogan (2004) and references cited therein. Table 7.16 below presents estimates of the percentage of women who are seropositive and the associated standard error and confidence interval.

80. Note that the point estimates are the same for all programmes that use weights, but there is a clear difference between the weighted and unweighted estimates. Furthermore, standard errors produced by the appropriate software are nearly twice that produced by standard software packages that assume simple random sampling. In other words, the standard software packages seriously underestimate the variances of survey estimates. This could have important implications for policy. For instance, if some intervention was being planned based on a percentage of seropositivity of 65%, or less, then such intervention will be implemented as a result of the analysis based on the special software packages, but may not be implemented on the basis of analysis using standard software packages. Finally, Table 7.16 reveals that the software packages that appropriately estimate the variances of survey estimates produce approximately the same results. In the next section, we provide a brief overview of some publicly available statistical software packages that are used for the analysis of household survey data.

Table 7.16: Estimation of Percentage of Women Who Are Seropositive among Women with Recent Birth, Burundi, 1988-1989

Software Package	**Percent Seropositive**	**Standard error**	**95% Confidence Interval**
SAS 8.2 MEANS[1] Without weights	74.9%	2.1%	(70.8, 79.0)
SAS 8.2 MEANS[2] With weights	67.2%	2.3%	(62.7, 71.7)
SAS 8.2 SURVEYMEANS	67.2%	4.3%	(58.8, 75.6)
SUDAAN 8.0	67.2%	4.3%	(58.8, 75.6)
STATA 7.0	67.2%	4.3%	(58.8, 75.6)
EPI INFO 6.04d	67.2%	4.3%	(58.8, 75.6)
WESVAR 4.1	67.2%	4.3%	(58.8, 75.6)

7.9 Computer software for sampling error estimation

81. The above methods of sampling error estimation have been in use for a long time in developed countries, implemented mainly by customized computer algorithms developed by government statistical agencies, academic institutions, and private survey organizations. Recent advances in computer technology have led to the development of several software packages for implementing these techniques. Many of these software packages are now available for use on personal computers. The software packages use one or the other of the general approaches to variance estimation discussed in Section 7.7. Most of these software packages produce the most widely used estimates, such as means, proportions, ratios, and linear regression coefficients. Some software packages also include approximations for a wide range of estimators, such as logistic regression coefficients.

82. In this section, we present a brief overview of some publicly available software for the estimation of sampling errors for household survey data. This is by no means an exhaustive list of all available programmes and packages. We restrict attention to a few statistical software packages that are currently available on personal computers for use by the general survey data analyst. Each software package is briefly reviewed, specifying the applicable sample designs and variance estimation methods. The advantages and disadvantages of using the software are also highlighted. No attempt is made to provide the technical and computational procedures underlying the packages. Such details can be obtained from the websites of the packages, and some of the references cited at the end of the chapter.

83. The six packages reviewed here are CENVAR, EPI INFO, PC CARP, STATA, SUDAAN, and WESVAR. SUDAAN (Shah et. al., 1995), STATA (StataCorp, 1996), PC CARP (Fuller, et. al., 1989), and CENVAR all use the linearization method for estimating the sampling errors for nonlinear statistics. WESVAR uses replication methods only. Recent versions of SUDAAN also implement BRR and Jackknife techniques. Also, SAS and SPSS (which are not reviewed here) have developed new modules for the analysis of survey data. The replication-based programs offer many of the basic methods, except the bootstrap. We provide a brief comparison of the general features of these software packages. A thorough comparison of these software packages would require more extensive comparisons across sample surveys of different sizes and for more statistics, but this is beyond the scope of the limited review conducted here.

84. Internet links to the several statistical software packages reviewed in this chapter, and many others, can be found at the website:

www.fas.harvard.edu/~stats/survey-soft/survey-soft.html.

85. Brogan (2004) provides a detailed comparison of several statistical software packages, including the ones reviewed here, based on data from one household survey.

86. We now provide brief overviews of the software packages. Interested readers can obtain more details from the current manuals or the websites indicated below.

CENVAR

U.S. Bureau of the Census; contact International Programs Center,
U.S. Bureau of the Census,
Washington, DC 20233-8860;
E-mail IMPS@census.gov
Website: www.census.gov/ipc/www/imps

87. CENVAR is a component of a statistical software system designed by the U.S. Bureau of the Census for processing, management, and analysis of complex survey data, the Integrated Microcomputer Processing System (IMPS). It is applicable to most sample designs, such as simple random sampling, stratified random sampling, and multistage cluster sampling. CENVAR uses Linearization approximation for variance estimation.

88. Estimates produced by CENVAR include means, proportions, and totals for the total sample as well as specified subclasses in a tabular layout. In addition to the sampling error, the 95% confidence interval limits, coefficients of variation, design effects, and unweighted sample sizes are provided.

EPI INFO
U.S. Centers for Disease Control and Prevention
Epidemiology Program Office, Mailstop C08,
Centers for Disease Control and Prevention,
Atlanta, GA 30333
E-mail: EpiInfo@cdc1.cdc.gov
Website: http://www.cdc.gov/epiinfo/

89. EPI INFO is a statistical software system designed by the U.S. Centers for Disease Control and Prevention for processing, managing, and analyzing epidemiological data, including complex survey data (CSAMPLE component). Relevant documentation is available on-line in the program, and can be printed chapter-by-chapter. It is designed specifically for stratified multistage cluster sampling through the ultimate cluster sampling model.

90. EPI INFO produces sampling error estimates for means and proportions for the total sample as well as for subclasses specified in a two-way layout. The printed output includes unweighted frequencies, weighted proportions or means, standard errors, 95% confidence interval limits, and design effects.

PC CARP
Iowa State University
Statistical Laboratory,
219 Snedecor Hall,
Ames, IA 50011
Website: http://cssm.iastate.edu/software/pccarp.html

91. PC CARP is a statistical software package developed at Iowa State University for the estimation of standard errors for means, proportions, quantiles, ratios, differences of ratios, and

analysis of two-way contingency tables. The program is designed to handle stratified multistage cluster samples. PC CARP uses the Linearization approach for variance estimation.

STATA
Stata Corporation
702 University Drive East, College Station, TX 77840
E-mail stata@stata.com;
Website: http://www.stata.com

92. STATA is a statistical analysis software package designed for the estimation of sampling errors for means, totals, ratios, proportions, linear regression, logistic regression, and probit analysis procedures. Additional capabilities include the estimation of linear combinations of parameters and hypothesis tests, as well as the estimation of quantiles, contingency table analysis, missing data compensation, and other analyses. STATA uses the Linearization approach for variance estimation.

SUDAAN
Research Triangle Institute
Statistical Software Center,
Research Triangle Institute,
3040 Cornwallis Road,
Research Triangle Park, NC 27709-2194
e-mail SUDAAN@rti.org;
Website http://www.rti.org/patents/sudaan.html

93. SUDAAN is a statistical software package for analysis of correlated data, including complex survey data. It is applicable to a wide variety of designs, including simple random sampling and multi-stage stratified designs. It provides facilities for estimation of a range of statistics and their associate sampling errors, including means, proportions, ratios, quantiles, cross-tabulations, odds ratios; linear, logistic, and proportional hazards regression models; and contingency table analysis. The program uses the Linearization approach for variance estimation.

WESVAR
Westat, Incorporated;
1650 Research Blvd., Rockville, MD 20850-3129
E-mail WESVAR@westat.com
Website: http://www.westat.com/wesvar/

94. WESVAR is a statistical software system designed by Westat, Inc. for the analysis of complex survey data, including contingency table analysis, regression, and logistic regression. It is applicable to most sample designs but is specifically designed for stratified multistage cluster samples based on the ultimate cluster sampling model.

95. WESVAR uses replication techniques for variance estimation, including jackknife, balanced half sample, and the Fay modification to the balanced half sample method. It requires

that a new version of the data set be created in a special WESVAR format and the specification of replicate weights.

7.10 General comparison of software packages

96. The software packages reviewed here have a lot of features in common. All programs require the specification of weights, strata, and sampling units for each sample element. They do not all handle every conceivable sample design in an unbiased fashion. For example, primary sampling units in most stratified multistage sample designs are selected with probabilities proportionate to size and without replacement. Only one program in the list, SUDAAN, has features to handle explicitly this type of design. However, all listed programs will handle such a design under an ultimate cluster sample selection model (see Kalton, 1979 and Section 7.7 above). Furthermore, SUDAAN also has features to estimate variances for designs employing without replacement selection of primary sampling units. Stata is the only package with estimation features to account for the stratification and multistage selection employed in the design.

97. All of the packages estimate sampling variances and related statistics (design effects, intra-class correlation, etc.) for means, totals, and proportions for the total sample, for subclasses of the total sample, and for differences between subclasses. Most of them estimate sampling variances for regression and logistic regression statistics. All of them estimate test statistics based on the sampling variances they produce.

98. CENVAR, EPI INFO, PC CARP, and WESVAR are available either for free, or at a nominal charge. Interested users should use the e-mail addresses and other contact information provided to obtain further information on how to acquire the software and associated documentation.

7.11 Concluding remarks

99. In this chapter, we have presented a brief overview of procedures for calculating sampling errors of estimates based on both standard sample designs and more complex designs used for household surveys. The calculation of sampling errors is a critically important aspect of the analysis and reporting of results derived from household surveys. Ideally, sampling errors should be calculated for all characteristics in the tabulation package of the survey. In practice, however, a set of key characteristics of interest is designated for the calculation sampling errors for each domain. The characteristics chosen should be those regarded as substantively important to the survey, but they should also include a representative choice of items that have certain statistical properties, namely those thought to be highly clustered (for example, variables indicating ethnicity or access to services); those thought to have low clustering effects (such as marital status). In addition, the choice should be guided by other features, such as characteristics comprising a high or low proportion of the population, or of important domains of interest.

100. The chapter also advocates the use of special computer software for the estimation of sampling errors for survey data. We have provided examples of situations in which serious errors are committed in the estimation of sampling errors when standard statistical software

packages are used. In general, the use of standard statistical packages for household survey data analysis will understate the true variability of the survey estimates. These smaller estimates of standard error can lead to the drawing of misleading conclusions regarding the results of the survey, for instance erroneously declaring significant differences between the means of two groups or incorrectly rejecting a hypothesis.

101. The chapter also provides a catalogue of some publicly available statistical software packages, along with basic contact information and an overview of their application. The lack of knowledge or expertise in sampling error estimation is one of the impediments to sophisticated analysis of data in developing countries. Many analysts are not aware of the need to use specialized software or, if aware, prefer not to do so because of the need to learn a new software package.

102. It must be emphasized that the chapter represents only an introduction to the vast and growing field of variance estimation for complex survey data. The reader is encouraged to read some of the references cited at the end of the chapter for a more detailed and systematic treatment of the subject. For a more extensive review of these and other software packages, including computer code and output for some of the available software, see Brogan (2004) and references cited therein.

103. Finally, it must be recognized that with rapid advances in technology, a lot of software packages either become obsolete or are improved beyond the specifications provided in this review, in a relatively short time. Indeed it is possible that some of these specifications will be obsolete by the time this handbook is published. It is therefore important to remember that the most accurate information regarding the software packages should be obtained from their respective manuals or websites at the time of use.

References and further reading

An, A and Watts, D. (2001), New SAS procedures for analysis of sample survey data*, SUGI paper No. 23*, SAS Institute Inc., Cary, NC [http://support.sas.com/rnd/app/papers/survey.pdf]

Binder D.A. (1983), "On the Variances of Asymptotically Normal Estimators from Complex Surveys," *International Statistical Review* 51, 279-92

Brick J.M., Broene P., James P. and Severynse J. (1996), *A User's Guide to WesVarPC*, Westat, Inc., Rockville, MD

Brogan, D. (2004), *Sampling Error Estimation for Survey Data*, Technical Report on Surveys in Developing and Transition Countries, United Nations Statistics Division

Burt V.L. and S.B. Cohen (1984), "A Comparison of Alternative Variance Estimation Strategies for Complex Survey Data." *Proceedings of the American Statistical Association Survey Research Methods Section*

Carlson B.L., A.E. Johnson, and S.B. Cohen (1993), "An Evaluation of the Use of Personal Computers for Variance Estimation with Complex Survey Data," *Journal of Official Statistics* 9(4), 795-814

Dippo C.S., R.E. Fay, and D.H. Morganstein (1984), "Computing Variances from Complex Samples with Replicate Weights." *Proceedings of the American Statistical Association Survey Research Methods Section*

Hansen M.H., W.N. Hurwitz, and W.G. Madow (1953), *Sample Survey Methods and Theory, Volume I: Methods and Applications*. New York: Wiley (Section 10.16)

Hansen M.H., W.G. Madow, and B.J. Tepping (1983), "An Evaluation of Model-Dependent and Probability-Sampling Inferences in Sample Surveys," *Journal of the American Statistical Association* 78(384), 776-793

Kish L. and M.R. Frankel (1974), "Inference from Complex Samples," *Journal of the Royal Statistical Society* B(36), 1-37

Landis J.R., Lepkowski J.M., Eklund S.A., and Stehouwer S.A. (1982), A Statistical Methodology for Analyzing Data from a Complex Survey: the First National Health and Nutrition Examination Survey. *Vital and Health Statistics, 2(92),* DHEW, Washington, DC

Lehtonen, R., and E. J. Pahkinen (1995), *Practical Methods for Design and Analysis of Complex Surveys*. New York: Wiley

Lepkowski J.M., J.A. Bromberg, and J.R. Landis (1981), "A Program for the Analysis of Multivariate Categorical Data from Complex Sample Surveys." *Proceedings of the American Statistical Association Statistical Computing Section*

Levy, Paul S. and Stanley Lemeshow (1999), Sampling of Populations: Methods and Applications. Third edition. John Wiley & Sons, New York

Potthoff, R.F., Woodbury, M.A. and Manton, K.G. (1992), "Equivalent Sample Size" and "Equivalent Eegrees of Freedom" Refinements for Inference Using Survey Weights Under Superpopulation Models." *Journal of the American Statistical Association* 87, 383-396

Rust K.F. and Rao J.N.K. (1996), "Variance Estimation for Complex Surveys Using Replication Techniques", Statistical Methods in Medical Research, 5, 283-310

Rust K. (1985), "Variance Estimation for Complex Estimators in Sample Surveys," *Journal of Official Statistics* 1(4), 381-397

Shah B.V., Barnwell B.G. and Bieler G.S., (1996). *SUDAAN User's Manual: Release 7.0*, Research Triangle Institute, Research Triangle Park, NC

Tepping B.J. (1968), "Variance Estimation in Complex Surveys," *Proceedings of the American Statistical Association Social Statistics Section*, pp. 11-18

United Nations (1993), Sampling Errors in Household Surveys. NHSCP Technical Study. UNFPA/UN/INT-92-P80-15E

Wolter K.M. (1985), *Introduction to Variance Estimation*. New York: Springer-Verlag

Woodruff R.S. (1971), "A Simple Method for Approximating the Variance of a Complicated Estimate," *Journal of the American Statistical Association* 66(334), 411-414

Chapter 8

Nonsampling Errors in Household Surveys

8.1 Introduction

1. Both sampling and *nonsampling error* need to be controlled and reduced to a level at which their presence does not defeat the usefulness of the final survey results. In the preceding chapters on sample design and estimation methodology, much attention was focused on sampling error and somewhat less on other sources of variation in surveys, such as non-response, which collectively make up the class of errors known as *nonsampling error*. The latter are particularly harmful when they are non-random, because of the bias they cause in survey estimates from household surveys.

2. All survey data are subject to error from various sources. The broad fundamental distinction of errors is between errors in the measurement process and errors in the estimation of population values from measurement of a sample of it, thus, sampling error.

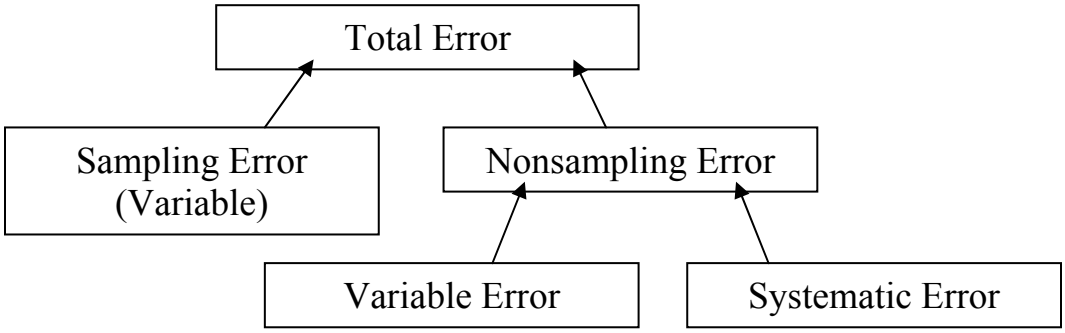
3. In the preceding chapters it has been assumed that each unit Y_i in the population is associated with a value y_i called the true value of the unit for characteristic y. It has also been assumed that whenever Y_i is in the sample the value of y reported or observed on it is y_i. In many situations this will be the case but not always. For example in countries with viable and comprehensive registration of vital events through birth certificates "true" values may be easily attainable when *yi* is defined as age. However, in other situations such as a qualitative assessment of one's health true values are much harder to obtain or even define. For example a sick person may rate himself/herself fit depending on the circumstances.

4. In survey practice the supposition that the value reported or observed on unit Y_i is always y_i irrespective of who reports it or under what circumstances it is obtained is unwarranted. Actual survey experience contains numerous examples to show that errors of measurement or observation, as well as errors from erroneous response, non-response and other reasons will occur, whenever a survey is carried out.

5. In addition to response errors, surveys are subject to errors of coverage, processing, etc. The quality of a sample estimator of a population parameter is a function of total survey error, comprising both sampling and nonsampling error. As has already been pointed out, sampling error arises solely as a result of drawing a probability sample rather than conducting a complete enumeration. Nonsampling errors, on the other hand, are mainly associated with data collection and processing procedures.

Figure 8.1 depicts the relationship of sampling and nonsampling error as components of total survey error.

Figure 8.1. Total Survey Error



6. *Nonsampling errors*, therefore, arise mainly due to invalid definitions and concepts, inaccurate sampling frames, unsatisfactory questionnaires, defective methods of data collection, tabulation and coding and incomplete coverage of sample units. These errors are unpredictable and not easily controlled. Unlike in the control of *sampling error* this error may increase with increases in sample size. If not properly controlled *nonsampling* error can be more damaging than sampling error for large-scale household surveys.

