UNITED NATIONS SECRETARIAT **ESA/STAT/AC.93/8**
Statistics Division **03 November 2003**

English only

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

D R A F T

Data processing, analysis, and dissemination*

by

Maphion Mungofa Jambwa**

* This document is being issued without formal editing.

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

Table of contents

Chapter Eight: Data processing, analysis, and dissemination 3
8.1. Introduction 3
8.2. A visual summary of the phases and processes of a household survey 3
8.3. Survey planning 4
8.3.1. Survey objectives 4
8.3.2. The tabulation and analysis plan 5
8.3.3. Design of the survey questionnaire 6
8.3.4. The conceptualization and general considerations for systems design for household surveys 7
8.3.5. The options for configuring data processing systems for household surveys 12
8.3.6. The development and uses of databases for household surveys 16
8.4. Survey operations 20
8.4.1. Data collection and data management 20
8.4.2. Data preparation 20
8.4.3. File structure and datasets for tabulation and analysis 29
8.4.4. Estimation and assessment of the accuracy of estimates 33
8.4.5. The analysis of survey data 38
8.4.6. The reporting, presentation and dissemination of survey data 51
Annex 1. Software options for different steps of survey data processing 58
Annex 2. Documentation Structure for a survey and its production system 59

Chapter Eight: Data processing, analysis, and dissemination

8.1. Introduction

1. Information technology (IT) has developed rapidly during the last two decades or so. Its development has, in turn, impacted significantly on the techniques for designing and implementing survey processing systems.

2. The main development in hardware has been the shift from mainframe systems to Personal Computer (PC) platforms. The PC has become increasingly powerful both in terms of processing speed and storage capacity. PCs can now perform all kinds of processing, ranging from small-scale surveys to large-scale statistical operations such as population censuses.

3. Parallel to the developments in hardware have been the significant improvements in the quality and user friendliness 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.

4. 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. Annex 1 may serve as a rule of thumb for choosing software for the different steps of survey data processing.

5. Having the appropriate hardware and software is necessary but no guarantee for the successful processing of survey data. How the processing system is conceptualized and implemented is also critical for the timely delivery of the statistical products.

8.2. A visual summary of the phases and processes of a household survey

6. In principle all surveys run through the same kind of cycle and the typical phases are as follows:

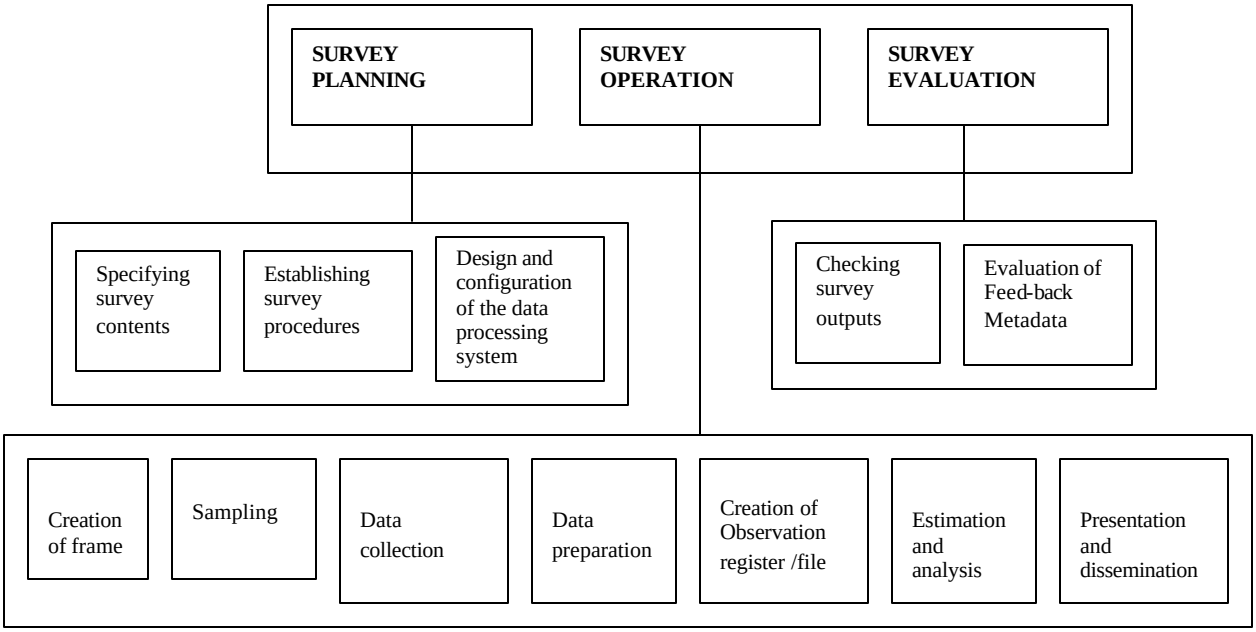
Survey planning: the designers of the survey make decisions about the major purposes, users and uses of the survey, its major outputs and major inputs, procedures for obtaining the inputs (the design and preparation of the questionnaire and related survey instruments) and transforming them into outputs, and the design of the data processing and documentation system.

Survey operations: this consists of the creation of the sampling frame, sampling, data collection (measurement), data preparation (data entry, coding, editing and imputation), and the creation of the observation file, estimation, analysis and presentation and dissemination of results.

Survey evaluation: consisting 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. It should, however, be noted that there is a tendency of those who design surveys to be either preoccupied with component parts, loosing sight of the overall system or they execute the survey in an undesirable sequence of operations. The desired sequencing is that: survey objectives should determine the output design (e.g. the tabulation plan). 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. A theme to be emphasized in this chapter is the need to follow through and document all these processes in a clear, coordinated, and comprehensive manner.



Source: Sundgren (1999).

8.3. Survey planning

8.3.1. Survey objectives

9. The first step in the design of any survey should be the articulation, agreement on, 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 on questions about the target population. The respective survey questionnaire should, therefore, provide the data that can answer such questions.

10. However, the data required to serve the survey objectives needs to be meticulously determined, given the limited resources normally availed to the national statistics agency in a developing country to undertake such surveys. The process should begin with a set of questions for which the stakeholders for the survey would like to have answers. Typically, there are four types of such questions.

11. One set of questions is that which seeks to establish the fundamental characteristics (the status) of the population under study (the proportion of the population that is poor; the rate of unemployment; etc.).

12. Another set tries to link household characteristics with government policies and programs in order to examine the coverage (impact) of such programs. For example: the proportion of households participating in a particular program, and how their characteristics compare to those of households not participating in the program.

13. A third set of questions concerns changes (impact) in households' characteristics over time. Governments and other agencies often want to know whether the living conditions of households are improving or deteriorating.

14. Then, there are questions about determinants (i.e. seeking to establish causal relationships) of household circumstances and characteristics. These are questions on what is happening and why it is happening. They seek to understand the impact of current policies or programs.

15. Once the set of questions to be answered has been agreed upon, they can be expressed as objectives of the survey. For example, a question about the current rate of unemployment implies that one objective is to measure unemployment. The next step is to rank the objectives in order of importance. If the number of objectives is too large, it may be necessary to drop those with low priority ratings. Overloading the survey data collection phase can be very risky. For example it can compromise the quality from the field due to fatigue endured by respondents during data collection interviews. It can also lead to unnecessary complications for the data processing step and may even lead to no output being delivered.

8.3.2. The tabulation and analysis plan

16. 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.

17. 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.

18. 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 data will prompt new tables.

19. 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. The analysis plan should also be the main reference point to guide the analysis of the survey results.

8.3.3. Design of the survey questionnaire

20. 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.

21. The size and format of the questionnaire need very serious consideration. Indeed, it is rather tragic that household surveys, tend to fail because of the tendency of overloading the data collection phase of the survey. Often there is so much disparity between the amount of data collected and the data products of the respective survey. It is often difficult to resist the temptation and pressure to include many questions in the questionnaire. However, the cost and risk of complicating the data processing need to be very carefully considered. In some instances the survey results have never come out because of this factor. Some tough prioritization process has, therefore, to be adopted, and this is also where the tabulation plan can be an important platform.

22. It is important to analyze whether the information recorded on the questionnaire can be processed easily. An identification code should uniquely identify each questionnaire and should always be numerical. It should distinguish between different questionnaires, information for assignment of expansion factors (strata, primary sampling units, area segments, distinction between administrative areas needed for tabulation, etc.). From the

point of data processing, the ease with which questions in the questionnaire are amenable for processing is critical. For example, it is easier to handle pre-coded versus post-coded questions.

23. Survey data management begins concurrently with questionnaire design. It is important to ensure that there are no flaws on e.g. the definition of observation units, skip patterns, etc., right from that stage. Every household survey collects information about a major statistical unit **(the basic object**) – the household – as well as about a variety of subordinate units (**associated objects**) within the household – persons, budget items, plots, 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 is by means of a simple serial number written or stamped on the cover page of the questionnaire, preprinted by the print shop. This and the rest of the data on the cover page (geographic location, urban/rural status, sampling codes) usually become important attributes of the household, included in the survey datasets.

24. The layout of the questionnaire is very important because it affects the layout of the input screens (forms). The screen should, as far as possible, have the same layout as the questionnaire. This will make it easier for the persons entering the questionnaire. This issue is taken up later in the discussion of data entry screens.

8.3.4. The conceptualization and general considerations for systems design for household surveys

25. The process of developing a survey processing system is intertwined to the underlying survey phases described above. The rest of the chapter discusses some of these steps and issues in as far as they impact on the processing of household surveys within the environment of developing countries.

8.3.4.1. General considerations for systems design for household surveys

26. One of the first major activities to be undertaken when planning a census or household survey should be the system design. During this step the survey data to be collected and generally speaking the whole data processing system are specified according to some formalized scheme.

27. The design of the processing system for a statistical survey should be made and involve close co-operation between the statisticians, subject matter experts, and systems analysts/programmers. Such co-operation would be strengthened and facilitated by the use of a common model (scheme) for the design, development, and documentation of all systems within a statistical agency. At least, a common approach for the processing of household surveys is desirable, if such an approach is not yet established within the agency.