8.2 Bias and variable error

Table. 8.1. Classification of Errors

Variable errors	Sampling error
	Nonsampling error
Bias	Sampling error
	Nonsampling error

7. As Table 8.1 shows, survey errors may be classified as variable errors and bias. Variable errors come primarily from sampling error though nonsampling error also contributes to variable error, the latter chiefly from data processing operations such as coding and keying. By contrast, bias comes about mainly from nonsampling error due to such causes as invalid definitions, erroneous measurement procedures, erroneous responses, non-response, under-coverage of the target population, etc. Some bias may also be attributable to sampling error but this would be that which arises from calculations of sampling variances using a variance estimator that does not validly reflect the sample design – thus resulting in over- or under-estimates of the sampling errors.

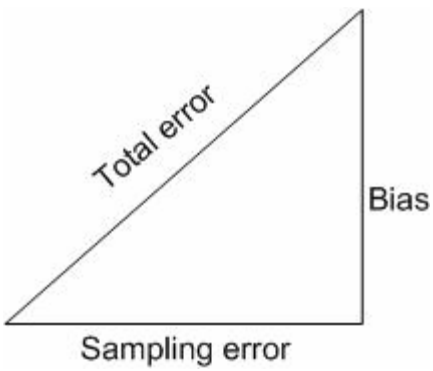
8. Bias generally refers to systematic errors that affect any survey taken under a specified sample survey design with the same constant error. As implied in the preceding paragraph, sampling errors ordinarily account for most of the variable errors of a survey, while biases arise

mainly from *nonsampling* sources. Thus, bias arises from flaws in the basic survey design and procedures, while variable error occurs because of the failure to consistently apply survey designs and procedures.

9. The statistical term for total survey error is called *mean square error* (MSE) and is equal to the variance plus the squared bias (see Figure 8.2). If for arguments sake the bias were zero, the MSE would therefore simply be the variance of the estimate. In household surveys, however, bias is never zero. As earlier indicated, however, measuring total bias in surveys is virtually impossible. That is partly because its computation requires the knowledge of the true population value which in most cases is not known or practically obtainable. The sources of bias are so numerous and complex that attempts are rarely made to estimate it in total.

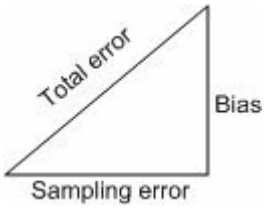
10. The triangle in Figure 8.2 below depicts the graphical representation of total error and its components. The height of the triangle represents bias while the base represents variable error. The hypotenuse measures total error, derived from the concept that root mean square error (that is, total error) equals the square root of the product, sampling variance plus squared bias.

Figure 8.2. Depiction of Mean Square Error



11. If either variable error or bias is reduced, total error is accordingly reduced. Figure 8.3 depicts a situation where both variable error and bias are considerably reduced. Consequently total error is significantly reduced as can be seen from the length of the hypotenuse compared to that in Figure 8.2.

Figure 8.3. Depiction of Reduced Mean Square Error



12. The aim of having good sample survey design coupled with a good implementation strategy is to reduce both variable error and bias in order to obtain relatively accurate sample results.

13. In general, high precision is achieved from large and otherwise well-designed samples, while accuracy can only be achieved if both variable error and bias are minimized (reduced). This implies that a precise design may nevertheless be highly inaccurate if it has a large bias. It is important to recognize in this context that estimates of standard errors that are often included in household survey reports under-estimate total survey error, because those estimates do not account for the impact of bias.

14. In practice nonsampling errors can be further decomposed into the variable component and systematic errors, according to Biemer and Lyberg (2003). Systematic errors are generally non-compensating errors and therefore tend to agree (mostly in the same direction, that is, positive), while variable errors are compensating errors that tend to disagree (canceling each other).

8.2.1 Variable component

15. The variable component of an error arises from chance (random) factors affecting different samples and repetition of the survey. In the case of the measurement process we can imagine that the whole range of procedures from interviewer selection, data collection to data processing can be repeated using the same specified procedures, under the same given conditions, and independently without one repetition affecting another. The results of repetitions are affected by random factors, as well as systematic factors, which arise from conditions under which repetitions are undertaken and affect the results of the repetition the same way.

16. When the variable errors (VE) are caused only by sampling errors, VE squared equals sampling variance. The deviation of the average survey value from the true population value is the bias. Both variable errors and biases can arise either from sampling or nonsampling operations. The variable error will measure the divergence of the estimator from its expected value and it comprises both sampling variance and nonsampling variance. The difference of the expected value of the estimator from its true value is total bias and comprises both sampling bias and nonsampling bias.

17. Variable errors can be assessed on the basis of appropriately designed comparisons between repetitions (replications) of survey operations under the same conditions. Reduction in variable errors depends on doing more of something such as increasing the sample size or using

more interviewers. On the other hand bias can be reduced only by improving survey procedures such as introducing quality control measures at various stages of the survey operation.

8.2.2 Systematic error (bias)

18. Systematic error occurs when, for example, there is a tendency either to consistently under-report or over-report in a survey. For example, in some societies where there are no certificates for birth registration, there is a tendency among men to report their ages older than their actual ages. This practice would obviously result in systematic bias – an over-estimate of the average age in the male population.

8.2.3 Sampling bias

19. Sampling biases may arise from inadequate or faulty conduct of the specified probability sample or from faulty methods of estimation. The former includes defects in frames, wrong selection procedures and partial or incomplete enumeration of selected units. Chapters 3 and 4 of this handbook provide detailed discussion of the numerous circumstances under which sampling bias can occur from inadequate implementation of even a near-perfect sample design.

8.2.4 Further comparison of bias and variable error

20. In general, biases are difficult to measure, which is why we emphasize their rigorous control. Their assessment can only be done by comparing the survey results with external reliable data sources. On the other hand variable error can be assessed through comparisons between sub-divisions of the sample or repetition of the survey under the same conditions. Bias can be reduced by improving survey procedures.

21. According to Verma (1991) some sources of error appear mainly in the form of bias, among them coverage, non-response, and sample selection. On the other hand errors in coding and data entry may appear largely as variable error.

22. Although both systematic and variable errors reduce overall accuracy and reliability, bias is more damaging in estimates such as population means, proportions and totals. These linear estimates are sums of observations in the sample. As already noted, variable nonsampling errors like sampling errors can be reduced by increasing the sample size. For nonlinear estimates such as correlation coefficients, standard errors and regression estimates both variable and systematic error can lead to serious bias (Biemer and Lyberg, 2003). However, in many cases in household surveys the main aim is to provide descriptive measures of the population such as means, population totals and proportions; therefore the emphasis here is on reducing systematic error.

23. In summary, bias arises from shortcomings in the basic survey design and procedures. It is harder to measure than variable error and can only be assessed on the basis of comparison with more reliable sources outside the normal survey or with information obtained by using improved procedures.

8.3 Sources of nonsampling error

24. The various and numerous causes of nonsampling error are present right from the initial stage when the survey is being planned and designed to the final stage when data are processed and analyzed.

25. A household survey programme may be considered as a set of rigorous rules that specify various operations. These rules, for instance, describe the target population to be covered, specify subject-matter concepts and definitions to be used in the questionnaire and lay out methods of data collection and measurements to be made. If the survey operations are carried out according to the rules laid down, it is theoretically possible to obtain a *true value* of the characteristic under study. However, nonsampling error makes this an unattainable ideal.

26. In general, nonsampling errors may arise from one or more of the following factors:

a. Data specification being inadequate and/or inconsistent with respect to objectives of the survey
b. Duplication or omission of units due to imprecise definition of the boundaries of area sampling units
c. Incomplete or wrong identification particulars of sampling units[23] or faulty methods of interviewing
d. Inappropriate methods of interview, observation or measurement using ambiguous questionnaires, definitions or instructions
e. Lack of trained and experienced field interviewers including lack of good quality field supervision
f. Inadequate scrutiny of the basic data to correct obvious mistakes
g. Errors in data processing operations such as coding, keying, verification, tabulation etc.
h. Errors during presentation and publication of tabulated results.

The list above is by no means exhaustive.

8.4 Components of nonsampling error

27. Biemer and Lyberg (2003) identify five components of nonsampling error, namely specification, frame, non-response, measurement and processing error. We may add that estimation error is another that should be considered. These types or error are briefly discussed below:

8.4.1 Specification error

28. Specification error occurs when the concept implied by the question is different from the underlying construct that should be measured. A simple question such as how many children

[23] Note that even though this kind of error occurs in the sample selection operation, it is nevertheless a type of nonsampling bias.

does a person have can be subject to different interpretations in some cultures. In households with extended families the respondent's biological children may not be distinguished from children of brothers or sisters living in the same household. In a disability survey, a general question asking people whether or not they have a disability can be subject to different interpretations depending on the severity of the impairment or the respondent's perception of disability. People with minor disabilities may perceive themselves to have no disability. Unless the right screening and filter questions are included in the questionnaire, the answers may not fully bring out the total number of people with disabilities.

8.4.2 Coverage or frame error

29. In most area surveys primary sampling units comprise clusters of geographic units such as census enumeration areas (*EA*s) (see chapter 4 for a full discussion on frames). It is not uncommon for the demarcation of *EA*s to be improperly carried out during census mapping. Thus households may be omitted or duplicated in the second stage frame. Such imperfections can bias the survey estimates in two. If units are not present in the frame but should have been, this results in zero probability of selection for the omitted units and an under-estimate will result. On the other hand if some units are duplicated, this results in over-coverage and hence an over-estimate.

30. Errors associated with the frame may, therefore, result in both *over-coverage* and *under-coverage*. The latter is the most common in large-scale surveys in most African countries.

31. In multi-stage household surveys, sampling involves a number of stages, such as selection of area units in one or more stages; listing and selection of households; and listing and selection of persons within selected households (see chapter 3). Coverage error can arise in any of these stages.

32. It is important to re-emphasize that neither the magnitude nor the effect of coverage errors is easy to estimate because it requires information not only external to the sample but also, by definition, external to the sampling frame used.

33. *Non-coverage* denotes, as implied above, failure to include some units of a defined survey population in the sampling frame (see chapter 6 for more discussion on coverage error including non-coverage). Because such units have zero probability of selection, they are effectively excluded from the survey results.

34. It is important to note that we are not referring here to deliberate and explicit exclusion of sections of a larger population from survey population. Survey objectives and practical difficulties determine such deliberate exclusions. For example attitudinal surveys on marriage may exclude persons under the minimum legal age for marriage. Residents of institutions are often excluded because of practical survey difficulties. Areas in a country infested with landmines may be excluded from a household survey to safeguard the safety of field workers. When computing non-coverage rates, members of the group deliberately and explicitly excluded should not be counted either in the survey target population or under non-coverage. In this

regard defining the survey population should be part of the clearly stated essential survey conditions (see chapter 3 for discussion of the survey target population).

35. The term *gross-coverage error* refers to the sum of the absolute values of *non-coverag*e and *over-coverage error ra*tes. The *net non-coverage* refers to the excess of non-coverage over over-coverage. It is, therefore, their algebraic sum. The net coverage measures the gross-coverage only if over-coverage is absent. Most household surveys in developing countries suffer mainly from under-coverage errors. Most survey research practitioners agree that in most social surveys under-coverage is a much more common problem than over-coverage. Corrections and weighting for non-coverage are much more difficult than for non-responses, because coverage rates cannot be obtained from the sample itself, but only from outside sources.

36. The non-coverage errors may be caused by the use of faulty frames of sampling units, as is discussed at length in chapter 4. If the frames are not up-dated and old frames are used in order to save time or money, it may lead to serious bias. For example, in a household survey if an old list of housing units is not up-dated from the time of its original preparation (which could be as long as 10 years prior to the current survey) then newly added housing units in the selected enumeration area will not be part of the second-stage frame of housing units. Similarly, abandoned housing units will remain in the frame as blanks. In such a situation, there may be both omission of units belonging to the population and inclusion of units not belonging to the population.

37. At times there is also failure to locate or visit some units in the sample. This is a problem with area sampling units, especially, in which the interviewer must identify and list the households to be sampled. This problem arises also from use of incomplete lists. In addition, weather or poor transportation facilitates may sometimes make it impossible to reach certain units during the designated period of the survey.

38. As discussed in chapter 3, the underlying objective of a household sample survey is to obtain objective results to facilitate making valid inferences about the desired target population from the observation of units in the sample. Survey results can, therefore, be distorted if the extent of non-coverage differs among geographical regions and sub-groups such as male-female, age categories, ethnic and socio-economic classes.

39. Non-coverage errors differ from non-response. The latter, as discussed previously, results from failure to obtain observations on some sample units due to refusals, failure to locate addresses or find respondents at home, loss of questionnaires, etc. The extent of non-response is measured from the sample results by comparing the selected sample with the realized sample. As noted above the extent of non-coverage, by contrast, can only be estimated by some kind of check external to the survey operation.

8.4.2.1 Sample implementation errors

40. Implementation error in sampling refers to losses and distortions within the sampling frame, for example, erroneous application of the selection rates or procedures. Another example

is the inappropriate substitution in the field of the selected households by others that are more convenient or cooperative.

8.4.2.2 Reducing coverage error

41. The most effective way to reduce coverage error is to improve the frame by excluding erroneous units and duplicates. This is best accomplished by ensuring that old frames are up-dated adequately (see chapter 4 for detailed discussion). It is also important to ensure that the area sampling units and the households within them can be easily located, best accomplished by having good mapping operations in place when the original frame, usually the latest population and housing census, is constructed.

8.4.3 Non-response

42. As noted repeatedly in this handbook, non-response refers to the failure to obtain responses from some of the sample units. It is instructive to think of the sample population as split into two strata, one consisting of all sample units for which responses are obtained and the second for which no responses could be obtained.

43. In most cases non-response is not evenly spread across the sample units but is heavily concentrated among subgroups. As a result of differential non-response, the distribution of the achieved sample across the subgroups will deviate from that of the selected sample. This deviation is likely to give rise to non-response bias if the survey variables are also related to the subgroups.

44. The *non-response* rate can be accurately estimated if counts are kept of all eligible elements that fall into the sample. The r*esponse rate* for a survey is defined as the ratio of the number of questionnaires completed for sample units to the total number of eligible[24] sample units (see also chapter 6). Reporting of non-response in all public releases of the survey data is recommended practice and it should be mandatory in official surveys. Non-response can be due to selected sample persons not being at home, refusing to participate in the survey or being incapacitated to answer questions. Non-response can also occur due to lost schedules/ questionnaires and to inability to conduct the survey in certain areas because of weather, terrain or lack of security. All categories of non-response refer to eligible respondents and should exclude ineligibles, as implied by the footnote below. For example, in a fertility survey, the target population in the selected *EA*s will comprise only women in reproductive age groups, thus excluding females outside the age group and all males.

45. As mentioned in chapter 6, there are two types of non-response: *unit non-response* and *item non-response*. *Unit non-response* implies that no information is obtained from a given sample units, while *item non-response* refers to a situation where some but not all the information is collected for the unit. Item non-response is evidenced by gaps in the data records for responding sample units. Reasons may be due to refusals, omissions by interviewers and incapacity. Refusal by a prospective respondent to take part in a survey may be influenced by

[24] Some units that are sampled may be found to be out-of-scope for the survey and hence ineligible, such as vacant, condemned or abandoned dwellings

many factors, among them, lack of motivation, shortage of time, sensitivities of the study to certain questions, etc. Groves and Couper (1995) suggest a number of causes of refusals, which include social context of the study, characteristics of the respondent, survey design (including respondent burden), interviewer characteristics and the interaction between interviewer and respondent. With specific reference to item non-response, questions in the survey may be perceived by the respondent as being embarrassing, sensitive or/and irrelevant to the stated objective. The interviewer may skip a question or ignore recording an answer. In addition, a response may be rejected during editing.

46. The magnitude of unit (total) non-response, among other reasons, is indicative of the general receptivity, complexity, organization and management of the survey. Thus it is indicative of the complexity, clarity and acceptability of particular items sought in a questionnaire and the quality of the interviewer work in handling those items.

47. Non-response introduces bias in the survey results, which can be serious in situations in which the non-responding units are not "representative" of those that responded, and that is usually the case. Non-response increases both the sampling error, by decreasing the sample size, and nonsampling errors.

48. Efforts to increase response are often procedural, with respect to the choice of survey operations. For example, in order to increase response in the 1978 Fertility Survey in Zambia, female teachers were recruited as interviewers to ask questions on contraception etc. It was thought that if young men were used as interviewers there would have been a higher rate of refusals as it is taboo for young men to ask especially elderly women questions about sex-related matters including contraception.

49. Non-response cannot be completely eliminated in practice; however, it can be minimized by persuasion techniques, through repeated visits to "not-at-home" households and other methods. See chapters 6 and 9 for more information on treatment of item non-response in survey data.

8.4.4 Measurement error

50. These errors arise from the fact that what is observed or measured departs from the actual values of sample units. These errors center on the sustentative content of the survey such as definition of survey objectives, their transformation into usable questions, and obtaining, recording, coding and processing responses. These errors thus pertain to the accuracy of measurement at the level of individual units.

51. For example at the initial stage wrong or misleading definitions and concepts on frame construction and questionnaire design lead to incomplete coverage and varied interpretations by different interviewers leading to inaccuracies in the collected data.

52. Inadequate instructions to field staff are another source of error. For some surveys instructions are vague and unclear leaving interviewers to use their own judgment in carrying out fieldwork. The interviewers themselves can be a source of error. At times the information collected on a given item for all units may be wrong; this is mainly due to inadequate training of field workers.

53. Age reporting in Africa is a common measurement problem through age heaping and digital preference. These and other examples of measurement error may be attributable to respondents or interviewers or both. At times there may be interaction between the two, which may contribute to inflating such errors. Likewise, the measurement device or technique may be defective and may cause observational errors.

Respondents can introduce errors because of the following reasons:

- Failure to understand the survey question(s)
- Careless and incorrect answers due to, for example, lack of adequate understanding of the objective(s) of the survey; in particular, the respondent may not give sufficient time to think over the questions.
- Desire to "cooperate" by answering questions even when they do not know the correct answer
- Deliberate inclination to give wrong answers, for example, in surveys dealing with sensitive issues, such as income and stigmatized diseases.
- Memory lapses if there is a long reference period, a case in point being the collection of information of non-durable commodities in expenditure surveys.

54. The cumulative effect of various errors from different sources may be considerable since errors from different sources may not cancel. The net effect of such errors can be a large bias.

8.4.5 Processing errors

Processing errors comprise:

- Editing errors
- Coding errors
- Data entry errors
- Programming errors, etc.

55. The above errors arise during the data processing stage. For example in coding open-ended answers related to economic characteristics, coders may deviate from the prescribed procedures in coding manuals, and therefore assign wrong codes to occupations.

8.4.6 Errors of estimation

56. Errors in estimation are chiefly due to failure to apply correct formulas in calculating the survey weights. Errors may also arise by calculating the weights erroneously even when the correct formula is used. Estimates of sampling variance (sampling error) arise when the variance estimator used is not faithful to the actual sample design, thus creating errors in the confidence

intervals associated with the survey point estimates. In each such instance mentioned in this paragraph the results are biased.

8.5 Assessing nonsampling error

57. The sources of nonsampling error are numerous and varied, as has been discussed at length in this chapter. Accordingly, it is virtually impossible to assess the totality of nonsampling error that arises in a survey. It is possible, however, to study and assess some of the components of nonsampling error, as discussed in the sub-sections below.

8.5.1 Consistency checks

58. In designing the survey instruments (questionnaires), special care has to be taken to include certain auxiliary items of information that will serve as a check on the quality of the data to be collected. If these additional items of information are easy to obtain, they may be canvassed for all units covered in the survey, otherwise, they may be canvassed only for a sub-sample of units.

59. For example, in a post census enumeration survey (*PES*), where the *de jure* method is followed it may be helpful to also collect information on *de facto* basis, so that it will be possible to work out the number of persons temporarily present and the number of persons temporarily absent. A comparison of these two figures will give an idea of the quality of data. Similarly, inclusion of items leading to certain relatively stable ratios such as sex ratios may be useful in assessing the quality of survey data.

60. Consistency checks should also be used in the processing stage of the survey. Cross-checks can be introduced to ensure, for example, that persons coded as head of households are not younger than a pre-specified age or that females with fertility histories are not younger than, say, 13 years old. See much more discussion on consistency checks in the data processing stage in chapter 9.

8.5.2 Sample check/verification

61. One way of assessing and controlling some types of nonsampling errors in surveys is to independently duplicate the work at the different stages of operation with a view to facilitating the detection and rectification of errors. For practical reasons the duplicate checking can only be carried out on a sample of the work by using a smaller group of well-trained and experienced staff. If the sample is properly designed and if the checking operation is efficiently carried out, it is possible, not only to detect the presence of errors, but also to get an idea of their magnitude. If it were possible to completely check the survey work, the quality of the final results could be considerably improved.

62. With the sample check, rectification work can only be carried out on the sample checked. This limitation can be improved somewhat by dividing the output at different stages of the survey, that is, filled-in schedules, coded schedules, computation sheets, etc., into lots and checking samples from each lot. In this case, when the error rate in a particular lot is more than

the specified level, the whole lot may be checked and corrected for errors, thereby improving the quality of the final results.

8.5.3 Post-survey or re-interview checks

63. An important sample check, which may be used to assess response errors consists of selecting a sub-sample, or a sample in the case of a census, and re-enumerating it by using better trained and more experienced staff than those employed for the main investigation. For this approach to be effective, it is necessary to ensure that;

- The re-enumeration is taken up immediately after the main survey to minimize recall error.
- Steps are taken to minimize the *conditioning effect that* the main survey may have on the work of the post-survey check.

64. Usually the check-survey is designed to facilitate the assessment of both *coverage* and *content errors*. For this purpose, it is first desirable to re-enumerate all the units in the sample at the high stages, that is, *EA*s and villages, with the view of detecting coverage errors and then to re-survey only a sample of ultimate units ensuring proper representation for different parts of the population which have special significance from the point of view of nonsampling errors.

65. A special advantage of the check-survey is that it facilitates a unitary check, which consists first, of matching the data obtained in the two enumerations for the units covered by the check-sample and then analyzing the observed individual differences. When discrepancies are found, efforts are made to identify the cause of their presence and gain insight into the nature and types of nonsampling errors.

66. If a unitary check cannot be mounted due to time and financial constraints, an alternative but less effective procedure called aggregate check, may be used. This method consists in comparing estimates of parameters given by check-survey data with those from the main survey. The aggregate check gives only an idea of net error, which is the resultant of positive and negative errors. The unitary check, by contrast, provides information on both net and gross error.

67. In a post-survey check, the same concepts and definitions as those used in the original survey should be followed.

8.5.4 Quality control techniques

68. There is ample scope for applying statistical quality control techniques to survey work because of the large scale and repetitive nature of the operations involved in such work. Control charts and acceptance-sampling techniques can be used in assessing the quality of data and improving the accuracy of the final results in large-scale surveys. To illustrate, work of each data entry clerk could be checked 100 percent for an initial period of time, but if the error rate falls below a specified level, only a sample of the work thereafter may be verified.

8.5.5 Study of recall errors

69. Response errors, as earlier mentioned in this chapter, arise due to various factors such as:

- The attitude of the respondent towards the survey.
- Method of interview.
- Skill of the interviewer.
- Recall error.

70. Of these, *recall error* needs particular attention as it presents special problems often beyond the control of the respondent. It depends on the length of reporting period and on the interval between the reporting period and the date of the survey. The latter may be taken care of by choosing for the reporting period a suitable interval preceding the date of survey or as near that period as possible.

71. One way of studying recall error is to collect and analyze data relating to more than one reporting period in a sample or sub-sample of units covered in a survey. The main problem with this approach is the effect of certain amount of *conditioning effect* possibly due to the data reported for one reporting period influencing those reported for the other period. To avoid the conditioning effect, data for the different periods under consideration may be collected from different sample units. Note that large samples are necessary for this comparison.

72. Another approach is to collect some additional information, which will permit estimates for different reporting periods to be obtained. For example in a demographic survey one may collect not only age of respondent, but also date, month and year of birth. The discrepancy will reveal any recall error that may be present in the reported age.

8.5.6 Interpenetrating sub-sampling

73. This method involves drawing from the overall sample two or more sub-samples, which should be selected in an identical manner and each capable of providing a valid, estimate of the population parameter. This technique helps in providing an appraisal of the quality of the information, as the interpenetrating sub-samples can be used to secure information on nonsampling errors such as differences arising from differential interviewer bias, different methods of eliciting information, etc.

74. After the sub-samples have been surveyed by different groups of interviewers and processed by different teams of workers at the tabulation stage, a comparison of the estimates based on sub-samples provides a broad check on the quality of the survey results. For example, in comparing the estimates based on four sub-samples surveyed and processed by different groups of survey personnel, if three estimates are close to each other and the other estimate differs widely and beyond that which could reasonably be attributed to sampling error, then one would suspect the quality of work in the discrepant sub-sample.

8.6 Concluding remarks

75. Nonsampling errors should be given due attention in household sample surveys because they can cause huge biases in the survey results if not controlled. In most surveys very little attention is given to the control of such errors at the expense of producing results that may be unreliable. The best way to control nonsampling errors is to follow the right procedures of all survey activities from planning, sample selection up to the analysis of results. In particular, careful and intensive training of field personnel should be standard practice and survey questions, especially those which have not been validated by past survey efforts, should be fully pre-tested.

References and further reading

Biemer, P. and Lyberg, L. (2003), *Introduction to Survey Quality*: Wiley Series in Survey Methodology, Wiley, Hoboken.

Biemer, P. et al (editors) (1991), *Measurement Errors in Surveys:* Wiley Series in Probability and Mathematical Statistics, Wiley, New York.

Groves, R. and Couper, M. (1995), "Theoretical Motivation for Post-Survey Non-response Adjustment in Household Surveys," *Journal of Official Statistics*. Vol.11, No.1, pp. 93-106.

Groves, R. et al (editors) (2000), *Survey Non-response*, Wiley-Interscience Publication, John Wiley & Sons, Inc., New York.

Hansen M., Hurwitz, W. and Bershad, M. (2003), "Measurement Errors in Censuses and Surveys:" *Landmark Papers in Survey Statistics*, IASS Jubilee Commemorative Volume.

Kalton, G. and Heeringa, S. (2003), *Leslie Kish: Selected Papers*, Wiley Series in Survey Methodology, Hoboken.

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

Murthy, M. (1967), *Sampling Theory and Methods*, Statistical Publishing Society, Calcutta.

Onsembe, J. (2003), "Improving Data Quality in the 2000 Round of Population and Housing Censuses," UNFPA Country Services team, Addis Ababa, Ethiopia.

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

Raj, D. and Chandhok, P. (1998) *Sample Survey Theory*, Narosa Publishing House, London.

Som, R. (1996), *Practical Sampling Techniques*, Marcel Dekker Inc., New York.

Cochran, W. (1963), *Sampling Techniques*, Wiley, New York.

Shyam, U. (2004), "Turkmenistan Living Standards Survey 2003," Technical Report, National Institute of State Statistics and Information, Ashgabad, Turkmenistan.

Sukhatme, P. et al (1984), *Sampling Theory of Surveys with Applications*, Iowa State University Press and Indian Society of Agricultural Statistics, Ames and New Delhi.

United Nations Statistics Division (1982), *Nonsampling Errors in Household Surveys: Sources, Assessment and Control: National Household Survey Capability Programme*, United Nations, New York.

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

Whitfold, D and Banda, J. (2001), "Post Enumeration Surveys: Are they Worth it?" United Nations Symposium on Global Review of 2000 Round of Population and Housing Censuses: Mid –Decade Assessment and Future Prospects, New York, 7-10 August.

Chapter 9

Data Processing for Household Surveys

.

9.1 Introduction

1. This chapter discusses data processing with respect to national household surveys. It starts with a description of the typical household survey cycle. It then discusses the preparations for data processing as an integral part of the survey planning process

2. Information technology has developed rapidly during the last two decades or so. Its development has, in turn, impacted significantly on the techniques for designing and implementing statistical survey processing systems.

3. The main development in hardware has been the shift from mainframe systems to personal computer platforms. The personal computer has become increasingly powerful both in terms of processing speed and storage capacity. Personal computers can now perform different kinds of processes with respect to statistical operations, ranging from small-scale surveys to large-scale statistical operations such as population censuses and household surveys with very large samples.

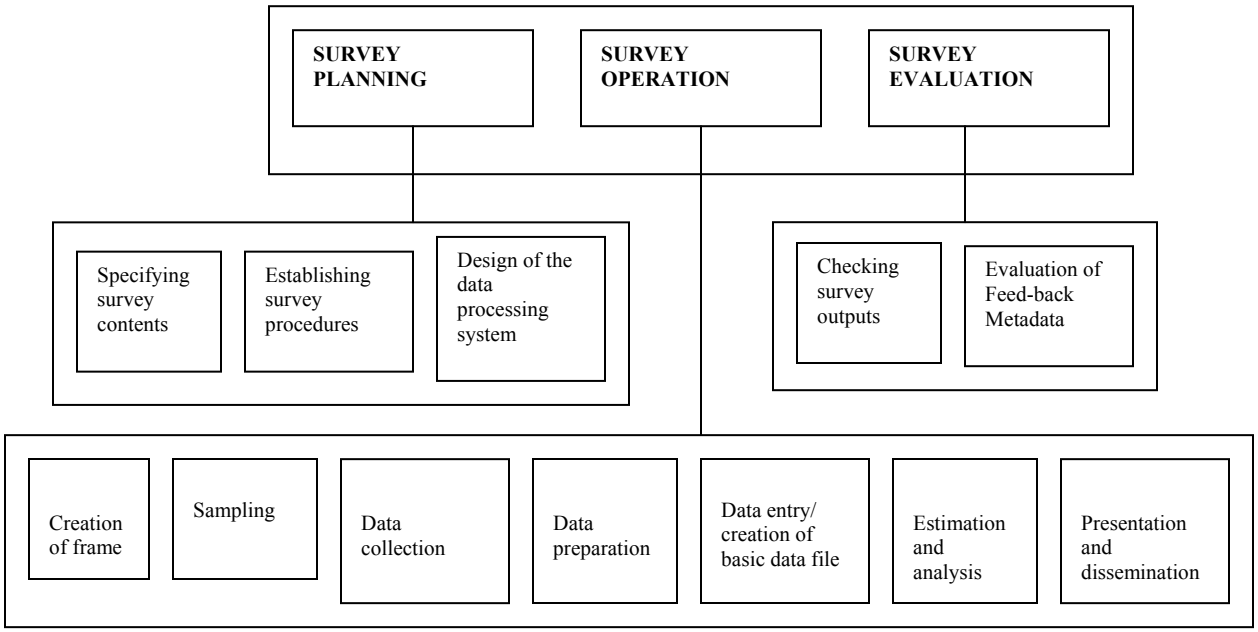
4. Parallel to the developments in hardware have been the significant improvements in the quality of software for statistical data processing, analysis, and dissemination. This has also made it possible for some of the processing tasks to move from computer experts to subject matter specialists.

5. A number of software packages for the processing of statistical surveys have emerged over the years. The relative strengths for each of these software products differ with the different steps of data processing. The annex to this chapter may serve as a rule of thumb for choosing software for the different steps of survey data processing. The annex also provides a summarized description of each of the software products cited in the chapter.

9.2 The household survey cycle

Figure 9.1 depicts the typical household survey cycle.

Figure 9.1. Household Survey Cycle



Source: Sundgren (1991).

6. In principle all surveys run through the same kind of cycle, with the typical phases as follows:

Survey planning: The designers of the survey make decisions about the major purposes and uses of the survey results, its major outputs and major inputs, procedures for obtaining the inputs (the design and preparation of the questionnaire and related survey instruments) and transforming them into outputs, and the design of the data processing and documentation system.

Survey operations: This consists of the creation of the sampling frame, sample design and selection, data collection (measurement), data preparation (data entry, coding, editing and imputation), creation of the observation (the basic raw data) file, estimation including the computation of weights, creation of derived variables, analysis, and presentation and dissemination of results.

Survey evaluation: This consists of checking and evaluating whether the specified end-products have been delivered, the output properly published and advertised, the metadata documented and stored, etc.

7. Before embarking on the design and implementation of the data processing system for a particular survey, it is important to visualize the overall system for the particular survey. The proper sequencing of operations and processes is also critical for the successful implementation of any household survey. The desired sequencing is that survey objectives should determine the output design (for example, the tabulation plan and databases). That in turn would dictate the

subsequent activities of survey design, data collection, data preparation and processing, and, ultimately, the analysis and dissemination of the results.

8. Data processing can be viewed as the process through which survey responses obtained during data collection are transformed into a form that is suitable for tabulation and analysis. It entails a mixture of automated and manual activities. It can also be time-consuming and resource-intensive and has an impact on the quality and cost of the final output.

9.3 Survey planning and the data processing system

9.3.1 Survey objectives and content

9. As discussed in chapter 2, the first step in the design of any survey should be the articulation and documentation of its main objectives. Household surveys provide information about households in the population. They are implemented to answer questions that the stakeholders may have about the target population. The objectives of a particular survey can be seen as an attempt to obtain answers to such questions and the respective survey questionnaire should, therefore, provide the relevant data.

10. Typically, the questions that stakeholders may need to be addressed through the use of household survey data can be classified into a number of categories as follows (Glewwe, 2003):

11. One type of questions is that which seeks to establish the fundamental characteristics of the population under study (*the proportion of the population that is poor; the rate of unemployment; etc.*).

12. Another set is one that seeks to assess the impact of interventions on or general developments in household characteristics (for example, *the proportion of households participating in a particular program, or how their characteristics compare to those of households not participating in the program, whether the living conditions of households are improving or deteriorating over time, etc)*.

13. Finally, there is the category of questions about determinants or relationships between household circumstances and characteristics *(that is, questions on what is happening and why it is happening)*.

9.3.2 Survey procedures and instruments

9.3.2.1 Tabulation plans and expected outputs

14. A useful technique to assist the survey designer in bringing precision to the user's need for information is to produce tabulation plans and dummy tables. Dummy tables are draft tabulations, which include everything except the actual data. As a minimum the tabulation

outline should specify the table titles, column stubs, identifying the substantive variables to be tabulated, the background variables to be used for classification, and the population groups (survey objects or elements or units**)** to which the various tables apply. It is also desirable to show the categories of classification in as much detail as possible, though these may be adjusted later when the sample distribution over the response categories is better known.

15. The importance of a tabulation plan can be viewed from a number of perspectives. One is that the production of dummy tables will indicate if data to be collected will yield usable tabulations. They will not only point out what is missing, but also reveal what is superfluous. Furthermore, the extra time that is spent on producing dummy tables is usually more than compensated for at data tabulation stages by reducing time spent on the design and production of actual tables.

16. There is also the close relationship between the tabulation plan and the sampling design employed for a survey. For example, geographical breakdown in the tables is only possible if the sample is designed to permit such breakdown.

17. The United Nations (1982) Manual on **National Household Survey Capability Programme - *Survey Data Processing: a Review of Issues and Procedures*,** provides a more exhaustive description of what a tabulation plan entails and its various benefits.

18. The aim of the above is to stress the important role that a tabulation plan can play with respect to the effective planning of the particular survey and the respective data processing system. It is important, however, to stress that a tabulation plan only represents the skeleton of some of the output that can be expected from the respective survey. Household surveys have the potential of generating a wealth of information. The cleaned micro dataset of the survey can be seen as the main and basic output. Such a micro dataset often needs to be packaged and made available to stakeholders in a user-acceptable form through appropriate distribution channels.

9.3.2.2 Form design and printing

19. Once the survey objectives and tabulation plan have been determined, the relevant questionnaire(s) can be developed. The questionnaire plays a central role in the survey process in which information is transferred from those who have it (the respondents) to those who need it (the users). It is the instrument through which the information needs of the users are expressed in operational terms as well as the main basis of input for the data processing system for the particular survey.

20. According to Lundell (2003), the physical layout of the questionnaire to be used during enumeration has implications on the data capture and vice-versa. If scanning techniques have been found advantageous for data capture, special form designs are required and they will be different depending upon whether key-to-disk or manual data entry techniques are decided upon.

21. Regardless of the data entry technique, every questionnaire must be uniquely identifiable. There should be a unique form identification printed on every form. Since misinterpretation of

the form identification may cause duplicate entries and other problems measures should be taken to minimize the risk. Use of bar codes will obviously be the best choice when using scanning techniques. If manual data entry is considered, the identification of the form should still contain information such as a check-digit to prevent incorrect entry. The identification code should uniquely identify each questionnaire and should always be numerical. Typically, information for assignment of sample weights or expansion factors (strata, primary sampling units, area segments, distinction between administrative areas needed for tabulation, etc.) is also attached to the form identification.

22. The forms are usually bound in books (for example, book for each Enumeration Area or Ward, etc.). Every book must be uniquely identifiable as the case with the forms, and there must be a clearly specified relationship between each book and its forms, so that you can always find form X belongs to book Y, and only book Y. Book identification information will be used throughout the data processing phase. This will start from indicating the arrival in the storeroom of a book from the field, to retrieving a form when necessary, for example, to check something during tabulation or analysis of the data. Therefore, the risk of misinterpreting the identification information of a particular book must be minimized in the same way as the identification information of a particular form.

23. All fields must be designed to accommodate the maximum number of characters possible for its variable, for example, the maximum possible number of household members must be made absolutely clear in order to size the field correctly.