28. One of the greatest requirements and benefits from a systems design model such as described above is that all survey operations would be explicitly described and written down and could be referenced at a latter stage. The resulting documentation (i.e. set of metadata) is

important both for the development and maintenance of the respective statistical production systems. It is also a very important quality aspect, for example, through enabling easier access to the survey results. One or more final data files will be produced for every survey processing system. The files should be documented such that even 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. Annex 2 provides an example of such a template.

29. An additional benefit, from the documentation cited above, emanates from the fact that the costs for systems development and maintenance are always quite high in a statistical agency, where there are many different systems. Such costs tend to go down with increased possibilities of combining and using micro-data files from different surveys for purposes other than the original objectives of the respective systems even long after the datasets were produced. The structure of household surveys (save for the content) 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. As indicated earlier, the template in Annex 2 can be an invaluable tool for facilitating such a beneficiation process. The adoption of this kind of 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.

30. 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 the various surveys if they have the same meaning. This is particularly important for variables that are used to identify the records (the objects) within the file, as these variables may also be used when combining (joining) data from different systems.

8.3.4.2. A systems design model for household surveys

31. It should be clear enough, from the above, that there are significant benefits to be reaped from a policy of uniform systems design within a statistical agency. The method used should, of course, be based on modern Information Systems theory and should support all kinds of statistical production systems, improve the data analysis, and facilitate systems documentation.

32. Systems design methods are commercially available as concepts or software products (e.g. data modelling tools, Computer Aided Systems Engineering tools, etc.). Most of these use bases that are usually variants of the 'entity-relationship-attribute' (ERA) model. The latter is also sometimes labelled as the 'Object-Property-Relationship (OPR)' approach. The models are usually very general, in order to support any type of business. Models or products that specifically support the type of systems design issues that face a statistical agency are not readily found on the open market. As a result, several statistical agencies around the world have developed their own methods, or adapted general methods for their statistical production systems.

33. Statistics Sweden uses an in-house developed systems design method. The method is based on the ERA model, adapted for statistical systems. It has also been used by agencies that have received technical support from Statistics Sweden such as the national statistics offices in Laos, Lesotho, Namibia, South Africa, Tanzania, and Zimbabwe. It has been used, for example, for the design and documentation of systems for household surveys, population and housing censuses, business registers, etc.

34. The method is based on three concepts: objects, properties, and relations. It strictly separates the so-called infological phase (i.e. the subject matter analysis, output/table analysis, object analysis, etc.) from the so-called datalogical phase (the database design, file design, application design, etc.). It stresses the statistics design and documentation. The infological phase can be said to be contents-oriented while the datalogical phase is technique-oriented.

35. During the infological phase the contents and structure of the planned information system is specified in terms of the notions: objects, relations and variables. It is user-oriented and requires very close co-operation and collaboration between the subject-matter statisticians and the data processing specialists. It is mainly concerned with the contents and purposes of the system, i.e. in answering the questions WHAT and WHY and not the technical aspects of the data processing system.

36. During the datalogical phase the resulting infological model is systematically transformed into a model of the data files and data processes. The production system is modeled and the main concern in this step is HOW. The transformation is done in such a way that all files, eventually, become flat files, so that the processes can be optimally performed using generalized software. It also includes the specification of the file structure and record layouts, the updating operations, retrieval and tabulation processes, etc.

37. A number of reports, for example, those by Sundgren (1984 and 1986), and Jambwa (1989) provide some more details of the model. The following provides an indication of some of its general and fundamental concepts.

38. *Data Structure:* The key notions used to describe the structure of the data are: basic objects, relations, and variables.

39. An **object** is any concrete or abstract entity (physical object, living creature, organization, event, etc.) that the users may want to have information about. **Objects** for the particular household survey are items (elements or units) that the users would like to have information on (e.g. household, person, etc). For most household surveys the basic object is the HOUSEHOLD. There may be several associated objects related to the basic object, and these will depend on the particular surveys. The example below shows the object system formulated for the 1987 Zimbabwe Intercensal Demographic Survey (ZICDS).

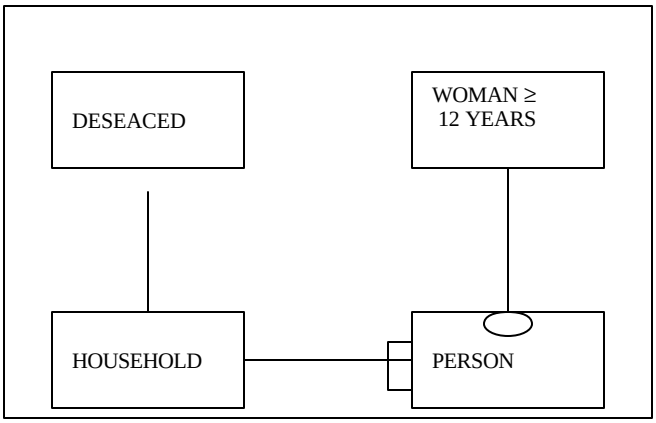
40. The objects associated with HOUSEHOLD were: PERSON, WOMAN 12 YEARS OR ABOVE, and DECEASED.

41. In a typical agricultural survey the objects associated with HOUSEHOLD are: PERSON, and FIELD *(representing plots of land operated by the household)* while in the typical Household Budget Survey HOUSEHOLD is associated with: PERSON, INCOME, *(and items)* FOOD, GOODS, DURABLES, ETC.

42. For every object, there will be several variables of interest. **Variables** being **properties** (attributes or characteristics) of the objects e.g. the object PERSON can have age, income, occupation, marital status, etc., as variables. Variables may be qualitative or quantitative.

Object	Identifying Variables	Object Definition	Important Variables	Related Objects	
				Object	Foreign key
HOUSEHOLD	HID (AREA, DIVISION, SUBDIV, EANR, HHNR)	A house is a group of persons who normally live and eat together, and excludes visitors.	SOH (size of household) – derived variable. STARTUM AREA	PERSON DEACASED	HID HID
PERSON	HID, PID	The person is a usual member of the household or a visitor last night	SEX, AGE, MARSTAT, ETHNIC, USEM, RELTH, ETC.	HOUSEHOLD WOMAN ≥12 YEARS OLD	HID HID, PID
DECEASED	HID, DID	The deceased who was a usual member of the household during the last 12 months.	SEXD AGED	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

43. The logical link between an object and a variable is called an association. Objects may also be linked to other objects. Such links are called object **relations**. The key relations between objects are visualized in the so-called object graph of the model. Two or more objects would be related to one another in a certain way (e.g. two persons may be married to each other; and one person may be employed by a company; etc.). The following shows an object graph formulated for the 1987 ZICDS.



44. The relation between objects may be a one-to-one (**example**), one-to-many (e.g. one household comprising of two or more persons), or many-to-many (**example**). One object may also be subset of another (e.g. WOMAN ≥12 YEARS is a subset of the object PERSON).

45. Every object should also have a unique identification. The identification of an associated object indicates the basic object it relates to, e.g. PERSON would be related to HOUSEHOLD and would be identified by the combination of household id and the person serial number (within the household roster).

46. *Input to the data processing system:* The input consists of values observed and measured by enumerators according to the survey questionnaire, and the enumeration in the case of household surveys, is household based.

47. *Output of the data processing system:* The output of the system consists mainly of statistical tables (based on some tabulation plan), databases containing micro and macro data, etc., and 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.

48. *File organization:* 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.

49. *System flow chart:* A reasonably detailed flow chart should be set up for each survey. The chart is important for many reasons, one being as an instrument for making time schedules and estimation of human resources needed to complete the processing of the survey. Typically, the main activities in data processing for any survey, would include:

a. Data checking, editing, coding.
b. Data entry and verification.
c. Transformation of the data structure used at the input stage to a data structure suitable for tabulations.
d. Tabulation.

50. The systems flow chart would also include the fundamental file operations like: selection, projection, sorting and matching of files, derivation of new variables, aggregation, tabulation, and graphic presentation.

51. Experience in applying the above systems design model, in the various countries, shows that most of the household surveys within the typical national household survey program can be fitted and effectively implemented through this common approach. However, the details must be separately developed for each particular survey. The documentation template in Annex 2 provides a guide on the steps and procedures that need to be followed when designing the production and metadata for each particular system. The report by Rosen and Sundgren (1991) provides a detailed description of the template. The template provides a good and comprehensive checklist of the steps and procedures necessary for the efficient development and documentation of data processing systems for statistical surveys. The subsequent sections of the chapter present some of these items from the point of view of current and best practices vis-à-vis the design and implementation of such systems.