24. It is important to ensure that there are no flaws with respect to the definition of observation units, skip patterns and other aspects of the questionnaire. Every household survey collects information about a major statistical unit (the basic object) – the household – as well as a variety of subordinate units (associated objects) within the household – persons, budget items, agricultural plots and crops, etc. The questionnaire should be clear and explicit about just what these units are, and it should also ensure that each individual unit observed is properly tagged with a unique identifier. A typical way to identify households, which is also an important feature for the manual data entry system, is by means of a simple serial number written or stamped on the cover page of the questionnaire or pre-printed by the print shop. Usually, the serial number also represents the form identification.

25. Given the increasing adoption of image scanning and processing as a rapid means of data capture some discussion of the special features relating to the design of questionnaires for image scanning is also important. When questionnaires are designed to be processed by an image scanner, for example, by Optical Character Recognition (OCR), Intelligent Character Recognition (ICR) or Optical Mark Reading (OMR) software, some issues of questionnaire design that arise include those discussed in the next paragraph.

26. Two bar codes are used on the questionnaire. The first is the code that identifies each page of the questionnaire. This is important particularly where pages are very similar in layout and format. It is the main means by which the imaging software distinguishes the different pages of the questionnaire. A second bar code, usually with the associate interpretation, is placed on every page of the questionnaire and is exactly the same on every page of the questionnaire, but

will differ sequentially from every subsequent questionnaire. This bar code ties the pages of a particular questionnaire together since preparation for scanning usually separates the pages of the questionnaire. As a result, this bar code is very important.

27. The exact layout of the fields on the printed form represents the data dictionary format of the data to be collected from the questionnaire. If the intention is to capture a five-digit enumeration district code, for example, a constrained print field with a five-digit partitioned box is designed on the printed questionnaire to capture the enumeration district code which must be entered into this box. If the capture screen were designed for manual data entry this precision in the printing of the questionnaire would not be necessary and an open box capable of accepting five digits would be all that is required.

28. Designing forms for data entry by scanning or keypunching also differs considerably because the scanner relies entirely on the positions of the data fields for identification. Contrary to manual entry it needs no printed field identifications on the questionnaire apart from some adjustment fields on each page. To create a high accuracy level for the scanning system's interpretation of the image the fields should be neither too small nor too big. The interviewers should be encouraged in training as well as during the management of the fieldwork to write distinct, clear characters centered in the data field.

29. There is also the fundamental difference in the way the questionnaire is designed to capture, for example, occupation codes between manual coding processes suitable for manual keying data entry as opposed to on-screen coding using a computer-aided lookup table containing the occupation code book on-screen. When the questionnaire is designed for keying of data, the layout of the code is printed on the questionnaire for use by the person responsible for coding of the questionnaire after reviewing the open-ended response to the question, *"What is your occupation?*". In the case of a questionnaire to be scanned no such allocation of space on the face of the printed questionnaire is absolutely necessary, because the code book is built into the printing template designed on the computer and the role of the coder and data entry clerk is collapsed into one verifier role. At the time of data entry of the occupation code the verifier coder operator is presented with the open-ended response from the scanned image and a drop down of the indexed occupation codes is presented for rapid selection.

30. When use of scanning techniques is considered, the quality of printing of the questionnaire also becomes an important issue. The scanners are more sensitive to imperfections during printing than the human eye. Problems may occur, for example, because of use of certain colours or combination of some colours, variations in hue and sharpness, skewed or misplaced prints, errors in automatic numbering of pages and binding errors.

31. Catherine (2003) and Lundell (2003) present some of the key issues in relation to use of scanning technology for the processing of statistical surveys and censuses.

9.3.3 Design for household surveys data processing systems

9.3.3.1 Generalized approach to household surveys data processing

32. Systems design is one of the major activities entailed in the planning of a household survey. Essentially, during this step the survey data to be collected and the whole data processing system are specified according to some formalized scheme.

33. According to Jambwa et al (1989) the benefits that accrue from the adoption and use of a formalized scheme for the design, development, and documentation of all systems within a statistical agency, particularly for household surveys include the following:

(i) It can be the common platform for co-operation required between the statisticians, subject matter experts and systems analysts/programmers.

(ii) All survey operations would be explicitly described and documented and could be referenced at a latter stage. The resulting documentation (that is, set of metadata) is important both for the development and maintenance of the respective statistical production systems.

(iii) The costs for systems development and maintenance tend to be quite high in a statistical agency, given the many different systems that are usually required. The structure of household surveys tends to follow the same pattern and principles. For example, they tend to share the same file and data structures, coding systems, etc. Subsequent surveys can therefore benefit from data processing systems developed for previous ones, and this can be expected to bring down their development and maintenance costs.

(iv) The adoption of a formalized approach is also important for survey integration if, for example, there is a wish to conduct some combined analysis of data from different surveys or different survey rounds.

9.3.3.2 General features of a formalized scheme for systems design

34. This sub-section provides an indication of some of the general and fundamental aspects for the formalized scheme cited above.

Data Structure

35. Decisions about the social issue to be analyzed, the data to be used, and the statistical technique to be applied are fundamental to good analysis. However, an even more basic question is the identification and definition of the objects or units of analysis of the survey. This issue has already been emphasized in the sub-section above on Form Design and Printing. During the design of the data processing system for the survey, a more formalized and detailed description of the unit (objects) of analysis and variables should be undertaken.

36. According to Sundgren (1986) the object or unit of analysis can be defined as any concrete or abstract entity (physical object, living creature, organization, event, etc.) that the users may want to have information about. The definition of an object is intimately tied to the social issue (survey objectives) about which data are collected and analysed. Similarly, for household surveys objects or items or elements or units that the stakeholders would like to have information on are for example, household, person, plots, income, etc. In most cases the basic object is the household and there are usually a number of associated objects related to the basic object, and these will depend on the particular survey.

37. The example below shows the objects/units defined for the 1987 Zimbabwe Intercensal Demographic Survey (ZICDS). The household was the basic object and its associated objects were "person," "woman 12 years or above" and "deceased" (Lagerlof, 1988).

Table 9.1. Example of Objects/Units from Zimbabwe Intercensal Demographic Survey (ZICDS) 1987

Object/Unit	**Identifying Variables**	**Object/Unit Definition**	**Important Variables**	**Related Objects**	
				Object	Foreign key
HOUSEHOLD	HID = household identification (AREA, DIVISION, SUBDIV = subdivision, EANR = *EA* number, HHNR = head of household number)	A house is a group of persons who normally live and eat together, and excludes visitors.	SOH = size of household STRATUM AREA	PERSON DECEASED	HID HID
PERSON	HID, PID . PID = person identification.	The person is a usual member of the household or a visitor last night	SEX, AGE, MARSTAT = marital status, ETHNIC = ethnic group, USMEM = usual member of household, RELTH = relationship to head of household.	HOUSEHOLD WOMAN ≥12 YEARS OLD	HID HID, PID
DECEASED	HID, DID DID = deceased identification.	The deceased who was a usual member of the household during the last 12 months.	SEXD = sex of the deceased AGED = age of deceased	HOUSEHOLD	HID
WOMAN≥12 YEARS OLD	HID, DID	Every woman who is 12 years or above and who is a usual member of household or visitor last night	Number of children born	PERSON	HID, DID

38. For every object, there will be several variables of interest. Variables are properties (attributes or characteristics) of the objects. For example, the object, "person," can have age, income, occupation, marital status, etc., as variables. Variables may be qualitative or quantitative.

39. Every object should also have a unique identification. The identification of an associated object indicates the basic object it relates to, for example, "person," would be related to

"household" and would be identified by the combination of household identification (HID) and the person identification (PID), that is, the serial number within the household roster (PID).

Input to the data processing system

40. The input consists of values obtained and recorded by interviewers according to the survey questionnaire.

Output of the data processing system

41. The output of the system consists mainly of statistical tables (based on the tabulation plan), databases containing micro and macro data, etc. These will vary with respect to the type of object, type of variable and type of statistical measure. The tabulated variables are usually 'original' but may also be derived from original variables.

File organization

42. Usually one should have different file structures at the input stage and at the stage before tabulation. For example, the variable length file (versus the flat file) could be preferred for data entry for household surveys. This is because households differ in size and composition, and hence the need for variable length records during the data entry. This method uses space efficiently but is inconvenient for later processing. Eventually, however, it is often preferred that data should be organized in flat files to facilitate tabulation and the optimal use of different types of generalized software.

System flow chart

43. A reasonably detailed flow chart should be set up for the household survey. The chart is important for many reasons, one being as an instrument for making time schedules and estimation of resources needed to complete the processing of the survey. Typically, the main activities in data processing for any survey include:

(i) Data checking, editing, and coding.
(ii) Data entry, verification and validation.
(iii) Transformation of the data structure used at the input stage to a data structure suitable for tabulations.
(iv) Tabulation.

44. The systems flow chart should also include the fundamental file operations such as selection, projection, sorting and matching of files, derivation of new variables, aggregation, tabulation, and graphic presentation.

Documentation system

45. Comprehensive, clear documentation (that is, set of metadata) is important both for the development and for the maintenance of the respective data processing systems for household surveys. It is, therefore, important to document files and various operations so that persons not involved in the implementation of the original system can also use them. To ensure that the documentation is sufficient, a standardized template should be used and stored electronically together with its data.

46. As far as possible the same names, the same codes and the same data format should be used for variables in the data processing systems for all the various surveys of the survey organization if they have the same meaning. This is particularly important for variables that are used to identify the records (objects) within the file, as these variables may also be used when combining (joining) data from the different systems.

47. For the design of the data entry screens or form scanning, template tools are built into software to assist with the documentation and they need to be fully utilized to achieve effective documentation. For example, CSPro (Census and Survey Processing System) or Integrated Microcomputer Processing System (IMPS) data dictionary formats define the position of each variable in the data file - the start and end points, whether the variable is numeric or character in nature, whether it is recurring and if so how many times and the structure of the variable corrected within it. The dictionary also labels the values contained within the variable (Catherine, 2003)[25].

9.4 Survey operations and data processing

9.4.1 Frame creation and sample design

48. As discussed in chapters 3 and 4 the first stage sampling units for many household surveys are the census enumeration areas (*EA*s) defined by the most recently available national census. Creating a computer file with the list of all *EA*s in the country is a convenient and efficient way to develop the first stage sample frame. The best way to do it is with a spreadsheet program such as Microsoft Excel, with one row for each *EA* and columns for all the information that may be required.[26]

49. The frame must be easy to access and to use for various manipulations like sorting, filtering and production of summary statistics that can help in sample design and estimation. Microsoft Excel is easy to use, many know how to use it, and it has functions for sorting, filtering and aggregation that are needed when samples are prepared from the frame. The

[25] See the Annex to Chapter 9 for more information on CSPro and IMPS.
[26] See the Annex to Chapter 9 for more information on Excel.

worksheets could easily be imported into most other software packages. It is generally more convenient to create a different worksheet for each of the sample strata.

50. The contents of the records for frame units should be as follows:

• A primary identifier, which should be numerical, should be included. It should have a code that uniquely identifies all the administrative divisions and subdivisions in which the frame unit is located. It will be an advantage if the frame units are numbered in geographical order. Usually, *EA* codes already have the above properties;

• A secondary identifier, which will be the name of the village (or other administrative subdivision) where the frame unit is located, is also important. Secondary identifiers are used to locate the frame unit on maps and in the field;

• A number of sampling unit characteristics, such as measure of size (population, households), urban or rural, population density, etc. should be included on the file. Such characteristics may be used for stratification, assigning selection probabilities, and as auxiliary variables in the estimation;

• Operational data such as information on changes in units and indication of sample usage should be included.

51. The selection procedures and the selection probabilities for all of the sample units at every stage must be fully documented. When master samples are used, there should be records showing which master sample units have been used in samples for particular surveys. A standard identification number system must be used for the sampling units.

52. The Master Sample in Namibia based on the Population and Housing Census of 1991, can serve as an example of what is discussed above (Namibia Central Statistics Office, 1996).

Example:
To be able to select a random sample of geographical areas in Namibia it was necessary to create a sample frame of geographical areas. For this purpose a frame of geographical areas – Primary Sampling Units (*PSU*s) - was created. The areas on average contained about 100 households and most were in the range of 80-100 households. The areas were built from enumeration areas of the 1991 Population and Housing Census. Small *EA*s were combined with adjacent *EA*s to form *PSU*s of a sufficient size. The rule of the thumb was that a *PSU* should be at least 80 households. In total there were about 1685 such *PSU*s stratified into strata of *PSU*s by region and rural, small urban and urban areas.

53. The stratification was based on a classification of *EA*s conducted during the preparations for the 1991 Census. A total of 32 strata were created and within the 32 strata the *PSU*s were listed in a geographical order. In the urban and small urban areas the *PSU*s were also listed by income level of the areas. First in the lists for urban and small urban strata were the high-income areas followed by the middle/low income areas.

54. The Central Bureau of Statistics prepared Microsoft Excel files of the master sample frame of *PSU*s. The frame contained:

- Region
- A unique *PSU* number
- Income level (only for urban areas)
- District
- *EA* number(s)
- Number of households as per 1991 Census
- Cumulative number of households by stratum
- Population by sex according to the Census
- Master sample status (whether the *PSU* is in the master sample or not)
- Master sample *PSU* number (only for *PSU*s in the master sample)
- Weights (raising or inflation factors) (only for *PSU*s in the master sample)

55. There was one Microsoft Excel file for each region and within each Excel file the *PSU*s were grouped according to rural, small urban, urban - high income and urban - middle/low income.

56. Pettersson (2003) discusses detailed issues relating to Master Samples. Munoz (2003) further discusses how a computerized frame such as described above may also prove instrumental in implementing the sampling procedure for a household survey, by guiding it through its main stages: organization of the first stage frame, usually built on the latest Population and Housing Census results (*EA*s); selection of primary sampling units with probability proportional to size, (size measured by the number of households, dwelling units or population); updating the spreadsheet on the basis of listing of selected households and calculation of selection probabilities and the corresponding sampling weights. The sub-section on point estimation procedures and the calculation of weights, which comes later in this chapter, provides some detailed discussion of the computation of selection probabilities and the corresponding weights. The required data for these calculations would be obtained from such a spreadsheet as discussed above, and the calculation of weights can be done using the spreadsheet as demonstrated by Munoz.

9.4.2 Data collection and data management

57. Household surveys can produce large amounts of questionnaires. The procedures for the physical handling and accounting for these masses of documents need to be well thought out and set up at an early stage, if chaos is to be avoided. The routines for the manual handling (filing and retrieval) of questionnaires must be carefully planned and operational well before the data start arriving from the field. One important part of such a system is to estimate the data expected, so that files, boxes, etc. can be acquired, and space on shelves or in cupboards can be allocated. A second part of the system is a log, where information regarding the questionnaires is entered on arrival, and where the flow of the data through the system can be followed. All these are key aspects of data management, and are important prerequisites for the successful management and implementation of any survey data processing strategy.

58. The physical security of data (questionnaires) is also an important issue at this stage. This has been seen as one area where the use of image scanning is found attractive. If upon arrival at the office the questionnaires are scanned there is a reduction in the risk of the data on the questionnaires being lost through any possible mishaps. The scanned questionnaires can be backed up onsite and offsite after they have been scanned and this is an additional level of security which image scanning allows (Edwin, 2003).

9.4.3 Data preparation

59. The data collected need to be entered into a data file. Transferring data on questionnaires into computer-readable data is termed data entry. In this connection it is often necessary to categorize variable values, which have been given as open answers; this categorization process is referred to as coding. By editing the data obtained, one may identify data which are erroneous. Then appropriate measures may be taken to check the suspected errors, for example, by making renewed contact with the source of information. Such checks may be followed by an update (correction). The processing steps include: data entry, coding, editing, checking, and update/correction. Collectively, they are referred to here as the data preparation step of the survey processing.

9.4.3.1 Strategies for data preparation

60. Munoz (2003) tackles the various aspects and configurations for data preparation in detail. The most prevalent organizational set-up for household surveys entails the undertaking of data preparation in central locations, after the data collection in the field. An alternative arrangement involves integrating data entry to field operations. The more recent innovation is the computer-assisted interviewing technique.

Centralized data preparation

61. This is the only option that existed prior to the advent of personal computers. It largely remains the main approach used for surveys in developing countries, with some modification due to the introduction of microcomputers. Under the approach, data entry is taken as an industrial process to be undertaken in one or a number of locations after the interviews. This could be done at the headquarters of the national statistics offices or in its regional offices.

Data preparation in the field

62. More recently, the integration of computer-based quality controls to field operations has been seen as one of the keys to improving the quality and timeliness of household surveys. Under this strategy, data entry and consistency controls are undertaken as an integral part of field operations.

63. One form which this can take is having the data entry operator work with a desktop computer in a fixed location (for example, in the regional office of the national statistics office) and organizing fieldwork so that the rest of the team visits each survey location (generally a primary sampling unit) at least twice, to give the operator time to enter and verify the consistency of the data in between visits. During the second and subsequent visits, interviewers re-ask the relevant households the questions where errors, omissions or inconsistencies are detected by the data entry program.

64. Another approach is having the data entry operator work with a notebook computer and join the rest of the team in their visits to the survey locations. The whole team stays in the location until all the data are entered and is qualified as complete and correct by the data entry program.

65. The perceived relative advantages of integrating data collection and data preparation include the scope for higher data quality since errors can be corrected while interviewers are still in the field, the possibility to generate databases and undertake tabulation and analysis soon after the end of field operations, and greater scope for standardizing the data collection by the interviewers.

66. Under the two approaches described above the need for consistent availability of electric power supply, where the operations are to take place, is critical. In countries with poor supplies of electricity both options would simply not be viable, and this is the case in most developing countries, especially the rural areas.

Computer-assisted interviewing

67. Computer Assisted Personal Interviewing (CAPI) is a form of personal interviewing, but instead of completing a questionnaire on paper, the interviewer brings a laptop or hand-held computer to enter the data directly into the database. This method saves time involved in the processing of the data, as well as saving the interviewer from carrying around hundreds of questionnaires. However, although the technology has been available for many years, very little has been done to seriously apply this strategy to complex surveys in developing countries. This type of data collection method can be expensive to set up and requires that interviewers have computer and typing skills. Computer-based interviewing also requires well-structured interviews, with a beginning and an end. However, most surveys in developing countries require multiple visits to each household, separate interviews for each member of the household, etc., in a process that is not strictly structured but rather intrinsically driven by the interviewer.

9.4.3.2 Coding and editing of survey data

68. Data checking, editing, and coding represent, probably, the most difficult phase of data processing. It is organizing data management and data preparation where newly trained survey professionals often encounter great difficulties.

9.4.3.2.1 Coding

69. The objective is to prepare the data in a form suitable for entry into the computer. The coding operation mainly involves assigning numerical codes to responses recorded in words (for example, geographic location, occupation, industry, etc). It may also entail transcription, in which numeric codes already assigned and recorded during interview are transferred onto coding sheets.

70. A manual should be prepared to give explicit guidance to the coders. Such a manual should contain a set of disjoint categories, which cover all acceptable responses to the questions under consideration. For a large-scale household survey, it is desirable to maximize the extent to which the questions are closed and pre-coded.

9.4.3.2.2 Editing and checking of data

71. The aim of checking and/or editing questionnaires is (i) to achieve consistency within the data and consistency within and between tables and (ii) to detect and verify, correct or eliminate outliers, since extreme values are major contributors to sampling variability in the survey estimates.

72. Editing involves revising or correcting the entries in the questionnaires. It might be viewed as a validating procedure, where inconsistencies and impossibilities in the data are detected and corrected, or as a statistical procedure, where checks are undertaken based on a statistical analysis of the data. The trend is that the computer does an increasing part of the editing, either at data entry or in special edit runs of the data. Such edit runs may or may not be interactive. Interactive means that the operator may perform the immediate correction of the errors. However, the rectification of the more complex errors requires more in-depth analysis and time before the right correction can be found and non-interactive edit runs would be more suitable. The reference material by Olsson (1990) provides some detailed discussion of the various aspects of checking and editing of survey data.

Checking and manual editing

73. The main task of checking or manual editing is to detect omissions, inconsistencies, and other obvious errors in the questionnaires before subsequent processing stages. Manual editing should begin as soon as possible and as close to the data source as possible, such as the provincial, district, or lower level offices. Ideally, the majority of errors in the data should be detected and corrected in the field before the forms are sent to the processing center. Thus, the training and manual of instructions usually instruct the interviewer and supervisor to check questionnaires and correct any errors while in the field before the data are sent away. This is an important and difficult task whose performance becomes a function of the quality of field materials, the effectiveness of the supervision, survey management, etc.

Computer-assisted editing

74. Computer editing can be done in two ways: (i) interactively at the data entry stage or (ii) using batch processing after data entry, or some combination of the two. Interactive editing tends to be more useful in the case of simple errors, for example, keying errors, otherwise it would delay the data capture process in the case of errors that need consultation with supervisors. The handling of such errors, including non-response, should be left to a separate computer editing operation.

75. Programs for computer-assisted editing are often designed using database programs such as Integrated Microcomputer Processing System (IMPS), Integrated System for Survey Analysis (ISSA), Census and Survey Processing System (CSPro), Visual Basic, and Microsoft Access.[27] The simplest programs scan through the data, record by record, and note inconsistencies based on edit rules written into the program. In more sophisticated editing programs, variables (for example identification variables) may be compared between files and discrepancies noted. The output from the systems consists of error lists, which often are manually checked against the raw data. The errors are corrected in a copy of the raw data life.

9.4.3.2.3 Types of Checks

76. Data on the questionnaires need to be subjected to different types of checks and the typical ones include range checks, checks against reference data, skip checks, consistency checks, and typographic checks (Munoz, 2003).

Range checks

77. These are intended to ensure that every variable in the survey contains only data within a limited domain of valid values. Categorical variables can only have one of the values predefined for them on the questionnaire (for example, gender can only be coded "1" for males or "2" for females). Chronological variables should contain valid dates and numeric variables should lie within prescribed minimum and maximum values (such as 0 to 95 years for age). A special case of range checking occurs when the data from two or more closely related fields can be checked against external reference tables.

Skip Checks. These verify whether the skip patterns and codes have been followed appropriately. For example, a simple check verifies that questions to be asked only to school children are not recorded for a child who answered "no" to an initial question on school enrollment. A more complicated check would verify that the right modules of the questionnaire have been filled in for each respondent. Depending on his or her age and sex, each member of the household is supposed to answer (or skip) specific sections of the questionnaire. Women aged 15 to 49 may be included in the fertility section, but men may not, for example.

[27] More details regarding the above software are provided in the Annex to Chapter 9.

Consistency checks

78. These checks verify that values from one question are consistent with values from another question. A simple check occurs when both values are from the same statistical unit, for example the date of birth and age of a given individual. More complicated consistency checks involve comparing information from two or more different units of observation. For example, parents should be at least 15 years older than their children, spouses should be of different genders, etc.

Typographic check

79. A typical typographic error is the transposition of digits (such as entering "14" rather than "41") in a numeric input. Such a mistake of age might be caught by consistency checks with marital status or family relations. For example, a married or widowed adult aged 41 whose age is mistakenly entered as 14 will show up with an error flag in the check on age against marital status. However, the same error in the monthly expenditure on meat may easily pass undetected since either $14 or $41 could be valid amounts. A typical measure of handling this is having each questionnaire entered twice, by two different operators.

9.4.3.2.4 Handling missing data

80. When the survey has reached the processing stage, there will most certainly remain a sizeable amount of missing data. Some households may have moved or refused to answer. Some questions in the questionnaire may not have been answered. Or some data may have been faked or be inconsistent with other information in the questionnaire. Whatever the reason, the effect is that the records are missing, empty or partly empty.

81. It is important to distinguish between missing data, that is, data that should have been there but for which the correct value is unknown, and zero data. For example, one questionnaire might be empty because the household refused to participate, whereas a second questionnaire may be empty because the household did not, for example, plant any crop on their fields. In the second case, the variable "area planted" should be zero. Such records must be retained in the file for analysis and tabulation.

82. The approach to take for genuinely missing data depends on which kind of data is missing. A selected sample element can be totally missing due to refusal by the household to take part in the survey or due to inability by the household respondent to answer the entire set of questions in the questionnaire. In such instances, 'unit non-response' is said to have occurred.

83. If a respondent is able to answer only some of the questions and not the others then 'item/partial non-response' has occurred because some, but not all, of the data have been obtained for the household.

84. Missing data of either type gives rise to biased survey estimates, as has been repeatedly emphasized in this handbook. For a detailed discussion on appropriate treatment of non-response including methods of adjusting for it see chapter 6.

85. In the case of partial non-response, it may be necessary to substitute the missing values with some reasonable estimate, in order to achieve consistency in the totals. This is known as imputation, as noted in chapter 6. There are several approaches that can be used to impute substitute values. Some of them are:

Mean value imputation: the mean value (in the *PSU* or whole data set) is used to impute the missing value.

"Hot deck'" imputation: a (donor) record similar to the incomplete record is sought. The missing values are borrowed from such a record. The donor record should have passed all edit tests.

Statistical imputation: the missing value is imputed using a relation (regression, ratio) with some other variable, derived from complete data.

86. The above are some of the methods available for imputation, and it should be noted that there are several more methods for that purpose. The efficiency of the imputation will, of course, depend on how successful the imputation model catches the non-response. In choosing the auxiliary information available, it is important that the variable correlates with the variable to be imputed; see Olsson (1990) for more information.

9.4.3.3 Data entry

87. The objective of data entry is to convert the information on the paper questionnaires into an intermediate product (machine-readable files) that must be further refined by means of editing programs and clerical processes in order to obtain so-called 'clean' databases as a final product. During the initial data entry phase the priority is speed and ensuring that the information on the files perfectly matches the information gathered on the questionnaires.

88. The method used for entering data from the questionnaires into the computer media should be decided upon at an early stage, since it will have a considerable impact on the basic workflow, the data storage technique, the form design as well as on staff composition.

9.4.3.3.1 Key-to-disk data entry

89. This involves keying of coded data onto, for example, disk, diskette, or compact disk. Many survey organizations in developing countries have gained considerable experience in this mode of data entry. It is the main approach in use and has been reinforced by the advent of personal computers and relevant software.

Data entry application

90. Normally the application comprises three modules. One module is where all the information is entered. The second module is for the verification of the entered data. This certifies that the quality of the information entered is good and it also keeps track of the performance of the data entry operators. The third module is for the correction of entered information as there may be a need to change errors on values that were not detected during data entry or the validation processes.

91. The data entry application usually has a main menu, where the data entry person can select between data entry, verification and correction. Before working on the main menu, the user must certify, with a user name and password, that he/she has permission to enter the application. If the login fails (that is, wrong user name or password is entered), the application will shut down immediately. All user-names and passwords are stored in a user table in the back-end, where the password is encrypted. When a user logs into the system with a valid password, tables in the back-end are updated.

Data entry module

92. The data entry module is the link between the questionnaire and the data file or database. This input system must be very simple to use for the data entry operator. There are some requirements that are important, as follows:

- The data entry screen should look as much as possible like the corresponding pages of the questionnaire. The operator should very quickly be able to find from the questionnaire the corresponding field on the screen.
- The speed for data entry is very important. An operator does not want to wait for the system to evaluate each entered value. The evaluation process must therefore be very fast, which implies that the system cannot have contact with the server more than necessary, which in this case means that values will not be saved to the database until all values of the a household are entered. The drawback with this is information for the currently entered household will be lost if for some reason the application should shut down. However, the benefit of the relatively high speed is more important.
- Each value in the questionnaire should have a numeric code to enable use of the numeric keypad, which is the basis for a high speed.
- The data entry module must have a variable validity control, where the operator immediately receives an error message when an invalid value is entered. The validity control should also take care of related values, for example, if ‘sex’ has value ‘1’ (male), then the fertility information must be disabled.
- The data entry program should of course flag as errors any situations that present logical or natural impossibilities (such as a girl being older than her mother) or are very unlikely (such as a girl being less than 15 years younger than her mother).
- It is important to keep track of the number of keystrokes and data entry time for later statistical use, for example, to predict the total data entry time.

Data verification module

93. The purpose of a verification system is to provide information on the quality of data entered and the failure rate for each data entry operator. The screen for this module has exactly the same layout as that of the data entry module, without any visible differences. Instead, the main difference is that not only is the number of keystrokes summarized, but also the number of errors is found. Options for the type of verification include *total verification*, where all *EA*s and questionnaires within an *EA* are verified, or *sample verification*, where only some of the *EA*s and some questionnaires are verified.

Data correction module

94. The data correction module is mainly used for correction of information that for some reason could not be completed in the data entry module. In this module it is possible to add, delete, or update information from a complete household down to a single value.

Supervisor administration application

95. The administration application will be the tool for the supervisors to accomplish changes in the database. The tool is mainly used for the correction of the Batch Master File (BMF) and to receive reports of user performance. It is important that:

- Supervisors have complete control of the BMF from the application. It should be possible for them to add, delete and update the BMF information.
- Users can be added and deleted and that a complete list of all users can be obtained. It should be possible to check the current status of all users, or just one single user.
- Keystroke statistics can be viewed and printed out. It should be possible to choose different time periods.
- The failure rate for a single user, and the average for all users, can be viewed and printed out.
- It is possible to reset an *EA* to data entry or data verification.
- All information that supervisors need to mange their work can be obtained from this application.

Svensson (1996) tackles the various aspects relating to key-to-disk data entry systems in detail.

Platforms for key-to-disk data entry systems

96. There are many data entry and editing program development platforms available on the market. For example, Census and Survey Processing System (CSPro) and its ancestor, Integrated Microcomputer Processing System (IMPS), have proven their ability to support the development of effective data entry and editing programs for complex national household surveys in many developing countries. They have also proved to be platforms that are easy to obtain and use (Munoz, 2003).

9.4.3.3.2 Scanning

97. The use of scanning in the data processing of censuses and surveys is growing rapidly. Just a few years ago the mainstream data entry method was synonymous to keyboard operated systems. Many competing systems were unavailable on the market. Today, the scene has changed and the best selling data entry systems are all based on scanning techniques. There are several sub-divisions of these techniques, all with their own advantages and disadvantages. The most commonly used acronyms include: OCR (Optical Character Recognition) which refers to recognition of machine printed characters; ICR (Intelligent Character Recognition) referring to recognition of hand written characters; OMR (Optical Mark Recognition) to recognition of pen or pencil marks made in predetermined positions, usually mark-boxes; and BCR (Bar Codes Recognition) referring to recognition of data encoded in printed bar codes.

98. According to Lundell (2003) the choice regarding the usage of scanning technology for statistical surveys and censuses is mainly between ICR and OMR. A country with large population would favour OMR, while a complex questionnaire favours ICR. OMR restricts the form design but offers fast processing and requires relatively less skilled staff. ICR allows for freedom in form design but processing is more demanding on computer capacity and staff skills. Bar codes are usually only used to print and retrieve identity information, for example, form numbers, since the bar code contains check-digit information to minimize errors.

99. During the scanning process, the questionnaires are scanned at speeds of between 40 and 90 sheets per minute duplex. Speed is the most important factor in the comparison of scanning over traditional forms of data entry involving keying of data. The scanning software is then used to identify the pages of the questionnaire and evaluate the contents of the questionnaire using ICR and OMR. Items queried or to be coded are sent to the verifier who reviews badly written items and codes open-ended questionnaires from the electronic lookup tables built into scanning template. There is a great deal of flexibility in how these verification checks are performed, depending on how the scanning template is set up. Critical variables can be completely or partially reviewed to maximize accuracy of the response being captured in the data-file.

100. It has been shown that use of the image scanning process can increase efficiency of data capture by 70 percent (Edwin, 2003). Many of the problems associated with scanning can be counteracted by proper technical organization of the process. For example, the problem of missing and mismatched pages can be dealt with by use of barcodes pre-printed onto the questionnaires and used as the vehicle for linking the various pages of the questionnaire. If proper maintenance and oversight of the equipment and software is maintained the long run cost of a scanning operation (including the purchase of equipment and software) compared to data keying operation can be shown to be significantly less.

101. Experience in the use of scanning for household surveys has been generally very limited especially in the sub-Saharan region. However, its use in the 2000 round of population and housing censuses has been quite significant and perhaps represents a turning point regarding its general adoption. For example, it was used in Kenya, Tanzania, South Africa, Namibia, and Zambia in their most recent censuses. Recently, it has also been used for all the Core Welfare

Indicator Questionnaire Surveys, driven by the World Bank. Countries such as Namibia and South Africa have also adopted it for their household survey programs.

9.4.3.4 File structure and organization of datasets

9.4.3.4.1 Data storage

102. For household surveys, which typically contain information on both the household level and the individual level, it could be efficient use of storage space to use a sequential or variable length form of the file because different households would have a different number of individuals attached to the household. A flat file would occupy unnecessary large proportion of empty space. A flat file would be adequate if all of the questions referred to the household as a statistical unit, but as discussed before, this is not the case. Some of the questions refer to subordinate statistical units that appear in variable numbers within each household, such as persons, crops, consumption items and so forth. Storing the age and gender of each household member as different household-level variables would be wasteful because the number of variables required would be defined by the size of the largest household rather than by the average household size.

103. The variable length file would normally be used for data entry for household surveys. This is because households differ in size and composition, and hence the need for variable length records during data entry. Although each type of record will be fixed in length and format, there will be different types of record within one file. Each file will be essentially a computerized image of the questionnaires as completed. Each line or block in the questionnaire will form a record. Each record will start with a string of identifiers linking the record to the household, unit of observation, and so on, and to the section of the questionnaire. This method uses space efficiently but is inconvenient for later processing, where cross-referencing of data from different files becomes critical.

104. CSPro, for example, uses a file structure that handles well the complexities that arise from dealing with many different statistical units, while minimizing storage requirements, and interfacing well with statistical software at the analytic phase.

105. The data structure maintains a one-to-one correspondence between each statistical unit observed and the records in the computer files, using a different record type for each kind of statistical unit. For example, to manage the data listed on the household roster, a record type would be defined for the variables on the roster and the data corresponding to each individual would be stored in a separate record of that type. Similarly, in the food consumption module a record type would correspond to food items and the data corresponding to each individual item would be stored in separate records of that type.

106. The number of records in each record type is allowed to vary. This economizes the storage space required, since the files need not allow every case to be the largest possible.

107. After the identifiers, the actual data recorded by the survey for each particular unit follow, recorded in fixed-length fields in the same order of the questions in the questionnaire. All data are stored in the standard ASCII (American Standard Code for Information Interchange) format.

108. Munoz (2003) and World Bank (1991) provide more detailed discussions regarding file management for household surveys.

9.4.3.4.2 Restructuring datasets for further operations

109. In order to facilitate appropriate analysis, the associated database must contain all information on the sampling procedure; labels for the sample design strata, primary sampling units, secondary sampling units, etc.; and sample weights for each sampling unit. The information will be needed for estimation of the required statistics and also for estimating the sampling errors of those estimates.

110. Following data entry, it is often necessary to restructure the data set and generate new files and to recode some of the existing data fields to define new variables more convenient for tabulation and analysis. This may be necessary to allow certain operations on the data including the estimation process.

111. The initial full survey data file may in fact contain information about units that are sampled from different populations (Rosen, 1991). For example, for a household budget survey, the data on sampled households as well as on sampled persons may be contained in the same initial file. To estimate statistical characteristics for the household population and the person population one needs a file with one record for each sampled household and a file with one record for each sampled person, respectively. Datasets or files based on households as units (objects) are used to produce statistics (tables, etc) on private households. Datasets or files based on individuals as units (objects) are used to produce statistics (tables) on persons from private households.

112. As already obvious from the above, there are, typically, two main types of files from household surveys - household files and individual (person-specific) files. In most cases, the files are household files in the sense that they carry values for household variables (variables relating to the observation unit or object “household”). Some of them are individual files (person files) in the sense that they carry values for variables on individuals (variables relating to observation unit or object “person”). The complete and final data files (datasets) will contain information on all responding households and individuals from each of the surveyed Primary Sampling Units.

113. The example in Figure 9.3 below illustrates how the big file for the 1987 Zimbabwe Intercensal Demographic Survey was reorganized to facilitate further processing. The second example in Figure 9.4 is for a typical household budget survey. These examples are based on material worked on by Lagerlof (1988) and Rosen (1991).