52. As noted, earlier, there are several options for systems design. The model described above represents only one such approach. It can also be stated that the approach followed in the SDA Integrated Survey (1991, Chapter 6), more or less goes along similar lines. This can be seen on the sections of the chapter that discuss the *Data Model and the Data Structures.*

8.3.5. The options for configuring data processing systems for household surveys

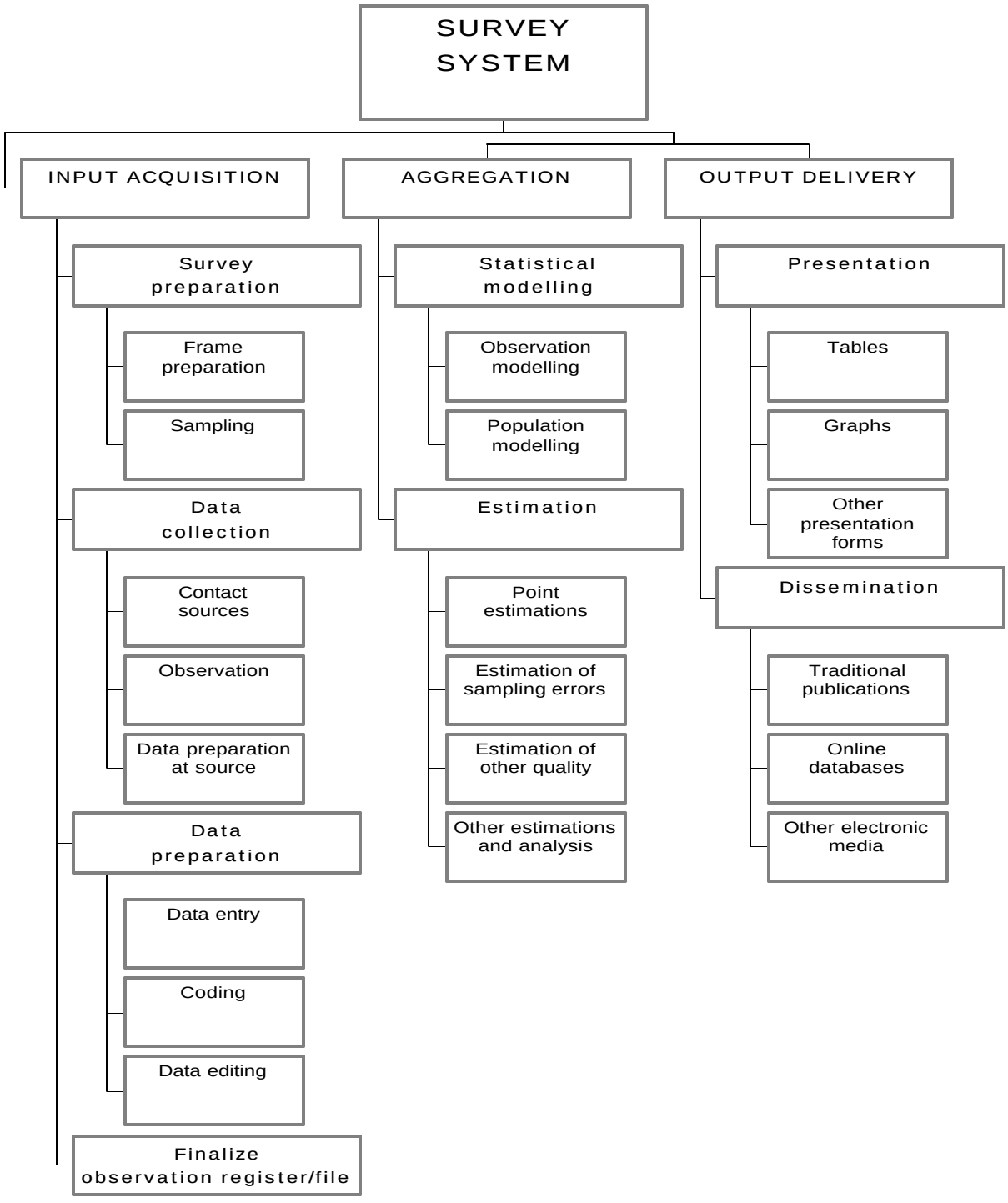
53. How the actual data processing system is designed is to some extent a function of the set up of the available data processing infrastructure. The set up determines the options whether the conventional or database oriented processing system or a hybrid of both can be adopted. This is the focus of the following discussion.

54. According to Sundgren (1995), there are three major functions entailed in a survey processing system:

a. An **input acquisition function**, which directly or indirectly observes (measures) certain object characteristics, and which prepares and stores the observation data obtained as microdata in an observation register (file).
b. An **aggregation function**, which transforms the microdata produced by the input acquisition function into macrodata, or 'statistics', which are estimated values of statistical characteristics.
c. An **output delivery function**, which makes macrodata (statistics) available to users, and which assists the users to interpret and analyze the data further.

8.3.5.1. The conventional survey data processing system

55. The following diagram illustrates a breakdown of the three functions into more concrete sub-functions and tasks.

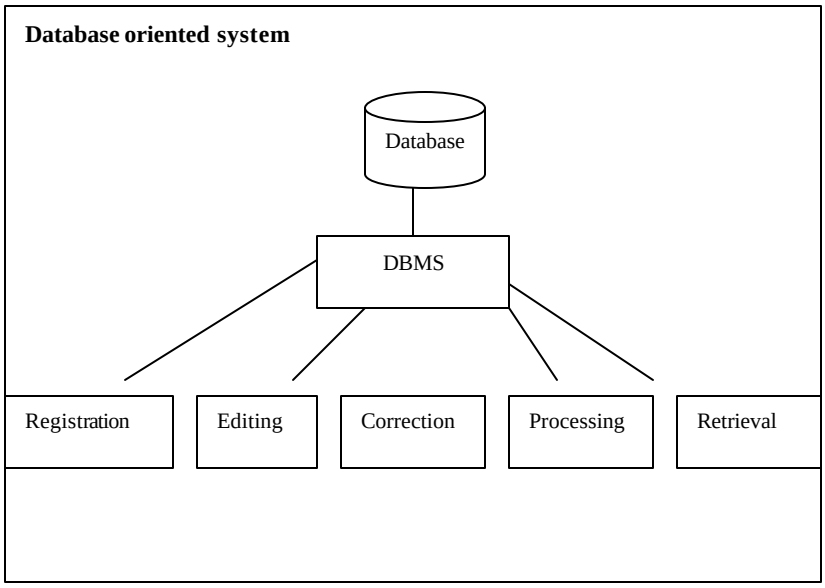


Source: Sundgren (1995)

56. In a traditional survey processing system, the functions, sub functions, and tasks are carried out more or less serially, and from top to bottom and from left to right.

8.3.5.2. The database oriented data processing

57. Modern technology permits a much more flexible organization of the processes for producing and disseminating statistics. This can be achieved through a database -oriented system, illustrated below.



Source: Sundgren (1985)

58. The microdata and macrodata, which are stored and processed, are communicated within and between the functions via a database and the database management system. The datasets are described by accompanying metadata, also stored within the database. The metadata are consistently updated, whenever the described data are transformed.

59. Databases can be used to handle different types of systems: survey processing systems; register management systems; user-driven retrieval systems; etc.

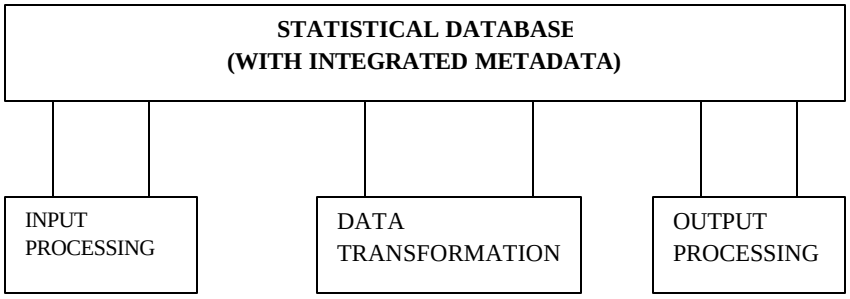
60. A survey processing system focuses on data collection and processing, resulting in a collection of microdata, which are aggregated into estimated values of certain statistical characteristics.

61. A user-driven retrieval system focuses on the dissemination of available macrodata and microdata from different surveys, which maybe relevant for the various categories of users. This is discussed later under the section on dissemination.

62. Regarding register management there are two kinds of registers: base registers, which establish and maintain an authorized list of objects (e.g. a register of population census enumeration areas or a business register) belonging to a certain population and; a code register, which establishes and maintains an authorized list of values belonging to the value set of certain variable or classification (e.g. some kind of look-up table.).

8.3.5.3. Subsystems for a database-oriented data processing system

63. According to Sundgren (1989) the subsystems for a database-oriented data processing system include the following:



- The database reflects and contains data about the object system of the survey.
- Both microdata and macrodata may appear in the physical database.
- The input subsystem contains functions for updating the database, and capturing, coding, and editing the input data.
- The output subsystem contains functions for retrieving and processing statistical information to end-users, and for initiating and presenting results of statistical analysis.
- The transformation subsystem contains functions for transforming data in the database e.g. between micro and macro data (aggregation); also functions for carrying out algorithms for statistical analysis.

8.3.5.4. The conventional versus database oriented production systems

64. Using the database technique in a data processing system implies: centralized file management, and data independence. In a conventional system, there is decentralized file management; every subsystem takes care of its own file management, although usually more or less the same data are processed in the different subsystems. In a database oriented system the file management system is concentrated in the database management (sub)-system, the DBMS. The DBMS software controls the operations of this subsystem.

65. In a conventional system, the different subsystems (steps) must be executed in a certain (sequential) order. In a database-oriented system, all subsystems operate against one and the same database, and the data may be physically distributed over several different files.

66. In the conventional system data are usually stored as physically integrated parts of the application software or as separate files or databases. Such systems, in general, have very low degrees of data independence. Even small changes in the data structure tend to entail changes to be made in several of the application programs.

67. The greater the extent to which changes can be made to the content and organization of the data, without having to rewrite the application program, the higher the degree of independence. Such relatively high degree of data independence is associated more with database-oriented systems under which it is the task of the DBMS to manage and deliver the data to the application programs in a standardized way.

68. The four basic functions of a database management system are to add, retrieve, update, or delete specified data. In order to be able to specify the data upon which such operations should be performed, there must be a data model describing the database, and the database management system must be able to interact with the database in terms of this data model.

69. Currently, it can be stated that the relational data model is the de facto standard for a wide range of commercial database management systems and database related software products. The Structured Query Language (SQL) is equally the widely accepted interface between the relational database management systems and database related software products.

8.3.6. The development and uses of databases for household surveys

70. In general there has been limited co-ordination of different data collection and data dissemination activities in developing countries, especially in Africa. Often, the different surveys that are published do not provide a comprehensive overview of society. Information that is required for planned development and the methodology for collecting, processing, and analysing the relevant data have been established for decades. What seems to be lacking is a technique that adequately integrates the several major dimensions of a national information system to support stakeholders in the various spheres of planning.

71. The goal of a database approach is to strengthen the coordination of data collection and dissemination. The underlying philosophy being that data should be collected and organized in such a way as to facilitate:

a. A comprehensive overview of socio-economic phenomena by allowing logical integration i.e. it should be possible to relate information from different surveys (and other sources) concerning the same phenomena.
b. Rapid access to information by allowing physical integration i.e. it should be easy to get an overview of the catalogue etc of the contents of the database and the information itself.

72. As discussed earlier, databases can be used to handle different types of systems: survey processing systems; register management systems; user-driven retrieval systems; etc.

The focus here is on how database techniques can be applied on household survey data. The three different types of system still apply.

73. One area is the data entry and editing phase of a household survey where there is a need to utilize database techniques to deal with embedded complex situations. Such complexity arises from the complex data structure often entailed in household surveys. One questionnaire can contain data from many different object types for example from the household, individuals, etc. The use of database techniques can enable the systems designer to have a simple file structure and at the same time be able to access different object types related to each other e.g. the household data together with the data from individuals belonging to the household.

74. Database techniques are also important when it comes to keeping an archive of clean micro-data files (data that has gone through the data entry and editing phase but has not been aggregated into statistical tables). The important aspect here is the integration of data descriptions (metadata) into the system and the existence of procedures (software and manuals routines) to facilitate ad-hoc retrievals from the stored data. In this situation the ambition is to facilitate easy and rapid overview and access described in a. above.

75. A third area for database techniques is to keep an archive of macro level data (i.e. statistical tables or time series). The most important areas, which have emerged here, are time series databases for storage and retrieval of economic time series and regional statistical databases for storage of data on regions, villages, wards, etc. Both the goals mentioned in a. and b. above are relevant in this regard.

76. Household survey programs represent a window where database techniques and the underlying benefits can be applied and reaped. They especially represent an opportunity where the full potential for an integrated approach can be used. However, some fundamental issues for such integration to be achieved need be sorted out at the outset.

8.3.6.1. Logical integration

77. There are two important aspects of logical integration that should be stressed. One is the level of detail for integration; the other is the coordination of the definitions of basic information units.

Level of detail

78. It has to be decided at which level longitudinal studies and cross-sectional studies are to be integrated. Should it be possible to study changes over time (or compare information from different surveys covering different aspects) on national, provincial, district or even household level? The level chosen will affect the analysis possible to undertake. There is a trade-off between level of detail and the complexity of both the survey design and the physical database design.

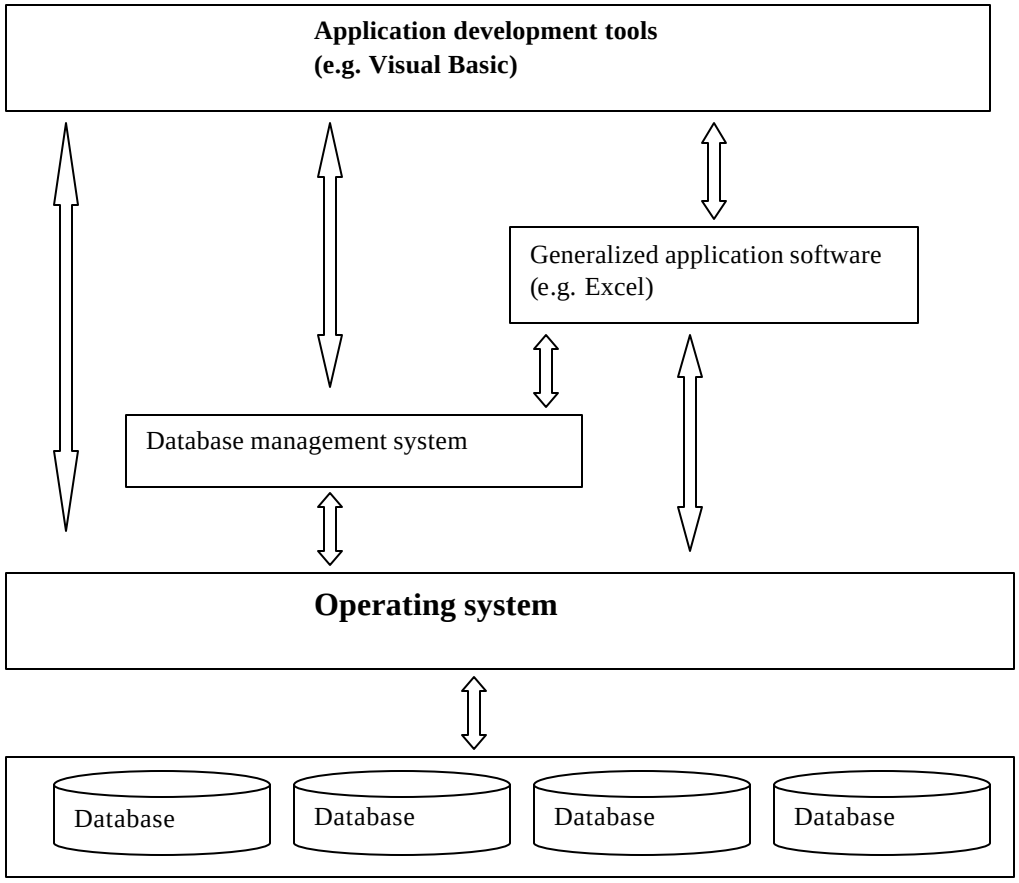
79. If longitudinal or cross-section studies on the household level are to be conducted, the same households have to be included in the samples of the different surveys. If the district level is chosen, the same districts have to be included in the area sample (but not necessarily the same households in the district).

80. The technical database design has to allow for easy access to the information at the level of detail chosen. If this level is very detailed the volume of information in the database will grow and hence also the complexity of the access mechanisms if any analysis is to be possible.

Coordination of definitions and classifications

81. The integration of information from different sources requires coordination of the definitions of the units of observation or measurement/classification. For example, the definition of household used in one survey should be identical with the household-definition used in another survey. The classifications of the units used in the different surveys also have to be compatible. For example the characteristics used to classify education in one survey should be comparable with the classification of education in another survey.

8.3.6.2. Physical integration



Source: Sundgren (1999)

82. The picture above shows the different pieces for a database system. It summarizes the interaction between the different kinds of software components, as well as between components and hardware component. Each interaction requires a well-defined interface.

83. The database should function as a clearing-house for statistical data used for socio-economic planning. Planners within the government and the other sectors of the economy as well as the research community should have access to the database.

84. Technically there should be interfaces to the different providers of information. Special attention has to be paid to the technical integration of the IT-systems used at the NSO and the database system. A unified structure of the IT-systems for the household surveys should be aimed for. The systems should interface to the database system in such a way as to facilitate a high degree of automation in the updating of the socio-economic database.

85. In the design of the database retrievals of different types should be allowed for. For some planning purposes the requirement is to retrieve time-series, analyze them or utilize them in econometric models. For other planning purposes it is also necessary to make aggregations and retrievals along other dimensions than time (for example by the geographical dimension).

86. The database system should be designed to store all kinds of information of statistical character, i.e. both tabular and time series oriented information.

87. In the database system there should be catalogue functions enabling the user to examine which information is stored. The system should be end—user oriented in the sense that it should be possible to use the system without any prior EDP-training, after a short course.

8.3.6.3. Technical requirements for the development of the database system

88. A prerequisite for setting up a database is that sufficient data processing capacity will be available. This includes hardware infrastructure for the database system. Many of the users-to-be of the database system should have PC's or terminals connected to the system. Additional disc capacity might have to be acquired when the contents of the database gradually builds up.

89. Regarding software development the strategy involves: the choice of appropriate DBMS software, Generalized software (e.g. Excel), and Application software (e.g. Visual Basic). The development should also be done step-by-step and linked to the development of the contents of the database, the training of the database management staff and the training of the users.

90. In the software development the following major steps can be distinguished: basic database design; design of retrieval functions for time series; design of updating functions; design of simple functions for analysis of time series data; design of interfaces to packages

for analysis, modeling and graphical presentation; design of tabular retrieval and presentation functions for cross-section analysis.

8.4. Survey operations

91. This phase follows from the specification of the substantive objectives including the type of information required, the population to be studied, the population domains for which separate estimates will be produced, the timing and precision requirements, the survey structure and sample design, questionnaire development, systems design, etc.

92. Its main sub-processes can be seen as data collection, data preparation, file organisation, estimation and analysis of the data, presentation and dissemination of the survey results.

8.4.1. Data collection and data management

93. Data collection can be viewed as starting with the recruitment, training, and the deployment of field staff for the enumeration process. Apart from the survey materials, the availability of adequate transport and communications facilities for the field staff is one critical factor for the success at this stage. Effective supervision and quality assurance measures are also important, if the survey is to produce the desired products. A significant part of the nonsampling errors can be minimized through the design and implementation of a properly thought out scheme for quality assurance.

94. 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 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.

8.4.2. Data preparation

95. 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, e.g. 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.

8.4.2.1. Configurations for data preparation

96. According to Munoz (2003), 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

97. This is the only option that existed prior to the advent of PCs. It, largely, remains the main approach used for surveys in developing countries, with some modification due to the introduction of microcomputers. It is also the approach assumed in the subsequent sections of the chapter, where coding and editing are discussed. 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 at the headquarters of the national statistics offices or in its regional offices. The two main techniques used for this approach are:

Interactive data entry

98. This involves keying of coded data onto disk, diskette, or CD. 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 PCs and relevant software.

Scanning

99. This involves the reading of the data directly by optically scanning the questionnaires. During scanning, the scanner creates an image of each questionnaire. The scanning software subsequently evaluates the scanned images and questionnaires with possible errors are subject to verification by the scanner operator. Typical errors include unidentified forms that cannot be evaluated, questionnaires with missing or mismatched pages, unrecognizable hand printed characters, etc. Scanning eliminates the need for operator controlled data entry, saving time and reducing errors and staff costs. However, the use of scanning requires expensive equipment including maintenance and servicing provisions, precision in the design and printing of questionnaires, and careful handling of the documents in the field. This, perhaps, explains why scanning has not been so prevalent despite the perceived gains from its adoption. Experience in the use of scanning for household surveys has been generally very limited especially in the sub-Saharan region. Recently, it has been used for the Core Welfare Indicator Questionnaire, driven by the World Bank.

Data preparation in the field

100. 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.

101. One form which this can take is having the data entry operator work with a desktop computer in a fixed location (e.g. in the regional office of the 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 from households for the questions where errors, omissions or inconsistencies are detected by the data entry program.