Table 9.3. Files Used for the 1987 Zimbabwe Intercensal Demographic Survey

File	Type	Contents
HOUSEHOLD	Household file	Household identification (Region, Province, District, etc.) Answers to all questions related to the household Derived variables, like household size (from the Members file), etc.
PERSON	Individual file	Household identification (HID) plus person identification (PID) Demographic characteristics: AGE, SEX, MARSTAT (marital status), USMEM (usual household member), RELTH (relationship to head of household).
DECEASED	Individual file	Household identification (HID) plus deceased identification (DID) Details of deceased who was usual member of household: SEX, AGED (age of deceased).
WOMAN≥12 years old	Individual file	HID, PID Details of every woman, in the household, at least 12 years old:

Table 9.4. Typical Files for a Household Budget Survey

File	Type	Contents
HOUSEHOLD	Household file	Household identification (Region, Province, District, etc.) Answers to all questions related to the household Derived variables, like household size (from the Members file), etc.
MEMBERS	Individual file	Household identification plus member identification Demographic characteristics: age, sex, marital status, education level, etc. of members Information on main activities: employment status, occupation, etc.
INCOME	Individual file	Household identification plus person identification plus income identification Income source:
FOOD	Household file	Household identification plus food item identification Food expenditures:
Other non-durable GOODS	Household file	Household identification plus goods item identification Goods expenditures:
DURABLES	Household file	Household identification plus durable item identification Durables expenditures:
AGRICULTURE	Household file	Household identification plus agriculture item identification Agriculture expenditures:
AGRICAPITAL	Household file	Household identification plus agriculture capital item identification Agriculture capital expenditures:

114. For tabulation purposes a flat file is necessary for most statistical software packages. Much of the available general software requires data in the flat format. In a flat file all records have the same set of variables or fields and are of the same length. A file is described as "flat" when exactly the same set of data fields exists for each respondent. The data fields are arranged identically within each record and a fixed number of records with identical layout are involved. An example is the format for a flat file of the Household File used for the 1987 Zimbabwe Intercensal Demographic Survey, Figure 9.5.

Table 9.5. Household File

Identification				**Sampling design Parameters**							**Variable values**			**Weight variable**
Stra-Tum	Sub-division	*EA*	Hh	S_h	a_h	R_h	b_{hr}	S_{hi}	M_{hi}	m_{hi}	x	y	z	w
h	r	i	j								x_{hrij}	y_{hrij}	z_{hrij}	w_{hrij}

115. The household file contains one record for each observed household, every record containing information on:

- Identification of the household
- Sampling design parameters
- Observed values of (household) variables.
- Weight variables.

Identification of the household – the combination $hrij$ says household j belongs to *EA* i in Subdivision r of Stratum h.

Sampling parameters – in this particular example these were as follows:

S_h = The 1982 number of households in the sampling stratum

a_h = The *EA* sample size in the sampling stratum

R_h = The number of sub divisions represented in the sample from the sampling stratum

b_{hr} = The number of sampled *EA*s from the (sub)-division

S_{hi} = The 1982 number of households in the *EA*

M_{hi} = The 1987 number of households in the *EA*

m_{hi} = The size of the household sample from the *EA*.

Observed variable values – x, y, z denote household variables**.**

Weight variable values – w denotes the weight variable for the household**.**

116. The Persons' file is organized analogously to the above. The minor difference is that the identification will have the person identification (Pid) and the index (k) for the individual person while "variables" refers to variables on the respective individuals.

117. The survey datasets only need to be organized as separate flat files (one for each record type) for dissemination, because the fixed-length field format of the native structure is also adequate for transferring the data to standard Database Management Systems (DBMSs) for further manipulation, or to standard statistical software for tabulation and analysis. Transferring the data to DBMSs is very easy because the native structure translates almost directly into the standard DBF format that all of them accept as input for individual tables (in this case, the record identifiers act as natural relational links between tables) (Munoz, 2003).

9.4.3.5 Estimation procedures and calculation of weights

118. Chapter 6 provides a detailed description of the rationale and the method of calculation of weights for household survey data (see also the various references by Rosen at the end of this chapter). A computation algorithm, leading from observed values to estimates of statistical characteristics is referred to as an (point) estimation procedure. In the first step, to point estimation, a weight is computed for each responding object. Then estimates of 'totals' are computed by summation of the weighted observation (observed value times the respective weight) values.

119. Munoz (2003) provides a good description of how a computerized system of Microsoft Excel spreadsheets can be used in implementing the sampling procedure for a household survey, by guiding it through its main stages: organization of the first stage frame; selection of primary sampling units with probability proportional to size and calculation of selection probabilities and the corresponding sampling weights.

120. The actual construction of weighted estimators is straightforward. One would start with the original sample dataset and create a new dataset by multiplying each observation the number of times specified by its weight. Then use the standard formulas for calculating the parameter using the weighted dataset.

121. It should be noted, however, that accurate weights must incorporate three components (Yansaneh, 2003) including various adjustments (see also chapter 6). Base weights account for the variation in the probabilities of being selected across different groups of households as stipulated by the initial design of the survey. The second adjustment is for variation in non-response rates across domains or sub-groups. Finally, in some cases there may be post-stratification adjustments to make the survey data conform to distributions from an independent source such as the latest population census.

122. Another complication for the estimation process is necessitated by the increased demand for domain level statistics. As discussed in chapter 3 a domain is a subset for which separate estimates are desired. Usually they may be specified at the sample design stage but may also be worked out from the derived data. A domain may also be a stratum, a combination of strata,

administrative regions (province, district, rural/urban level, etc.). It can also be defined in terms of demographic or socio-economic characteristics (for example, age, sex, ethnic group, poor, etc.). What follows is some attempt to describe how datasets can be constructed to facilitate estimation for domains.

123. We can start by visualizing a data (observation) file (for example, the household files) as shown for the Zimbabwe Intercensal Survey above. This file has one record for each sampled household. At the end of the survey process the file shall contain the following information for each household.

(i) Identification for the Household
(ii) Sampling parameters
(iii) Values for the study variables x, y, and z.
(iv) The value of the household's estimation weight.
(v) If the household belongs to category c or not.
(vi) If the household belongs to domain g or not.

124. These pieces of information (save for the sampling parameters) are denoted as follows:

HID = an identification label for sampled households. For simplicity we use the serial numbering, 1, 2,…, *n*. Hence, *n* stands for the total sample size.

x, y, and z are the observed values of variables of **X, Y,** and **Z** for the household.

c = 1 if the household is of category c, otherwise it is 0,

g = 1 if the household belongs to domain g, otherwise it is 0,

w = the estimation weight for the household.

125. The values of the indicator variables c and g are usually derived from values of other variables and not observed directly. For example we can have category c standing for "below the poverty line". Households are not asked if they belong to this category or not. The classification is derived, for example, from income data of the household and a stipulated poverty line. Similarly, often derivations from other variables are required to determine whether a household belongs to a specific study domain g or not (for example, domain g may consist of households with 3 + children). At the estimation stage the values of such indicators should be available in the observation file.

126. When all the data are available in the observation file, it will look as shown below, except that the sampling parameters are not included.

Table 9.6. Illustration of Household Survey Variables

Final data/observation file						
HID	**X**	**Y**	**Z**	**C**	**G**	**w**
1	x_1	y_1	z_1	c_1	G_1	w_1
2	x_2	y_2	z_2	c_2	G_2	w_2
3	x_3	y_3	z_3	c_3	G_3	w_3
.	.	.		.	.	.
w.	.	.		.	.	.
N	x_n	y_n		c_n	g_n	w_n

127. The above discussion has been confined to estimation of statistical characteristics for the household population. Estimations of statistical characteristics for the person population are carried out along the same lines. Generally, the estimation weight for a person is the same as that for the household to which the person belongs. Since all members in a typical household are listed in the questionnaire, a particular person is included in the person sample if and only if the person‘s household is included in the household sample. Hence the inclusion probability for a person is the same as the inclusion probability for the household to which the person belongs. Note, however, that the above is not true when sub-sampling within households is done. For example, in some sample designs the procedure may call for selecting only one adult per household or one male and one female; in those cases the weight for the selected individual(s) is independently calculated and is not equal to the household weight.

128. For completeness part of the estimation procedure for a household survey must include provision of estimates of the sampling (or standard) errors of the survey, especially for the most important statistics that are generated and released to the public. This subject, however, is covered in its own chapter of this handbook – chapter 7.

9.4.3.6 Tabulation, datasets for tabulation and databases

129. There are three major basic outputs from a statistical survey (Sundgren, 1995):

- *Macrodata* – ‘statistics’ representing estimates for certain statistical characteristics; these data are the primary purpose of the survey being carried.
- *Microdata* – ‘observations of individual objects’, underlying the macrodata produced by the survey; these data are essential for future use and interpretation of the survey results.
- *Metadata* – ‘data describing the meaning, accuracy, availability and other important features of the underlying micro and macro data’; these are essential for correctly identifying and retrieving relevant statistical data for a specific problem as well as for correctly interpreting and (re)-using the statistical data.

130. The household survey program should eventually produce a situation where the data archiving is based on a combination of micro-level and macro-level data. For multiple surveys where the same sample (the same households) has been used, one should aim to store the micro-data for these different surveys in an integrated fashion to facilitate combined use of the data. A prerequisite for this is very thorough documentation and description of the structure of the information collected.

131. Data storage should be considered in three phases (Lundell, 2003):

- *Storing* – during data entry data should be stored in a way that primarily works well with data entry and data cleaning methods used, as discussed earlier.
- *Warehousing* – when data have been entered and cleaned they should be added to a warehouse that is built to suit tools and methods for analysing and disseminating the data
- *Archiving* – the project data should be archived in a way that complies with long-lasting standards to ensure uncomplicated future retrieval of the data.

132. A data warehouse containing clean data can be created in several ways, using one of the following methods[5]:

- Flat files;
- Relational database (for example, Microsoft SQL Server); and
- Statistical software (for example, SAS or SPSS).

133. For long-term archiving of the final data there is one main option. The data must be saved as simple ASCII-format flat files with attached record descriptions. Most database systems and statistical software can export data to these files without much trouble and can also import data easily from these files.

References and further reading

An, A. and Watts, D. (2001), *New SAS Procedures for Analysis of Sample Survey Data, SUGI paper No. 23*, SAS Institute Inc., Cary, NC.

Arnic, A. etal (2003), "Metadata Production Systems within Europe – The Case of the Statistical System of Slovenia," Paper presented at Metadata Production Workshop, Luxembourg, Eurostat Doc 3331.

Australian Bureau of Statistics (2005), *Labour Statistics: Concepts, Sources and Methods* (6102.0), Statistical Concepts Library, Canberra.

Backlund, S. (1996), "Future Directions on IT Issues," Mission Report to National Statistical Center, Lao PDR, Vientiane.

Central Statistics Office, Namibia (1996), "The 1993/1994 National Household Income and Expenditure Survey, NHIES," Administrative and Technical Report, Windhoek.

[5] The Annex to Chapter 9 provides more information on these software packages.

Chromy, J. and Abeysasekara S. (2003), "Analytical Uses of Survey Data," United Nations Statistics Division, New York.

Chronholm, P. and Edsfeldt (1996), "Course and Seminar on Systems Design," Mission report to Central Statistics (CSS), Pretoria.

Brogan, D. (2003), "Comparison of Data Analysis Software Suitable for Surveys in Developing Countries," United Nations Statistics Division, New York.

Edwin, C. (2003), Review of Chapter on Data Processing, Analysis and Dissemination, UNSD Expert Group Meeting to Review the Draft Handbook on Designing of Household Sample Surveys, December 2003, New York.

Giles, M. (1996), *Turning Data into Information, A Manual for Social Analysis*, Australian Bureau of Statistics, Canberra.

Glewwe, B. (2003), "An Overview of Questionnaire Design for Household Surveys in Developing Countries," United Nations Statistics Division, New York.

Graubard, B. and Korn, E. (2002), "The Use of Sampling Weights in the Analysis of Survey Data," United Nations Statistics Division, New York.

International Labour Office (1990), *Survey of Economically Active Population, Employment, Unemployment and Underemployment, ILO Manual on Concepts and Methods*, ILO, Geneva.

Jambwa, M. and Olsson, L. (1987), "Application of Database Technology in the African Context," Invited Paper, 46th session of ISI, Tokyo.

Jambwa, M., Parirenyatwa, C. and Rosen, B., (1989), "Data Processing at the Central Statistical Office – Lessons from Recent History," Central Statistics Office, Harare.

Lagerlof, B. (1988), "Development of Systems for National Household Surveys – SCB R&D Report," Statistics Sweden, Stockholm.

Lehtonen, R. and Pahkinen, E. (1995), Practical Methods for Design and Analysis of Complex Surveys, Wiley & Sons, New York.

Lundell, L. (1996), "Information Systems Strategy for CSS – Report to Central Statistical Service (CSS)," Pretoria.

_____ (2003), "Census Data Processing Experiences - Report to Central Bureau of Statistics (CBS)," Windhoek.

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

Munoz, J. (2003), "A Guideline for Data Management of Household Surveys," United Nations Statistics Division, New York.

Olofsson, P. (1985), "Proposals for Survey Design, Kingdom of Lesotho," Report on short term mission on A Labour Force Survey, Bureau of Statistics, Masaru.

Olsson, U. (1990), *Approaches to Agricultural Statistics in Developing Countries – an Appraisal of ICO's Experiences*, Statistics Sweden International Consulting Office, Stockholm.

_____ (1990), "Applied Statistics Lecture Notes, Special Reports TAN 1990:1, Statistics Sweden International Consulting Office, Stockholm.

Pettersson, H. (2003), "The Design of Master Sampling Frames and Master Samples for Sample Surveys in Developing Countries," United Nations Statistics Division, New York.

Puide, A. (1994), Report on a mission to Takwimu, Dar-es-Salaam November 21 – December 21, 1994, TASTAT 1994:20, Statistics Sweden International Consulting Office, Stockholm.

Rauch, L. (2001), *Best Practices in Designing Websites for Dissemination of Statistics*, United Nations Statistical Commission and Economic Commission for Europe, Geneva.

Rosen, B. (1991), "Estimation in the Income, Consumption and Expenditure Survey," Report on Mission to Central Statistical Office, Harare, (ZIMSTAT 1991: 8:1), Harare.

Rosen, B. and Sundgren, B. (1991), Documentation for re-use of microdata from surveys carried out by Statistics Sweden, Working Paper for Research and Development Unit, Statistics Sweden, Stockholm.

Rosen, B. (2002), "Framework for the Master Sample, Kingdom of Lesotho," Report on short term mission to Bureau of Statistics, LESSSTAT 2002:7, Masaru.

Rosen, B. (2002), "Estimation Procedure for Master Sample Surveys," Report on short-term mission to Bureau of Statistics, Kingdom of Lesotho, Masaru.

Shah, B., Barnwell, B. and Bieler, G. (1996), *SUDAAN User Manual*: Release 7.0, Research Triangle Institute, Research Triangle Park, NC.

Silva, P. (2002), "Reporting and Compensating for Nonsampling Errors for Surveys in Brazil: Current Practice and Future Challenges," United Nations Statistics Division, New York.

SPSS, Inc. (1988), *SPSS/PC+V2.0 Base Manual*, Chicago.

Sundgren, B. (1984), *Conceptual Design of Databases and Information Systems*, P/ADB Report E19, Statistics Sweden, Stockholm.

_____ (1986), *User-Oriented Systems Development at Statistics Sweden*. U/ADB Report E24, Statistics Sweden, Stockholm.

_____ (1991), *Information Systems Architecture for National and International Statistics Offices – Guidelines and Recommendations*, United Nations Statistical Commission and Economic Commission for Europe, Geneva.

_____ (1995), *Guidelines: Modelling Data and Metadata*, United Nations Statistical Commission and Economic Commission for Europe, Geneva.

Svensson, R. (1996), "The Census Data Entry Application," Report from a mission to Central Statistical Service (CSS), Pretoria.

Thiel, L. (2001), "Design and Developing a Web Site," Report on short term mission LESSTAT: 2001:17, Kingdom of Lesotho, Bureau of Statistics, Masaru.

United Nations Statistics Division (1982), *National Household Survey Capability Programme, Survey Data Processing: A Review of Issues and Procedures*, New York.

_____ (1985), *National Household Survey Capability Programme, Household Income Expenditure Surveys: A technical study*. New York.

Verma V. (1982), *The Estimation and Presentation of Sampling Errors, World Fertility Survey*, Technical Bulletin No. 11, Voorburg, Netherlands.

Wallgren, A., Wallgren, B., Persson, R., Jorner, U. and Haaland, J. (1996), *Graphing Statistics and Data, Creating Better Charts*, Sage Publications Inc., California, London, New Delhi.

World Bank (1991), *The SDA Survey Instrument - An Instrument to Capture Social Dimensions of Adjustment*, Poverty and Social Policy Division, Technical Department, Africa Division, Washington.

Yansaneh I. (2003), "An Overview of Sample Design Issues in Household Surveys in Developing Countries," United Nations Statistics Division, New York.

Annex to Chapter 9

Software options for different steps of survey data processing

Type of operation	Software options
Database management system	Microsoft Structured Query Language (SQL) Server 2000, Standard Edition. Microsoft Access. Statistical Analysis System (SAS).
Data entry and editing	Visual Basic. Microsoft Access. Integrated Microcomputer Processing System (IMPS). Census and Survey Processing System (CSPro).
Data retrieval	Statistical Analysis System (SAS). Statistical Package for Social Sciences (SPSS). Microsoft Access. Microsoft Excel.
Tabulation, analysis, and presentation	Microsoft Word. Microsoft Excel. Statistical Analysis System (SAS). Statistical Package for Social Sciences (SPSS)
Variance estimation	CENVAR - variance calculation component of the Integrated Microcomputer Processing System (IMPS). Computation and Listing of Useful STatistics on Errors of Sampling (CLUSTERS). Integrated System for Survey Analysis (ISSA). Survey Data Analysis (SUDAAN). Statistical Analysis System (SAS). Statistical Package for Social Sciences (SPSS). Cluster Analysis and Regression Package (PC-CARP).