102. 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 till all the data are entered and is qualified as complete and correct by the data entry program.

103. 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 the more scope for standardizing the data collection by the interviewers.

104. 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

105. The use of palmtop computers to get rid of the paper questionnaires altogether is very appealing. 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. Computer-based interviewing requires well-structured interviews, with a beginning and an end. However, most surveys in developing countries require multiple visits to each household, separate interview to each member of the household, etc., in a process that is not strictly structured but rather intrinsically driven by the interviewer.

8.4.2.2. Coding and editing of survey data

106. Data checking, editing, and coding represent, probably, the most difficult phase of data processing. It is the organization of data management and data preparation that newly trained survey professionals often encounter great difficulties.

Coding

107. 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 (e.g. 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.

108. A manual should be written to give explicit guidance to the coders. Such a manual should contain a set of disjoint categories, which among them cover all acceptable responses to the questions under consideration. For a large-scale survey, it is desirable to maximize the extent to which the questions are closed and pre-coded.

Editing and checking of data

109. 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 errors in summaries.

110. 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 digging in time before the right correction can be found and non interactive edit runs would be more suitable.

Checking and manual editing

111. 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 request the enumerator and supervisor to check questionnaires and correct any errors 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

112. 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 e.g. 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, need to be left to a separate computer editing operation.

113. Programs for computer-assisted editing are often designed using database programs (IMPS, ISSA, CSPro, Visual Basic, etc.). 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.

Types of checks

114. Data on the questionnaires needs to be subjected to five kinds of checks: range checks, checks against reference data, skip checks, consistency checks, and typographic checks.

115. ***Range checks.*** 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.

116. ***Skip checks***. These verify whether the skip 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. For instance, children less than 5 years of age should be measured in the anthropometrical section but should not be asked the questions about occupation. Women aged 15 to 49 may be included in the fertility section, but men may not.

117. ***Consistency checks.*** 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. Some examples are as follows: *Demographic consistency of the household* (e.g. parents should be at least 15 years older than their children, spouses should be of different genders, etc); *Consistency of occupation* (e.g. the farming section should be present if and only if some household members are reported as farmers in the labor section); *Consistency of age and other individual characteristics* (the age of each person

should be consistent with personal characteristics e.g. marital status, relationship to the head of the household, etc.) and; *Control totals* (the data entry program should check that the control total equals the sum of the individual numbers).

118. ***Typographic checks.*** 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.

8.4.2.3. Handling of missing data

119. 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 questionnaire may not have been answered. Or some data in the questionnaire may have been faked or inconsistent with other information. Whatever the reason the effect is that the records are missing, empty or partly empty.

120. It is important to distinguish between missing data, i.e. data that should have been there but 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 e.g. 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.

121. The approach to take for genuinely missing data depends on which kind of data is missing. A selected sample element can be totally missing e.g. due to a refusal by the household to take part in the survey or due to inability by the household representative to answer the entire set of questions in the questionnaire. In such instances, 'unit non-response' is said to have occurred.

122. If a respondent is able to answer only some of the questions and not the others then 'item/partial non-response' has occurred because at least one item of the y vector is missing for that sample element.

123. Missing data of either type can give rise to biased and erroneous standard error estimates. The best would of course be to try and avoid non-response at the enumeration stage. A number of countries (e.g. Lesotho and others in Southern Africa) have a policy using substitutes for the non-respondents. Some substitute sample households are provided to the enumerator but the selection is done at head office. The pros and cons of such an approach can lead to some lengthy debate. Some cons for this include that it can be time consuming and prone to errors or bias.

124. In cases where all data is missing for the whole sample element, the easiest would be to base the tabulations on the remaining sample elements in the PSU, and to adjust the weights. For example, the sampling weight for each PSU can be recalculated as:

*Weight = (Old weight) * (Sample size)/(Number of responding holdings).*

125. 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 and there are several approaches that can be used:

Mean value imputation: the mean value (in the PSU or whole data set) is used to impute the missing value.

"Hot deck'" imputation: a record similar to the incomplete record is sought. The missing values are borrowed from such a record.

Statistical imputation: the missing value is imputed using a relation (regression, ratio) with some other variable, derived from complete data.

126. 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.

8.4.2.4. Data entry

127. The objective of data entry is to convert data the raw material (the information on the paper questionnaires) into an intermediate product (machine-readable files) that needs to 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. Sometimes double data entry procedures are used to ensure that this is the case but this is now rarely done. Nowadays, sampled verification tends to be more prevalent.

The data entry application

128. Normally the application comprises of 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.

129. 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 (i.e. wrong user name or password is entered), the application will shut down immediately. All user-names and passwords are store in a user table in the back-end, where the password is encrypted. When a user logs in to the system with a valid password, tables in the back-end are updated.

The data entry module

130. 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 to live up to:

- 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 questionable 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, e.g. 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).
- Keep track of the number of keystrokes and data entry time for later statistical use, e.g. predict the total data entry time.

The data verification module

131. 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 the number of keystrokes is summarized, but also the number of errors found. Some decision regarding the strategy for verification has to be decided upon. Options could be *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).

The data correction module

132. 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.

The supervisor administration application

133. 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.

Development platforms for data entry systems

134. According to Munoz (2003), there are many data entry and editing program development platforms available in the market, but few of them are specifically adapted to the data management requirements of complex household surveys. A World Bank review conducted in the mid 1990's found that at that time two DOS-based platforms were adequate: the World Bank's internally developed LSMS package and the U.S Bureau of the Census' IMPS program. Both platforms have progressed after that in response to changing hardware and operating system environments. CSPro – a Windows-based application that provides some tabulation capabilities, besides its primary role as a data entry and editing program development environment, has superseded IMPS. The LSMS package has evolved towards LSD-2000 – an Excel-based application that strives to develop the survey questionnaire and the data entry program simultaneously.

135. Both CSPro and LSD-2000 (or their ancestors) have proven their ability to support the development of effective data entry and editing programs for complex national household surveys in many countries. These platforms are also easy to obtain and use. Almost any programmer – in fact, almost anybody with a basic familiarity with computers – can be expected to acquire in a couple of weeks the techniques needed to initiate the development of a working data entry program.

8.4.3. File structure and datasets for tabulation and analysis

136. 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 the 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.

137. According to Yansaneh (2003) data from households in developing countries are not usually amenable to basic analysis (basic frequencies). One reason is that usually there is very little documentation on sample design for the surveys. Also the data files often do not have the format, structure, and requisite information that would allow any sophisticated analysis.

138. In order to facilitate appropriate analysis, the associated database must contain: all information on the sampling procedure; labels for the sample design strata, PSUs, SSUs, etc.; sample weights for each sampling unit; etc. This information will be needed for estimation of the required statistics and also for estimating the sampling errors of those estimates.

139. 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. At that stage it is likely that data lies in a big file not readily amenable for further operations and it would be necessary to split the data set for further operations.

140. One reason for splitting the initial full survey file is that it often contains so many variables that it is impractical to work with all of them simultaneously in the estimation process. It is more practical to split the information into different data sets.

141. The other reason is that the full survey file may in fact contain information about units that are sampled from different populations. For example, for a household budget survey the full file contains data on sampled households as well as on sampled persons. 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.

142. One approach is to divide the set of data files according to the objects specified by the object graph for the processing system. Typically, there are two main types of files from household surveys: household files and individual (person-specific) files. In most of the

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 full and final data files (datasets) will contain information on all responding households and individuals, from all the surveyed PSUs.

143. The pictures below are examples of how big files for different types of surveys can be reorganized to facilitate further processing.

- ***Example 1***: 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:
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:

- ***Example 2***: Typical files for an agriculture survey

File	Type	Contents
HOUSEHOLD	Household file	Household identification (Region, Province, District, etc.; Household id) 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 agricultural activities: work on holding, employment status, occupation, etc.
FIELDS	Household file	Household identification plus field identification Characteristics of the field: Area planted, crops planted, etc.

- ***Example 3***: Files used for the 1987 Zimbabwe Intercensal Demographic Survey (ZICDS)

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.

144. In general, after data entry, the processing of flat files is preferred. This is simplest file format and is judged to be more efficient for tabulation. Much of the available general software requires data in the flat form. 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. Examples of the format for a flat file are the file structures the 1987 ZICDS.

145. As explained earlier, four object types were identified for the survey. Correspondingly, four files were used and below are examples of the detailed layout of two of the files, the household and the person file.

- ***Example 4***: 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}

146. 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.

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 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 EAs 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**.**

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 stand for variables on the respective individuals.

- ***Example 5***: Person File

Identification					**Sampling design parameters**							**Variable values**			**Weight variable**
Stra-tum	Sub-division	EA	Hh	Pid	S_h	a_h	R_h	b_{hr}	S_{hi}	M_{hi}	m_{hi}	x	y	z	w
h	r	i	j	k								x_{hrijk}	y_{hrijk}	z_{hrijk}	w_{hrijk}

8.4.4. Estimation and assessment of the accuracy of estimates

8.4.4.1. Point estimation procedures and the calculation of weights

147. Weights (or raising factors) are needed in calculating national totals, averages, proportions, etc. for statistical parameters (characteristics). The sampling weight variable is required for estimation of a population parameter because of unequal selection probabilities for each sample element and/or because of non-response adjustments. The value of the sampling variable weight for a given respondent element in the dataset (denoted by w) is the number of elements represented by that w. The sum of the value of this weighting variable over all Ws in the dataset estimates the number of elements in the population. 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.

148. For a stratified two-stage sampling procedure with pps-sampling of PSUs and systematic sampling in the second stage, the calculation of weights can be explained as below.

First, an unbiased estimation of the population (the total) for the variable y is obtained as:

$$y-total=\sum y_k * w_k$$

k runs over sampled units

Where, the estimation weights w_k are:

$$w_k=\frac{1}{p_k}, k=1,2,\ldots\ldots, N$$

p_k = The (sample) inclusion probability for unit k (i.e. the probability that unit k is selected to the sample).

149. The first task in deriving weights is to derive the inclusion probabilities for sampled units. The inclusion probabilities are dependent on the sampling procedure used to select the sample. Generally, the household samples used in developing countries are selected using a two-stage sampling procedure with pps-sampling for the PSUs and systematic sampling for the second stage (i.e. for the selection of the households). Usually, the secondary sampling unit (SSU) is the household.

150. Since the two sampling stages are assumed to be carried out independently, the inclusion probability p_k for the $SSU_k = f_1(h,i) * f_2(j/h,i)$. That is, the inclusion probability

of the SSU is the product of the inclusion probability at the first stage $f_1(h,i)$ times the conditional inclusion probability at the second stage $f_2(j/h,i)$.

151. The two inclusion probabilities can be defined as follows:

- $f_1(h,i)$ = The probability that $PSU(h,i)$ is selected to the PSU-sample.
- $f_2(j/h,i)=$ The probability that $SSU(h,i,j)$ is selected in the second stage given that $PSU(h,i)$ was selected in the first stage.

Explicit expressions for f_1 and f_2 can be introduced as follows:

n_h = Number of PSUs selected in stratum h.

S_{hi} = The size value of $PSU(h,i)$

S_h = Total value size of stratum h, $S_{h1}+S_{h2}+S_{h3}+.....+S_{hNh}$

M_{hi} = Number of SSUs in $PSU(h,i)$

m_{hi} = Number of SSUs in sampled $PSU(h,i)$

When pps-sampling with sizes S_{hi} is used in the first sampling stage, we have:

$$f_1(h,i)=\frac{n_h * S_{hi}}{S_h}$$

And in the second stage, we have:

$$f_2(j/h,i)=\frac{m_{hi}}{M_{hi}}$$

152. Once the inclusion probabilities are calculated, we can obtain the estimation weights as the inverse of the joint inclusion probability as below. That is for stratified two-stage sampling with pps-sampling of PSUs and systematic sampling in the second stage, the estimation weights w_k for SSU_k, $k=(h,i,j)$ is:

$$w_k=\frac{1}{p_k}=\frac{1}{f_1 * f_2}=\frac{S_h}{n_h * S_{hi}} * \frac{M_{hi}}{m_{hi}}$$

153. 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.

154. Most software will allow one to specify a weight variable and to do this calculation more directly. However, it is not correct to simply calculate standard errors for estimated parameters using standard formulas applied to this modified (weighted) dataset. This is discussed in the next section of the chapter.

155. It should be noted, however, that accurate weights must incorporate three components. What is said above only incorporates one of the components i.e. the base weight. 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. For example, it is typical that wealthier households refuse to respond than middle income and lower income households. The base weights need to be inflated by the inverse of the response rate for all groups of households. Finally, in some cases there may be 'post-stratification adjustments'. The basic notion being that some data source, e.g. a census, may provide very precise estimates of the distribution of the population by age, sex, etc. If the survey estimates of these distributions survey do not match those given by the other, more accurate data source, the survey data should be re-weighted so that, with the new weights, the survey reproduces the distribution from the other data source.

156. Another complication for the estimation process is necessitated by the increased demand for domain level statistics. A domain being a subset for which estimates are desired. Usually they may be specified at the sample design stage but may also be worked out e.g. from the derived data. A domain may 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 (e.g. age, sex, ethnic group, poor, etc.). What follows is some attempt to describe how datasets can be constructed to facilitate estimation for domains.

157. We can start by visualizing an observation file (e.g. the household file) 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.

a. Identification for the Household
b. Sampling parameters
c. Values for the study variables *x*, *y*, and *z*.
d. The value of the household's estimation weight.
e. If the household belongs to category c or not.
f. If the household belongs to domain g or not.

158. 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.

159. 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 e.g. 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 (e.g. if the domain g consists of households with 3 + children). At the estimation stage the values of such indicators should be available in the observation file.

160. When all the data is available in the observation file, it will look as shown below, except that the sampling parameters are not included.

Final observation file

HID	**x**	**y**	**z**	**c**	**g**	**w**
1	y_1	x_1	z_1	c_1	g_1	w_1
2	y_2	x_2	z_2	c_2	g_2	w_2
3	y_3	x_3	z_3	c_3	g_3	w_3
.	.	.		.	.	.
w	.	.		.	.	.
n	y_n	x_n		c_n	g_n	w_n

161. Computation algorithms for estimates of totals, averages, proportions, etc. over the categories and domains are obtained from formulas similar to those for the estimation of statistical characteristics for the household population. The only modification is the inclusion of filters for categories and domains.

162. 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. Basically, the estimation weight for a person is the same as that for the household to which the person belongs. Since all members in a 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.

8.4.4.2. Assessment of the accuracy of estimates

163. The notion of *error*, regarding an estimate for a statistical characteristic (or parameter), refers to the *difference* between the estimate (say $\hat{Y}$) and the theoretical 'true parameter value' (say Y) that would be obtained if all sources of error are eliminated. Simply put, suppose that the estimate of the average monthly income for a certain population in a survey is 900 USD, and that the actual average obtained from a complete enumeration without errors of reporting and processing, is 850 USD. Then the error of the estimate would be 50 USD. Alternatively, *error* could be labeled as the *deviation*.

164. *Sampling errors* refer to differences between estimates based on a sample survey and corresponding population values that would be obtained if a census was carried out using the same methods of measurement, and are 'caused by observing a sample instead of the whole population'. *'Nonsampling errors'* include all other errors affecting a survey. They can and do occur in all sorts of surveys, including censuses. Earlier sections of this chapter go into how some of the nonsampling errors can be handled. The focus below is on the handling of sampling errors.

Handling sampling errors

165. In many surveys undertaken in developing countries, sampling error estimates are not usually computed nor published and commentary on survey estimates often leaves out the degree of accuracy.

166. To assess the accuracy of an estimate one first has to estimate the variance of the point estimator. A guiding principle is to use estimation procedures, which are at least approximately, unbiased, and which give, in the particular situation, a minimum standard deviation (or relevant variance).

167. Variance estimation is important because it indicates the precision of estimators, leading to confidence intervals and testing hypotheses about population parameters. The most common way of giving uncertainty measures is by means of confidence intervals or margin of error.

168. For example the formula for a 95% confidence for a total 't' is $t = \pm 1.96 * s(t)$, where s(t) = the estimated standard deviation for the estimator t.

169. In the same way the formula for a 95% confidence for an arithmetic mean (*am*) is

$$am = \pm 1.96 * s(am), \text{ where}$$

$s(am)$ = the estimated standard deviation for the estimator *am*.

170. And the formula for a 95% confidence for percentage *(p)* is

$$am = \pm 1.96 * s(p), \text{ where}$$

$s(p)$ = the estimated standard deviation for the estimator *p*.

171. The first step in the computation of the uncertainty measure is estimation of the variance of the (point) estimator. Hence the computation of uncertainty measure is often referred to as variance estimation.

172. Special software programs are needed to calculate estimates of the standard error that reflect the complexity of the sample design actually used. Such complexities include stratification, clustering, and unequal sampling probability (weighting). Standard software cannot be used in this regard because they almost always assume that the data has been produced under a scheme of random sampling. Thus in general they will understate the true value of variability of survey estimates. These in turn can lead to drawing invalid inference from the data. There are a number of software packages now available for variance calculation for complex survey data. Examples of such software are: WESVAR; SUDAAN; STATA; etc.

8.4.5. The analysis of survey data

173. The term 'analysis' is difficult to define. It is widely used but has no unique and precise meaning. Essentially, it is the turning of data into information and is a necessary extension of the data processing and the tabulation of survey results. The analysis of household survey data can be seen as the transformation of the respective data into information that throws light on the social conditions and social development of the surveyed population.

174. Analysis often involves data from the particular survey as well as relevant data and information from other sources within and outside the national statistics office (NSO). Many of the analyses of statistical data from the survey are likely to be carried outside the NSO by many different types of users: researchers, policy and decision makers, financial institutions, mass media, schools, etc. This is particularly the case in developing countries because of the lack of resources and capacity for further processing beyond tabulation.

175. Data analysis ranges from very simple summary statistics to extremely complex multivariate analysis. According to the Australian Bureau of Statistics (1996) the following is a list of the main types of analysis:

- *Simple derivations* (e.g. means, medians, quantiles, histograms, two and three way tables, scatter plots, etc.).

- *Complex derivations* (e.g. Gini coefficients, expectation of life, etc.).
- *Data models* (e.g. population estimates, population projections, etc.).
- *Social indicators* (e.g. fertility rate, dependence ratio, labor participation rate, etc).
- *Mathematical modeling* (e.g. linear regression, factor analysis, etc.).

176. The focus of this section is more on the simple derivations or what is generally referred to as basic descriptive statistics presented using summary measures (totals, means, ratios, percentages, relationships, etc.) through tables, graphs, etc., on one variable, two variables, and for more than two variables. These are what household surveys in developing countries are commonly designed to produce.

177. Examples of estimates of totals are number of persons economically active, number of persons employed, number of persons unemployed, etc. An example of a mean is the average income of women in the labor force. For ratios of total/means we have e.g. the proportion of households with total income below the poverty line or the average household income for female-headed households in relation to the population, respectively. The estimates can be at the domain or national level.

178. The first step towards analysis is to organize the dataset, for example, by main survey objects as earlier discussed in the chapter. Once the dataset has been created the next step is to generate the basic descriptive statistics and present them. The presentation can be in terms of a single variable, in terms of two variables, or for three or more variables. This can be in tabular or graphical form.

8.4.5.1. Descriptive versus explanatory purpose

179. If the survey has a descriptive purpose, then focus is to estimate the parameters (values such as total number, percentage, average, etc) in a population e.g. the number employed, unemployed, under-employed, etc. This can also be labelled the 'What' analysis that focuses on significant facts surrounding an issue.

180. A survey may also have an explanatory purpose. Apart from knowing the facts around a particular issue, one may also want to address the 'Why' questions i.e. not merely showing a trend moving in a certain direction (e.g. by a graph or indicator) but also to explain why it has so moved. In that case the purpose is to establish the causal relationship between two or more variables. This may be connected to theory, a model, or earlier findings. Due to the non-experimental character of a typical household survey this must be done with care. However, there are techniques of controlling for this problem. The control could be obtained by the so-called table analysis, as discussed later in the chapter.