(a) The Microsoft Office, which is developed by Microsoft Corporation, is a comprehensive package containing various software programs, for example:

- Microsoft Office Access, the Office database management program, offers an improved ease of use and an expanded ability to import, export, and work with XML data files;
- Microsoft Office Excel, the Office spreadsheet program, includes support for XML and new features that make it easier to analyze and share information.
- Microsoft Office Word is the Office word processor;
- Microsoft SQL Server 2000 is the server database for full enterprise project and resource management capabilities; and
- Microsoft Office Outlook, the Office personal information manager and communication program provides a unified place to manage e-mail, calendars, etc.

http://www.microsoft.com/office/system/overview.mspx#EDAA

(b) Microsoft released Visual Basic in 1987. Visual Basic is not only a programming language, but also a complete graphical development environment. This environment allows users with little programming experience to **quickly** develop useful Microsoft Windows applications which have the ability to use OLE (Object Linking and Embedding) objects, such as an Excel spreadsheet. Visual Basic also has the ability to develop programs that can be used as a front end application to a database system, serving as the user interface which collects user input and displays formatted output in a more appealing and useful form than many SQL versions are capable of.
Visual Basic's main selling point is the ease with which it allows the user to create nice looking, graphical programs with little coding by the programmer. The main object in Visual Basic is called a **form** and this facilitates the development of data entry screens.

http://www.engin.umd.umich.edu/CIS/course.des/cis400/vbasic/vbasic.html

(c) CENVAR is the variance calculation component of the Integrated Microcomputer Processing System (IMPS), a series of software packages for entry, editing, tabulation, estimation, analysis, and dissemination of census and survey data. The U.S. Bureau of the Census developed integrated Microcomputer Processing System (IMPS).

http://www.census.gov/ipc/www/imps/

(d) CENVAR is based on the PC CARP (Cluster Analysis and Regression Package for Personal Computers) software originally developed by Iowa State University. PC-CARP uses the linearization procedure for variance calculation.

http://www.census.gov/ipc/www/imps/

(e) CSPro (Census and Survey Processing System) is a public-domain software package for entering, editing, tabulating and mapping census and survey data. CSPro was designed and implemented through a joint effort among the developers of IMPS and ISSA: the United States Census Bureau, Macro International, and Serpro, S.A. Funding for the development is provided by the Office of Population of the United States Agency for International Development. CSPro is designed to eventually replace both IMPS and ISSA.

http://www.census.gov/ipc/www/imps/

(f) Computation and Listing of Useful STatistics on Errors of Sampling (CLUSTERS) package was originally developed to compute sampling errors for the World Fertility Survey (WFS) programme. It uses the Taylor linearization method for computing sampling errors. It has also been used to compute sampling errors for various household surveys, especially those under the Demographic Health Survey programmes, in many developing countries. Verma V. (1982): **The Estimation and Presentation of Sampling Errors, World Fertility Survey, Technical Bulletin No. 11.**

(g) Macro International Inc developed integrated System for Survey Analysis (ISSA) specifically for the Demographic Health Survey program. It has been used for all aspects of data processing,

data entry, editing, and tabulation. It also has a sampling error module to allow the calculation of sampling errors for complex demographic rates such as fertility and mortality rates using the Jacknife method. Macro International Inc. (1996), **Sampling Manual, DHS-III Basic Document No. 6**

(h) Statistical Analysis System (SAS) was developed by SAS Inc. in 1966. It is a computer package for data analysis, file management and the calculation of sampling errors. An, A and Watts, D (2001), present some of the latest features in SAS.

(i) Statistical Package for Social Sciences (SPSS) was developed by SPSS Inc. It is a computer package for data analysis and file handling, etc. SPSS, Inc. (1988), SPSS/PC+V2.0 Base Manual, presents some of the latest features.

(j) Survey Data Analysis (SUDAAN) is a comprehensive sample survey (and correlated data) software package with analytical strengths for both descriptive and modeling analyses. Research Triangle Institute developed it. More details can be obtained from Shah B. et al (1996).

Annex: Overview of sample survey design

1. Sampling is a technique by which a part of the population is selected and results from this fraction are generalised on the whole population from which the part or sample has been selected. In general there are two types of samples, namely probability and non-probability samples. Our focus in this handbook is on probability samples.

A.1 Sample design

2. At the outset it should be stressed that sample design cannot be isolated from other aspects of survey design and implementation. In general, sampling theory is concerned with how, for a given population, the estimates from the survey and the sampling errors associated with them are related to the sample size and structure. In practice sample design involves the determination of sample size, structure and takes into account costs of the survey.

Procedures of selection, implementation and estimation

- Each element in the population should be represented in the frame from which the sample is to be selected.
- The selection of the sample should be based on a random process which gives each unit a specified probability of selection.
- All and only selected units must be enumerated.
- In estimating population parameters from the sample, the data from each unit/element must be weighted in accordance with its probability of selection.

Significance of probability sampling to large-scale household surveys

- It permits coverage of the whole target population in sample selection.
- It reduces sampling bias.
- It permits generalization of sample results to the population from which the sample is selected.
- It has been argued that it allows the surveyor to present results without having to apologise for using non-scientific methods (Kish, 1965).
- It allows the calculation of sampling errors, which are reliability measures.

Basic requirements for designing a probability sample

- The target population must be clearly defined.
- There must be a sampling frames or frames in case of multi-stage samples.
- The objectives of the survey must be unambiguously specified in terms of:

    a. Survey content
    b. Analytical variables
    c. Level of dissagregation (e.g. do you need estimates or data at national, rural-urban, provincial, district, etc. levels?).

- Budget and field constraints should be taken into account.
- Precision requirements must be spelled out in order to determine the sample size.

3. Random selection of units reduces the chance of getting a non-representative sample. Randomisation is a safe way to overcome the effects of unforeseen biasing factors. The method of sample selection used depends to a greater extent on the sampling scheme being used. The more complex the sample designs the more demanding the selection procedures required.

Survey units and concepts

4. ***Elements***: Elements (units) of a population are units for which information is sought. They can be the elementary units comprising the population about which inferences are to be drawn. For example in a household survey fertility women in the reproductive ages usually ultimate elements. To facilitate data collection in a survey it is absolutely essential that elements should be well defined and physically easy to identify.

5. ***Population***: The population is the aggregate of elements defined above. Elements are, therefore, the basic units that comprise and define the population. It is essential to define the population in terms of:

- Content, this calls for the definition of the type and characteristics of the elements which comprise the population.
- Extent refers to geographic boundaries as they relate to coverage.
- Time would refer to the time period to which the population refers.

6. ***Observational unit***: These are units from which the observations are obtained. In interview surveys they are called respondents. Reporting units are often the elements, as in surveys of attitude. Note that in some cases observational and reporting units may be different. For example in a survey of children under five, parents will normally give, as proxies, information pertaining to their children. In such cases the selected children will be observational units while parents are reporting units.

7. ***Sampling unit***: The sampling units are used for selecting elements into the sample. In element sampling each sampling unit contains one element, while in cluster sampling, for instance, a sampling unit comprises a group of elements called a cluster. For example an enumeration area (EA) which may be a first stage sampling unit contains a cluster of households. It is possible for the same survey to use different sampling units. A good example is multi-stage sampling which uses a hierarchy of sampling units (see chapter 2).

8. ***Sample units***: Selected sampling units may be termed sample units and the values of the characteristics under study for the sample units are known as sample observations.

9. ***Unit of analysis***: This is a unit used at the stage of tabulation and analysis. Such a unit may be an elementary unit or group of elementary units. It should be noted that the unit of

analysis and the reporting unit need not necessarily be identical. The reporting unit in a survey of children under five may be parents while the unit of analysis will be children under five.

10. ***A sampling frame*** is used to identify and select sampling units into the sample and for making estimates based on sample data. This implies that the population from which the sample has to be selected has to be represented in a physical form. The frame ideally should have all sampling units belonging to the population under study with proper identification particulars. Frames should be exhaustive and preferably mutually exclusive. The commonly used types of frames in surveys are: list, area and multiple frames.

11. ***List frame*** contains a list of sampling units from which a sample can be directly selected. It is preferable that the frame should have relevant and accurate information on each sampling unit such as size and other characteristics. The additional information helps in designing and/or selecting efficient samples.

12. ***Area Frames*** multi-stage frames are commonly used in household surveys. In this connection the frame consists of one or more stages of area units. In a two stage sample design, for example, the frame will consist of clusters, which can be called primary sampling units (PSUs), in selected PSUs a list of households becomes the second stage frame. In general, frames are needed for each stage of selection. The durability of the frame declines as one moves down the hierarchy of the units.

13. ***Area units*** cover specified land areas with clearly defined boundaries which can be physical features such as roads, streets, rivers rail lines, or imaginary lines representing the official boundaries between administrative divisions. Census enumeration areas (EAs) are usually established within the smaller administrative units that exist in a country. This facilitates the cummulation of counts for the administrative units as domains.

14. The frame or frames used for a household survey should be able to provide access to all the sampling units in the survey population such that every unit has a known and non-zero probality of selection in the sample. Access can be achieved by sampling from the frames usually through two or more stages of selection. The frame for the first stage of sampling must include all the designated sampling units. At subsequent stages of sample selection frames are needed only for the sample units selected at the preceding stage.

15. Sampling frame can be stored either on hardy copy or electronic media.

A.2 Basics of probability sampling strategies

16. The presentation covers simple random, systematic, stratified and cluster sampling.

A.2.1 Simple random sampling

17. Simple random sampling (SRS) is a probability sample selection method where each element of the population has an equal chance/probability of selection. Selection of the sample can be with or without replacement. This method is rarely used in large-scale household surveys because it is costly in terms of listing and travel. It can be regarded as the basic form of

probability sampling applicable to situations where there is no previous information available on the population structure. SRS is attractive for being simple in terms of selection and estimation procedures (e.g. sampling errors).

18. While SRS is not very much used in practice, it is basic to sampling theory mainly because of its simple mathematical properties. Most statistical theories and techniques, therefore, assume simple random selection of elements. Indeed all other probability sample selections may be seen as restrictions on SRS, which suppress some combinations of population elements. SRS serves two functions:

- Sets a baseline for comparing the relative efficiency of other sampling techniques.
- It can be used as the final method for selecting the elementary units, in the context of the more complex designs such as clustering and stratified sampling.

19. Considering a finite population of 100 house holds

$H_{1,}H_{2,}$-----H_{i}----------H_{100}

with income values $X_{1,}X_{2,}$ $---,X_{i},---,X_{100}$

The probability of any particular unit being selected is $\frac{1}{100}$

20. In drawing the sample, households can be numbered serially and using random numbers a sample, say, of size 25 can be selected. For equal probability selection method (EPSEM) f is the overall sampling fraction for the elements. Thus, $f = n/N$. If n =25 and N=100, the sampling fraction is $\frac{25}{100} = \frac{1}{4}$

21. There are two common methods of sample selection, namely:

a. Simple random sampling with replacement (SRSWR).
b. Simple random sampling without replacement (SRSWOR).

22. It is better intuitively to sample without replacement as you get more information because there is no possibility of repetition of sample units.

A.2.1.1 Simple random sampling with replacement

23. SRSWR is based on a random selection from a population by replacing the chosen element in the population after each draw. The probability of selection of an element remains unchanged after each draw, and any selected independent samples are independent of each other. This property explains why SRS is used as the default sampling technique in many theoretical

statistical studies. In addition, because the SRS assumption considerably simplifies the formulas for estimators, such as variance estimators, it is used as a reference.

Estimation

24. Given a sample of n units selected using SRSWR for which information on variable x, has been collected the mean, variance and population estimate are given by

Mean

$$\bar{x} = \frac{1}{n}\sum_{i}^{n} x_i = \frac{1}{n}\left[x_1 + x_2 + --- + x_n\right] \qquad \textbf{(A.1)}$$

$x_1 = 24,\quad x_2 = 30,\quad x_3 = 27,\quad x_4 = 36,\quad x_5 = 31,\quad x_6 = 38,\quad x_7 = 23,\quad x_8 = 40,$
$x_9 = 25,\ x_{10} = 32$

$$\bar{x} = \frac{24 + 30 + 27 + --- + 25 + 32}{10} = 30.6$$

Variance

$$V(\bar{x}) = \frac{s^2}{n} \qquad \text{(A.2)}$$

where $s^2 = \frac{1}{n-1}\sum_{i}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}\left[\sum_{i}^{n} x_i^2 - \frac{x^2}{n}\right] = \frac{1}{n-1}\left(\sum x_i^{\,2} - n\bar{x}^{\,2}\right)$

and shows calculation $x^2 = (\sum x_i)^2 = 93{,}636$

$$s^2 = \frac{(9{,}684 - 93{,}364)}{9} = 35.56$$

$$V(\bar{x}) = \frac{35.56}{30} = 3.56$$

A.2.1.2 Simple random sampling without replacement

25. The simple random sampling without replacement (SRWOR) is the most frequently used form of simple random sampling procedure. In this procedure, the selection process is continued

until *n* distinct units are selected and all repetitions are ignored. This is the same as retaining the unit or units selected and selecting a further unit with equal probability from the remaining units in the population.

26. The following are some of the properties SRSWOR:

- It gives a fixed sample size.
- Results in equal probability of selection for every element/unit (EPSEM).
- Like in SRSWR the sample mean and variance are unbiased estimates of population parameters.

▪ ***Example***

27. Total number of primary schools in a region is 275. A sample of 55 is selected without replacement.. Figures below show number of employees (y_i) in each of the selected schools, using SRSWOR.

5	10	32	6	8	2
15	16	35	7	50	6.
2	6	47	20	20	6
7	6	35	6	16	2
21	2	48	4	15	2
7	5	46	6	7	
4	4	8	2	6	
7	2	7	8	2	
5	12	10	6	2	
2	40	7	7	19	

$$\sum y_i = 688$$

$$\sum y_i^2 = 18{,}182$$

The sample mean is

$$\bar{y} = \frac{688}{55} = 12.5$$

The variance of the sample mean is

$$V(\bar{y}) = 1\text{-f}\ \frac{s_y^{\ 2}}{n} \qquad \text{(A.3)}$$

Where *1-f* is the population correction factor and

$$s_y^2 = \frac{1}{n-1}\left[\sum y_i^{\ 2} - \frac{\left(y_i^{\ 2}\right)}{n}\right] \qquad \text{(A.4)}$$

$$= \frac{1}{54}\left[28{,}182 - \frac{(688)^2}{55}\right]$$

$$= 177.33$$

$$\text{i.e. } V(\bar{y}) = \left(1 - \frac{55}{275}\right)177.3/55$$

$$= 2.579$$

$$\text{Se}(\bar{y}) = \sqrt{2.579} \quad = 1.6$$

Coefficient of variation

28. The coefficient of variation (CV) measures the precision of the estimator. In some situations it is useful to consider some relative measures instead of absolute measures of variation. The absolute measures such as standard error appear in the units of measurement of the variable and that can cause difficulties in some comparisons. The common relative measure of variation is the coefficient of variation, in which the unit of measurement is cancelled by dividing the standard error with the mean.

$$cv(\bar{x}) = \frac{se(\bar{x})}{\bar{x}} \qquad \text{(A.5)}$$

- ***Example***

Given the mean as 13 and standard error as 3.55, the ***cv*** will be

$$\frac{3.55}{13} = 0.27 \quad \text{or } 27\%$$

29. Coefficients of variation are useful for variables that are always or mostly positive. Such occur frequently in surveys, especially as count data. Comparison of variability of these items becomes more meaningful when expressed in relative terms. For example the spread of a variable e.g. income between two countries can be best compared by the coefficients of variation rather than standard errors, because the latter will be measured using different currency units. In this case the coefficient of variation may provide a reasonable comparison of average income because of the absence of units of measurement.

Confidence interval

30. In general if t is unbiased and normally distributed, the interval $\{t - k\delta(t), t + k\delta(t)\}$ is expected to include the population parameter in P % of the cases.

31. For example a 95 per cent confidence interval strictly means that, if we assert that the true unknown mean θ lies in this interval, then we shall be correct (on the assumptions made) for 95 per cent of the time. General expression for the confidence limits for the mean of normal distribution based on a reasonably large sample (n>30) is given by $\bar{x} \pm kse(\bar{x})$. Thus 95 per cent range excludes 5 per cent probability for which k =1.96.

Define symbols

- ***Example***

32. Suppose an estimate of the proportion of persons with a particular ailement is 0.7 and the standard error has been calculated to be 0.02. The 95% confidence interval is given as:

$$0.7 \pm (1.96x0.02), \quad \text{or} \quad 0.7 \pm 0.039$$

This means that the chances are 95 in 100 that the population proportion of persons with the particular ailment is in the range 0.66 to 0.74.

A.2.2 Systematic sampling

33. Systematic sampling is a probability sample selection method in which the sample is obtained by selecting every k^{th} element of the population where ***k*** is an integer greater than 1. The first number of the sample must be selected randomly from within the first ***k*** elements. The selection is done from an ordered list.

34. This is a popular method of selection especially when units are many and are serially numbered from *1 to N*. Suppose that *N* the total number of units is an integral multiple of the required sample size *n* and *k* is an integer, such that *N = nk,* a random number is selected between *1 and k*. Let us suppose 2 is the random Start, then the sample will be of size *n*

35. the units serially numbered as follows:

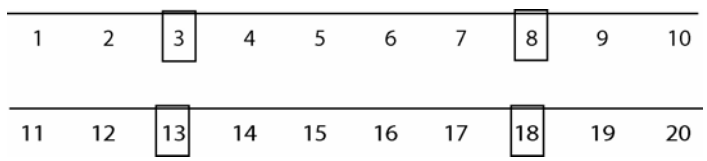
$$2, 2 + k, 2 + 2k, \ldots, 2 + (n - l)K$$

36. It will be observed that the sample comprises of the first unit selected randomly and every k^{th} unit, until the required sample size is obtained. The interval k divides the population into clusters or groups. In this procedure we are selecting one cluster of units with probability *1/k*. Since the first number is drawn at random from 1 to *k,* each unit in the supposedly equal clusters gets the same probability of selection l/k.

A.2.2.1. Linear systematic sampling

37. If N, the total number of units, is a multiple of n, thus if N = n*k*, is the sample size and k is a sampling interval, then the units in each of the possible systematic samples is n. In such a situation the system amounts to categorising the *N* units into *K* samples of n units each and selecting one cluster with probability *l/k*. When *N* = *nk*, $\bar{y}$ is unbiased estimator of the population mean $\bar{Y}$. On the other hand when N is not a multiple of k, the number of units selected using the systematic technique with the sampling interval equal to the integer nearest to N/n may not necessarily be equal to n. Thus when *N* is not equal to *nk* the sample sizes will differ, and the sample mean is a biased estimator of the population mean.

Figure A.1.: Linear systematic sampling

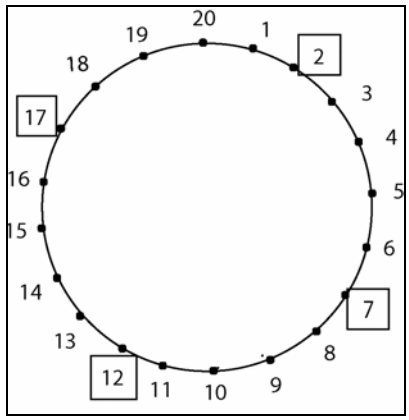


Random start is 3, *N* = 20, *n* = 4, and *K*=5 thus 3, 8, 13 and 18 are selected.

A.2.2.2 Circular systematic sampling

38. We note that in linear systematic sampling the actual sample size can be different from the desired and the sample mean is biased estimator of the population mean when N is not a multiple of n. However, a technique of circular systematic sampling overcomes the above mentioned limitation. In this method of selection you assume the listings to be in a circle such that the last unit is followed by the first. A random start is chosen from I to N. You then add the intervals *k* until exactly n elements are chosen. If you come to the end of the list, you continue from the beginning.

Figure A.2.: Circular Systematic Sampling



N = 20
n = 4, k = 5

Random start is 7
In the above case 7, 12, 17, and 2 are selected.