Table 1. Occupational distribution among the employed population in Tanzania LFS, 1990/91

Occupation	Population
Admin/Managers	214399
Professionals	17980
Ass. Professional	176115
Clerks/Cashiers	96435
Service/Shops	269435
Agriculture – Skilled	9114437
Craft. Etc. Workers	372567
Operators - Plant Machinery	120720
Sales/Labours	507117
Total	**10889205**

181. The table above is purely descriptive. To get more information out of the material the data could also be broken down to show some subpopulations (sexes, age-groups, etc.). Below the same material is broken down by sex. The table can be interpreted as describing the population in only one dimension (occupational distribution) but separately for two subcategories of the population (men and women).

Table 2. Occupational distribution among the employed men and women in Tanzania LFS, 1990/91

Occupation	Men	Women	Total
Admin/Managers	169371	45028	214399
Professionals	16148	1832	17980
Ass. Professional	123091	53024	176115
Clerks/Cashiers	51320	45115	96435
Service/Shops	149552	119883	269435
Agriculture – Skilled	4210747	4903690	9114437
Craft. Etc. Workers	335512	37055	372567
Operators - Plant Machinery	107646	13074	120720
Sales/Labours	291712	215405	507117
Total	**5455099**	**5434106**	**10889205**

182. In a statistical report a typical descriptive comment to the table could be:

> Agricultural work is the most frequent occupation among both men and women. 90 per cent of the women and 75 per cent of the men are employed in agriculture.

183. The table could, however, be interpreted in another way – as showing correlation between the variables sex and occupation. In this case we look at the table as two-

dimensional, one dimension being occupational distributions, the other being sex. One could surmise that sex affects occupation from the table:

> Women constitute about 50% of the employed. While women are often employed in agricultural work there are almost as many women as men in occupations as clerks/cashiers, service/shops and sales/labourers. Men are often employed in administration, management and professional occupations as well as in positions as craftsmen and operators.

184. In this case the survey data is used to explain some aspects of reality. We use the survey data for an explanatory analysis, relying for the interpretation on knowledge of sex patterns not evident from the survey itself.

8.4.5.2. Relationships between variables

185. The variable that one may wish to explain is generally labelled the dependent variable. The variable expected to explain the change in the dependent variable is referred to as the independent variable. Most of the phenomena that one may wish to analyse also often call for more than one independent variable to explain variation in the dependent variable. This happens because of the complexity of social phenomena. One independent variable usually explains only a certain amount of the variation in the dependent variable, and more independent variables have to be introduced in order to explain more of the variation.

186. Example: Assume that there is a relationship between underemployment and geographic area. In this case underemployment is considered to be the dependent variable and geographic the independent variable. When underemployment is studied as a dependent variable, geographic area would explain some of the difference in the number of persons being underemployed. This explanation is however, incomplete because there are reasons in addition to geographic areas that explain the variations in the dependent variable (underemployment). Among these additional independent variables maybe education, occupation, and age.

8.4.5.3. The object (counting unit or unit of analysis)

187. Decisions about the social issue to be analysed, the data to be used, and the statistical technique to be applied, are fundamental to good analysis. However, an even more basic level question is the object or unit of analysis.

188. Earlier in the chapter an **object** is defined as any concrete or abstract entity (physical object, living creature, organization, event, etc.) that the users may want to have information about. **Objects** for the particular household survey are items (elements or units) that the users would like to have information on (e.g. household, person, etc). For most household surveys the basic object is the HOUSEHOLD. There may be several associated objects related to the basic object, and these will depend on the particular survey.

189. The definition of an **object** is intimately tied to the social issue (survey objectives) about which data are collected and analysed. Below we explore examples of how some of the key objects in household surveys are generally defined:

190. **HOUSEHOLD -** A household is a group of persons who make common or joint provision for food and other essentials of living. A household may be single-person or multi-person. The concept of household is especially useful in the collection and analysis e.g. of data on income and expenditure. Many expenditures are household rather person based (e.g. expenditure on furniture, telephone and electricity bills, rates, etc.).

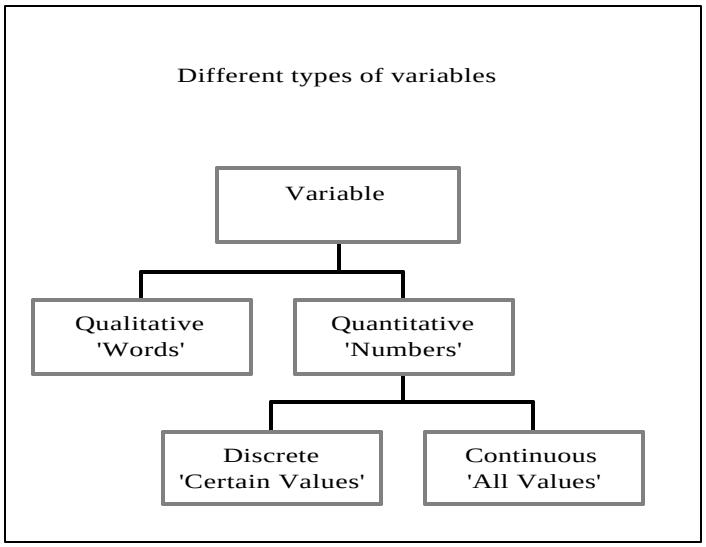
191. **PERSON** - The person is a usual member of the household or a visitor during the night previous to the survey interview. Aggregates or populations of people represent one of the most important outputs of household surveys. Sometimes sub-objects (sub-populations) of the larger object 'person' can be defined (e.g. Youth, Woman ≥12 Years Old, etc.).

192. **EVENT** – E.g. births, deaths, visits to the doctor, crime offences, etc. The event in aggregate is described as the incidence of that event.

193. **PHYICAL OBJECT** – E.g. dwelling, school, hospital, etc. While the actual number can be of great analytical interest (e.g. a social indicator of hospitals per capita), the characteristics of the object are also important (e.g. hospital beds, facilities, and equipment, staff numbers, etc.).

8.4.5.4. Different types of variables

194. For every object, there will be several variables of interest. **Variables** being **properties** (attributes or characteristics) of the objects e.g. for the object PERSON we can have age, income, occupation, marital status, etc., as variables. Variables (e.g. age, birthplace, etc.) of survey objects (e.g. person, household, etc) are coded according to classifications (e.g. 5-year age groups, countries of birthplace, etc.). Apart from the choice of the object of analysis, it is the variables and their classifications that determine the limits and richness of the analysis.



195. The above diagram summarises the division into types of variables for which we need to choose the correct type of table or chart. Variables may be qualitative or quantitative.

196. The first division whether a variable is numerical (i.e. a variable for which the value is indicated using numbers) and is called quantitative (e.g. the variable weight – 8 kg, etc.). If instead we use words then we speak of a qualitative variable e.g. the variable place of residence is described in words (urban-rural).

197. Among the quantitative variables we distinguish between those that can only take certain values and those that can take all the values in a range. Variables that can only take certain values (usually whole numbers) are called discrete – the number of children is, for example discrete. Variables that can take all numbers in a range (e.g. age) are called continuous.

198. Different types of tables or charts suit different types of variable. A typical example is the histogram, which is used for continuous variables. It would not be appropriate to use a histogram to illustrate a discrete variable.

8.4.5.5. Levels of measurement

199. Variables can be measured in a variety of ways. Measurement is the assignment of numerals or numbers to objects, events or variables according to rules. Many times we measure the indicators of concepts that cannot be observed directly. They must be inferred from the measurement of indicators of the concepts. Examples of such concepts are power, democracy, unemployment and need for treatment. Indicators are specified by operational definitions. After observation of the indicators, numerals or numbers are substituted for the values of the indicators, and a quantitative analysis is performed.

200. In some cases the type of variable determines the type of measurement whilst in some cases there is a choice. The variable 'age', for example, can be given in years but can also have measurement values such as youth-middle age-old. Different types of measurement give different types of information about a variable concerned. This influences our choice of the table or chart to use. Below are the different scales of measurement.

Level of measurement	**Characteristics, the measurement values can be …**
Nominal scale	Distinguished
Ordinal scale	Distinguished and ranked
Interval scale	Distinguished, ranked and measured with constant units of measurement

201. The Nominal gets its label from the Latin word 'nomen' meaning name. The ordinal scale is so called because the order in which the values are placed is important. Correspondingly intervals have a significant meaning in the interval scale.

Nominal level (scale)

202. This is the lowest level of measurement. The place of residence variable (urban-rural) is obviously measured on a nominal scale, since we are not able to (save for some meaningless alphabetical order) rank these two measurement values in any meaningful way. Other examples of variables on a nominal scale are gender, nationality, and geographic area.

Ordinal level (scale)

203. Variables at the ordinal level are not only classifiable but also exhibit some kind of relation. Typical relations are 'higher', 'greater', etc. If we measure according to the scheme youth-middle age-old we have an ordinal scale. We know which value is lowest or highest but not how big a difference there is between the different values.

Interval level (scale)

204. At the interval level we are able to rank a set of observations in terms of the greater than relation and we know the exact distance between each of the observations and this distance is constant. All the descriptive and inferential statistics are applicable to interval data. Examples of variables at an interval level are crime rates, voting participation rates, etc. Dates are also good examples of the interval scale. The unit of measurement, the year is constant.

205. Different levels of measurement give different charts. The level of measurement influences the choice of chart principally in that we ought to choose a chart that retains the characteristics of the measurement values. It is, for example, obvious that there ought to be equal distance between the years in a time series chart and that as far as possible we should retain a zero point in order to able to interpret ratios.

206. The level of measurement is also important in other areas of descriptive statistics. For example, in order for us to be able to calculate a mean value it is necessary that the variable we are working with is measure on at least the level of the interval scale.