A. 2.2.3. Estimation in systematic sampling

39. For estimating the total, the sample total is multiplied by the sampling interval.

$$\hat{Y} = k\sum y_i \qquad \text{(A.6)}$$

Estimate of the population mean is

$$\bar{y} = k\frac{\sum y_i}{N} \qquad \text{(A.7)}$$

40. Estimation of variance is intricate in that a rigorous estimate cannot be made from a single systematic sample. A way out is to assume that the numbering of the units is random in such a case a systematic sample can be treated as a random sample. The variance estimate for the mean is therefore given by

$$V(\bar{y}) = \frac{1}{n}\left(1 - \frac{n}{N}\right)\sum s^2 \qquad \text{(A. 8)}$$

where $s^2 = \frac{1}{n-1}\sum (y_i - y)^2$ and y = $\sum y_i$

41. A rigorous estimate of unbiased variance from a systematic sample can be computed by selecting more than one systematic sample from a particular population.

- ***Example***

42. There are 180 primary schools in a county area having an average of 30 or more people under the age of 21 per class. A sample of 30 schools was drawn using systematic sampling with an interval of k = 6.

Number of people under the age of 21 (yj) in the 30 selected villages.

60	200	45	50	40	79	35	41	30	120
300	65	111	120	200	42	51	67	32	40
46	55	250	100	63	90	47	82	31	50

$$\sum y_i = 2{,}542$$

Estimated number of students

$$\hat{Y} = k\sum y_i = 6 \text{ x } 2542 = 15,\ 252$$

Average number of students per farm

$$\bar{y} = k\frac{\sum y_i}{N} = \frac{6(2542)}{180} = 84.7$$

The variance of the sample mean will be calculated on the basis of the assumption that the numbering of Schools is random.

$$V(\bar{y}) = 1\text{-f}\ \frac{s^2{}_y}{n} \qquad \text{(A.9)}$$

$$\text{where } s_y^2 = \frac{1}{n-1}\left\{\sum y_i{}^2 - \frac{\left(\sum y_i\right)^2}{n}\right\}$$

$$= \frac{1}{29}\left(3{,}48700 - 215{,}392.13\right)$$
$$= 4{,}596.8$$

therefore $V(\bar{y}) = (0.833)\,(153.227)$
$= 127.64$

$$Se(\bar{y}) = \sqrt{127.64}$$

= 11.30

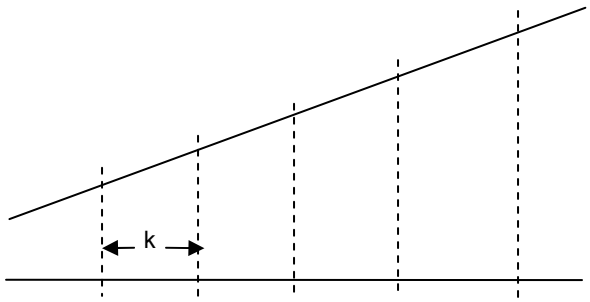
A. 2.2.4 Advantages of using systematic sampling

41. The selection of the first unit determines the entire sample. This aurgurs well for field operations as ultimate sampling units can be selected in the field by enumerators as they list the units.

42. The sample is spread evenly over the population when units in the frame are numbered appropriately. However, the sample estimate will be more precise if there is some kind of trend in the population.

43. Systematic sampling provides implicit stratification;

Figure A.3. Monotonic linear trend

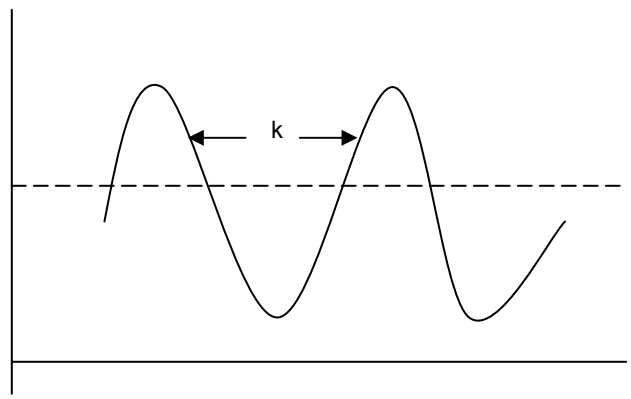


A. 2.2.5. Disadvantages of Systematic Sampling

44. If there is periodic variation in the population, systematic sampling can yield results that are either under-estimates or over-estimates. In such a case, the sampling interval falls in line with the data. For example, if you are studying transport flow for 24 hours on a busy street in a city, if your interval falls on pick hours, therefore you will consistently get high figures and you will not get representative results.

45. The selection method is prone to abuse by some enumerators/ field staff.

Figure A. 4. Periodic fluctuations



46. Strictly speaking, you cannot obtain a rigorous estimate of variance from a single systematic sample.

A.2.3 Stratification

47. Stratified sampling is a method in which the sampling units in the population are divided into groups called strata. Stratification is usually done in such a way that the population is sub-divided into heterogeneous groups which are internally homogeneous. In general when sampling units are homogeneous with respect to the auxiliary variable termed stratification variable, the variability of strata estimators is usually reduced. Further there is considerable flexibility in stratification in the sense that the sampling and estimation procedures can be rightly different from stratum to stratum.

48. In stratified sampling, therefore, we group together units/elements which are more or less similar, so that the variance δ_h^2 within each stratum is small, at the same time it is essential that the means $(\bar{x}_h)$ of the different strata are as different as possible. An appropriate estimate for the population as a whole is obtained by suitably combining stratum-wise estimators of the characteristic under consideration.

A.2.3.1. Advantages of stratified sampling

49. The main advantage of stratified sampling is the possible increase in the precision of estimates and the possibility of using different sampling procedures in different strata. In addition stratification has been found useful in the following situations:

- In case of skewed populations since larger sampling fractions may be necessary for selecting from the few large units, resulting in giving greater weight to few extremely large units for reducing the sampling variability.

- When a survey organization has several field offices in various regions into which the country has been divided for administrative purposes it may be useful to treat the regions as strata, so as to facilitate the organization of fieldwork.
- When estimates are required within specific margins of error, not only for the whole population, but also for certain sub-groups such as provinces, rural or urban, gender, etc. Through stratification such estimates can conveniently be provided.

50. If the sampling frame is available in the form of sub-frames, which may be for regions or specified categories of units, it may be operationally convenient and economical to treat sub-frames as strata for sample selection.

51. Summary of steps followed in stratified sampling:

- The entire population of sampling units is divided into internally homogeneous but externally heterogeneous sub-populations.
- Within each stratum, a separate sample is selected from all sampling units in the stratum.
- From the sample obtained in each stratum, a separate stratum mean (or any other statistic) is computed. The strum means are then properly weighted to form a combined estimate for the population.
- Usually proportionate sampling within strata is used when overall, e.g. national estimates are the objective of the survey and the survey is multipurpose.
- Disproportionate sampling is used when sub-group domains have priority, in cases where estimates for sub-national areas are wanted with equal reliabilities.

Notations

a. Population values
For H strata, total number of elements in each stratum will be denoted by N_1 N_2 _ _ _, N_h _ _ _ N_H such information is usually unknown.

Total population value is $\sum_{h}^{H} N_h = N$ (A.9)

b. $\overline{X}_{hi} = \frac{1}{N}\sum_{i}^{N_h} X_{hi} = \frac{X_h}{N}$ (A.10)

where X_{hi} is the value of the hth element in the hth stratum, X_h is the sum of the hth Stratum.

A. 2.3.2. Weights

52. The weights generally represent the proportions of the population elements in the strata and $W_h = \frac{N_h}{N}$ (A.11)

So, $\sum W_h = 1$

$$S_h^2 = \frac{1}{N-1}\sum_{i=1}^{N_h}\left(X_{hi} - \overline{X}\right)^2 \quad \text{(A.12)}$$

A. 2.3.3. Sample values

a. For H strata, the sample sizes in each stratum can be denoted by $n_1, n_2 ___ n_n$ where $\sum n_h = n$ the total sample size

b. x_{hi} is the sample element *i* in stratum *h*

c. $\bar{x}_h = \frac{1}{n_h}\sum_{i=1}^{n_h} x_{hi}$ (A.13)

d. $\bar{x}_{st} = \sum W_h \bar{x}_h$ (A.14)

e. $f_h = \frac{n_h}{N_h}$ is the sampling fraction for the stratum. (A.15)

Variance of n_h element in the hth stratum

$$v(\bar{x}_h) = \sum\left[1 - \frac{n_h}{N_h}\right]\frac{s_h^2}{n_h} \quad \text{(A.16)}$$

Where s_h^2 is the element variance for the hth stratum and is given by

$$s_h^2 = \frac{\sum\left(x_{hi} - \bar{x}_h\right)^2}{\left(n_h - 1\right)} \quad \text{(A.17)}$$

The variance of sample mean is given by

$$V(\bar{x}_{st}) = \sum W_h^2 (1 - f_h) \frac{s_h^2}{n_h} \tag{A.18}$$

A. 2.3.4. Proportional allocation

53. Proportional allocation in stratified sampling involves the use of a uniform sampling fraction in all strata. This implies that the same proportion of units is selected in each stratum. For example if we decide to select a total sample of 10 percent it means that we shall select 10 percent units from each stratum. Since the sampling rates in all strata are the same, the sample elements selected in the sample will vary from stratum to stratum. Within each stratum the sample size will be proportionate to the number of elements in the stratum.

54. In this case the sampling fraction is given by $f_h = \frac{n_h}{N_h} = \frac{n}{N}$ implying an EPSEM design.

Sample mean

$$\bar{x}_{st} = \sum W_h \overline{x_h} \tag{A.19}$$

Variance of the overall mean is

$$v(\bar{x}_{st}) = \frac{(1 - f)}{n} \sum W_h s_h^2 \tag{A.20}$$

A. 2.3.5. Optimum allocation

55. The method of disproportionate sampling involves the use of different sampling rates in various strata. The aim is to assign sampling rates to the strata in such a way as to obtain the least variance for the overall mean per unit cost.

56. In using this method the sampling rate in a given stratum is proportional to the standard deviation for that stratum. This means that the number of sampling units to be selected from any stratum will depend not only on the total number of elements but also on the standard deviation of the characteristics used as an auxiliary variable.

In optimum allocation, the notion of a cost function is also introduced. For example

$$C = C_o + \sum c_h n_h \tag{A.21}$$

where c_o is the fixed cost.

c_h is the cost of covering the sample in a particular stratum.

57. In many situations we may assume that c_h is a constant in all strata. Therefore, for our purpose we shall consider the Neyman's allocation.
Where c_h is constant and $n = \sum n_h$ the overall sample size which is fixed.

The number of units to be selected within a stratum is given by

$$n_h = \frac{W_h s_h n}{\sum W_h s_h} \quad \text{or} \quad n_h = \frac{N_h s_h . n}{\sum N_h s_h} \tag{A.22}$$

Variance is given by

$$v\left(\bar{x}_{st}\right) = \frac{\left(\sum W_h s_h\right)^2}{n} - \frac{1}{N}\sum W_h s_h^2 \tag{A.23}$$

58. The second term on the right is a finite population correction factor which may be dropped if you are sampling from a very large population, thus if the sampling fraction is small.

General observations

- Population values S_h and C_h are generally not known, therefore estimates can be made from previous or pilot sample surveys.
- Disproportionate allocation is not very efficient for selecting proportions.
- There may be conflicts on variables to optimize in the case of multi-purpose survey.
- In general, optimum allocation results in the least variance.

- ***Example***

The total number of primary schools in a province is 275. A sample of 55 schools is selected and stratified on the basis of number of employees.

Stratum	Number of employees per selected schools (y_{hi})	Total number of schools in each stratum (N_h)	Selected No. of schools by stratum		W_h	s_h^2	s_h	$W_h s_h$	$W_h s_h^2$
			Proportional allocation (n_h)	Optimum allocation (n_h)					
1	2,4,2,2,4, 2,2,4,2,2, 2,2,2,2,5,5	80	16	8	0.2909	1.663	1.289	0.3750	0.48
2	7,7,7,6,8, 7,7,6,7,6 6,8,6,7,8, 6,7,6,6,6	100	20	6	0.3636	0.537	0.733	0.2665	0.19
3	10,12,10,15, 21,16,20,20, 16,19,15	55	11	18	0.2000	15.564	3.945	0.7890	3.11
4	32,35,35,48, 46,47,50,40	40	8	23	0.1455	48.836	6.989	1.0169	7.10
		275	55	55	1.0000			2.4474	10.8

N = Total number of primary schools
n = total number of primary schools in the whole sample
N_h = Size of the hth stratum
n_h = sample size of the hth stratum

A. 2.3.6 Determination of within stratum sample sizes

a. Proportional allocation

$\frac{n}{N} = f$ which is the overall sampling fraction applied to the total number of units in the stratum in our example above, $f = \frac{55}{275} = 0.2$ or 20% for the distribution of sample sizes see column 4 in the table e.g. $n_h = 0.2x80 = 16$

b. Optimum allocation

The formula for obtaining sample sizes for different strata is given by

$$n_h = \frac{W_h s_h}{\sum W_h s_h}(n)$$

for example $n_h = \frac{0.3750}{2.4474} x55 = 8$

The rest of the results are given in column (5) in the table.

Example of how to compute variance based on proportional allocation and optimum allocation

a. Proportional allocation:

$$V(\bar{y}_{prop}) = \frac{1-f}{n} \sum w_h s_h^2 \qquad \text{(A.24)}$$

$$= \frac{(1-0.2)}{55}(10.8) = .0158$$

b. Optimum allocation:

$$V(\bar{y}_{opt}) = \frac{\left(\sum w_h s_h\right)^2}{n} - \frac{1}{N} \sum w_h s_h^2 \qquad \text{(A.25)}$$

$$= \frac{(2.4474)^2}{55} - \frac{10.8}{275} = 0.0693$$

c. In general

$$v(\bar{x}_{st})OP \le v(\bar{x}_{st})PROP \le v(\bar{x}_{st}) \le (\bar{x}_{st})SRS \qquad \text{(A.26)}$$

A.2.4 Cluster sampling

59. The discussions in the previous sections have so far been about sampling methods in which elementary sampling units were considered as arranged in a list from a frame in such a way that individual units could be selected directly from a frame.

60. In Cluster Sampling, the higher units e.g. enumeration areas (see chapter 2) of selection contain more than one elementary unit. In this case, the sampling unit is the cluster.

61. For example, to select a random sample of households in a city a simple method is to have a list of all households. This may not be possible as in practice there may be no complete frame of all households in the city. In order to go round this problem, clusters in the form of blocks could be formed. Then a sample of blocks could be selected, subsequently a list of households in the selected blocks made. If need be, in each block a sample of households say 10% could be drawn.

A.2.4.1. Some reasons for using cluster sampling

a. Clustering reduces travel and other costs of data collection.
b. It can improve supervision, control, follow-up coverage and other aspects that have an impact on the quality of data being collected.
c. The construction of the frame is made cheaper as it is done in stages. For instance in multi-stage sampling discussed in chapter 2 a frame covering the entire population is required only for selecting PSUs i.e. clusters at the first stage. At any lower stage, a frame is required only within the units selected at the preceding stage. In addition, frames of larger and higher stage units tend to be more durable and therefore usable over longer period of time. Lists of small units such as households and particularly of people tend to become obsolete within a short period of time.
d. There is administrative convenience in the implementation of the survey.

62. In general we should note that in comparing a cluster sample with an element sample of the same size, we shall find that in cluster sampling the cost per element is lower owing to lower cost of listing and/or locating of elements. On the other hand, the element variance is higher due to irregular homogeneity of elements (intra-class correlation) in the clusters. We illustrate the basic cluster sampling by considering a single stage design (multi-stage designs are presented in chapter 2).

A.2.4.2. Single stage cluster sampling

63. In a particular district, it may not be feasible to obtain a list of all households, and then select a sample from it. However, it may be possible to find a list of villages prepared during a previous survey or kept for administrative purposes. In this case we would obtain a sample of villages, then obtain information about all the households in the selected villages. This is a single-stage cluster sampling design because after a sample of villages has been selected all units in the cluster, in this case households, are canvassed.

64. Sample selection under clustering can be illustrated as follows:

Assume that from a population of EAs (clusters) a sample is selected with equal probability. For a single stage cluster sampling, all households from the EAs would be included in the sample.

Given that

$A =$ Total number of clusters
$B =$ Total number of households in the cluster
$a =$ A sample of clusters

i.e. $aB =$ n elementary units in the total sample
$AB = N$

65. The probability of selecting an element with equal probability is given by

$$\frac{a}{A} \; x \; \frac{B}{B} = \frac{n}{N} = f \qquad \text{(A.27)}$$

where N is the number of elementary units and f is the sampling fraction. In this case the probability of selection is simply $\frac{a}{A}$.

A.2.4.3. Sample mean and variance

$$\bar{y} = \frac{1}{aB} \sum_{\alpha=1}^{a} \sum_{\beta=1}^{B} \bar{y}_{\alpha\beta} = \frac{1}{a} \sum_{\alpha=1}^{a} \bar{y}_{\alpha} \qquad \text{(A.28)}$$

The sample mean is unbiased estimate of the population mean:

$$\text{E}(\bar{y}) = \frac{1}{A} \sum_{\alpha=1}^{a} \bar{y}_{\alpha} = \bar{Y} \qquad \text{(A.29)}$$

66. In fact because the sample size is fixed ($aB = n$) and the selection is of equal probability then the mean ($\bar{y}$) is unbiased estimate of the population mean $\bar{Y}$.

67. If the clusters are selected using a simple random selection the variance can be estimated as follows:

$$V(\bar{y}) = (l\text{-}f) \; s_{\alpha}^{2} \qquad \text{(A.30)}$$

where $s_{\alpha}^{2} = \frac{1}{a-1} \sum_{\alpha=1}^{a} (\bar{y}_{\alpha} - \bar{y})^{2}$

Appendix

List of experts: United Nations Expert Group Meeting to Review Draft Handbook on Designing Household Sample Surveys, New York, 3-5 December 2003 (see report ESA/STAT/AC.93/L.4).

List of experts

Name	Title and affiliation
Oladejo oyeleke Ajayi	Statistical Consultant, Nigeria
Edwin St. Catherine	Director, National Statistical Office, St. Lucia
Beverly Carlson	Division of Production, Productivity management, United Nations Economic Commission for Latin America and the Caribbean, Santiago, Chile
Samir Farid	Statistical Consultant, Egypt
Maphion M. Jambwa	Technical Adviser, SADC/EU, Gaborone, Botswana
Mr. Udaya Shankar Mishra	Associate Fellow, Harvard University, Boston, USA
Jan Kordos	Professor, Warsaw School of Economics, Warsaw, Poland
Anthony Turner	Sampling Consultant, U.S.A.
Ibrahim Yansaneh	Deputy Chief of Cost of Living Division, International Civil Service Commission, United Nations, New York
Shyam Upadhyaya	Director, Integrated Statistical Services (INSTAT), Nepal