8.4.5.6. Table construction

207. It should be acknowledged that most of the tables are based on Puide (1994).

(a) Frequency distributions – one dimensional tables

208. After data have gone through the data preparation phase (coding, data entry and editing, etc.), they are ready for analysis. The first task is to construct frequency distributions to examine the pattern of response to each and of the dependent and independent variables under investigation. To construct a frequency distribution, one simply lists the categories of interest for the variable and counts the number observations in each category as follows.

Category	Frequency (F)
I	F
II	F
III	F
IV	F
Total	**N**

209. The left column shows the categories of the variable (I-IV) and the right column shows the number or frequency (F) of observations in each category. The last row is the total (N) of all frequencies in the table.

210. The categories can be listed in any arbitrary order if the variable is the nominal type. The categories of ordinal variables represent different rankings and therefore must be arranged in a certain order.

211. When summarising interval variables in frequency distributions we must decide on the number of categories to use and the cutting point between them. Interval variables are ordinarily continuous; they can take any value within an interval. The classification into distinct categories may be quite arbitrary. For example, age may be classified into one-year, two-year, five-year groups.

212. The choice of categories should reflect the problem, the purpose of the survey and the questions being asked.

(b) Percentage distributions

213. Summarising the data by constructing frequency distributions of single variables is only the first step in data analysis. In the next step the frequencies must be converted into measures that can be interpreted meaningfully. An absolute frequency is meaningless in itself; it must be compared with other frequencies.

214. For example: the information about 20 women in a certain village is interesting only in relation to the number of men in the village, the number of children in the village, or the number of women living in the nearby village.

215. Proportions and percentages permit the comparison of two or more frequency distributions. Below we have two tables presenting the distribution of education among rural and urban populations.

Table 3. Education distribution: Rural population (in absolute frequencies and percentages)

Educational level attained	Frequency (F)	Percent (%)
Primary, not completed	60	15
Primary, completed	300	75
Above primary	40	10
Total	**400**	**100**

Table 4. Education distribution: Urban population (in absolute frequencies and percentages)

Educational level attained	Frequency (F)	Percent (%)
Primary, not completed	20	8
Primary, completed	200	80
Above primary	30	12
Total	250	100

216. Although there are more people with primary education in rural than urban areas (300 versus 200), a direct comparison of the absolute frequencies is misleading since the total N is different in each population. Instead the relative weight of the classes within each distribution, the frequencies should be expressed in percentages.

217. Whereas the population with primary education constituted about 75 percent of the rural population, it was 80 percent for the urban group. The new figures make it easier to compare the frequency distributions.

(c) Bivariate distributions (two dimensional tables, two way tables) – Cross Tabulation

218. A more challenging form of analysis is the discovery of 'significant' relationships between variables. Explanations and predictions involve relating the phenomena to be explained (dependent variables) to other explanatory phenomena (independent variables). In terms of analysis a relation implies a relation between two or more variables. When we say X and Y are related, we mean that there is something in common to both variables. For example, if we say that education and income are related, we mean that the two 'go together', i.e. that they covary. If the relation between education and income were positive we would find that increases in the level of education attained lead to higher income.

219. Typically, variables are connected through cross tabulation. The tabulations are then examined for significant relationships (sometimes aided by regression models, etc.). The

significance is interpreted in the light of the social issues under enquiry. Text, tables, graphs, social indicators, etc are used to communicate the results.

220. In a bivariate table, there are two cross-classified variables. Such a table consists of rows and columns. The categories of one variable are labels for the rows while the categories of the second variable are labels of the columns. Usually the independent variable is the column variable (listed across the top) and the dependent variable is the row variable (listed on the left side of the table).

221. The three tables below demonstrate the concept of covariation. The tables summarise information on two variables, income and nationality.

Table 5. Income by ethnic group (perfect covariation)

Income	**Nationality**			
	Indigenous Tanzanian	**Indian origin**	**European origin**	**Total**
High	0	0	8	8
Medium	0	8	0	8
Low	8	0	0	8
Total	**8**	**8**	**8**	**24**

222. Table 5 illustrates a perfect pattern of covariance of the variables. All Indigenous Tanzanians have low income, all persons of Indian Origin have a middle income and all persons of European Origin have a high income. The two variables, obviously, covary since specific categories in one variable go with specific nationality within the same income group.

Table 6. Income by ethnic group (moderate covariation)

Income	**Nationality**			
	Indigenous Tanzanian	**Indian origin**	**European origin**	**Total**
High	0	2	6	8
Medium	1	6	1	8
Low	7	0	1	8
Total	**8**	**8**	**8**	**24**

223. Table 6 is another case of covariation but moderate. Not all persons in a specific group are categorised in the very same income class. But it can still be stated that most persons of a specific nationality fall within the same income class.

Table 7. Income by ethnic group (near-zero covariation)

Income	**Nationality**			
	Indigenous Tanzanian	**Indian origin**	**European origin**	**Total**
High	2	3	3	8
Medium	3	2	3	8
Low	3	3	2	8
Total	**8**	**8**	**8**	**24**

224. Table 7 shows a pattern where the two variables are independent of each other. They do not covary. We cannot say anything about a person's income on the basis of nationality. Ethnic group in this case is a bad predictor of income.

225. Tables 5, 6, and 7 are bivariate distributions. The bivariate distributions consist of categories of two variables and their joint frequencies. Each table has two dimensions – the variables nationality and income. Each table has 9 frequency cells. They contain information with a specific combination of values in both variables. Each of the bivariate distributions in the 3 tables can be seen as the combination of three univariate distributions, one for each nationality.

(d) Percentages for bivariate distributions

226. To summarise a bivariate table it is useful to compare its univariate distributions. Presenting its frequencies in percentages can achieve this. It is also the predominant method of analysis. The scale or level of measurement has no importance in this case. The method can be used in the analysis of variables at nominal level, ordinal level, or interval level.

227. Below is a cross tabulation of two sub-groups of a given population, called here A and B, and marital status among public sector employees. We want to investigate if there is a difference between the two groups in the public sector regarding their marital status.

Table 8. Distribution of marital status in two groups among public employees (numbers)

Marital status	**Groups**	
	A	**B**
Married	49	59
Single	2	2
Divorced	10	5
Widowed	11	-
Total	**72**	**66**

Table 9. Distribution of marital status in two groups among public employees (percentages)

Marital status	Groups	
	A	**B**
Married	68.1	89.4
Single	2.8	3.0
Divorced	13.9	7.6
Widowed	15.3	-
Total	**100 (N=72)**	**100 (N=66)**

228. Each group has been treated as univariate distribution. The frequencies have been transformed into percentages by using the total number in each distribution as a base.

229. Now we can compare the two univariate distributions to see if there is a relationship between groups and marital status among public employees.

230. The computation of percentages goes down the columns- the independent variables. The comparison cuts across the row – the dependent variables.

231. By comparing the proportions of individuals in group A who are married with the proportion of individuals in group B who are married etc. we can decide whether there is a correlation between the two groups and marital status.

8.4.5.7. Control techniques

232. An association between two variables is not a sufficient basis for inferring that the two variables are causally related. Other variables must be ruled out as alternative explanations. For example, a relationship between height and income can probably be accounted for by the variable age. Age is related to both income and height, and this joint relationship produces a statistical relationship that has no causal significance. The original relation between height and income is said to be a spurious relation.

233. Cross tabulation is a method of control that can be compared to the mechanism of matching, used in experiments. In both techniques, the investigator attempts to evaluate the groups examined with respect to variables that may bias the results.

234. Cross tabulation involves the division of the sample into subgroups of the controlled variable. The original bivariate relation is then tested within each subgroup.

235. Only variables that are associated with both the independent and dependent variable can potentially bias the results and are to be selected as control variables.

236. To illustrate the logic of analytical table construction we have the following 2 tables that are two-dimensional tables. In the first table we have the variable 'poverty' tabulated

against 'level of education'. In the second table 'poverty' is tabulated for against two given sub-groups, called here A and B. In both, 'poverty is the dependent variable. The 'level of education' and the selected 'sub-group' are the dependent variables.

Table 10. Poverty by education

Poverty	**Education**	
	High	**Low**
Poor	28	50
Not poor	72	50
Total	**100**	**100**

237. The above table shows that poverty is lower among those with high education.

Table 11. Poverty by two sub-groups of the population

Poverty	**Group**	
	A	**B**
Poor	60	20
Not poor	40	80
Total	**100**	**100**

238. The above table shows that poverty is lower for those in group B.

Table 12. Education by two sub-groups of the population

Education	**Group**	
	A	**B**
High	20	75
Low	80	25
Total	**100**	**100**

239. The above table shows the relation between education and sub-groups. It is obvious that there is a relationship between education and the two sub-groups. The majority of the individuals in group A have a low education, while most of the individuals in group B have high education. Therefore the third selected variable (the selected sub-groups of the population) is related to both original variables (poverty and education).

240. The question then arises is if the chosen sub-groups and level of education both influence poverty or if poverty among individual in sub-group A is higher because these individuals do not get higher education to the same extent as individuals in subgroup B and poverty is lower among the more educated. Should this be the case when keeping the level of education constant, the correlation between the sub-groups and poverty should disappear. A

three-dimensional table combining poverty, education, and the two sub-groups could reveal this. The two background variables are inserted in the heading of the table.

241. To control for the education variable we have divided each level of education by the sub-groups, in the table below.

Table 13. Poverty by education controlling for sex

Poverty	**Groups**			
	A		**B**	
	Education High	**Education Low**	**Education High**	**Education Low**
Poor	60	60	20	20
Not poor	40	40	80	80
Total	**100**	**100**	**100**	**100**

242. The relationship between education and poverty is completely accounted for by the two sub-groups. There is no causal relationship between education and poverty.

243. In the above table we have studied the relation between poverty and two given sub-groups keeping the level of education constant. The same material could of course be re-arranged to show the relation between poverty and level of education keeping the sub-groups constant. The choice of control depends on what we want to highlight and what we want to comment upon.

8.4.6. The reporting, presentation and dissemination of survey data

244. Household surveys have the potential of generating a wealth of information. However, the various stakeholders can only realize the full potential impact and benefits of such information with its greater dissemination and utilization.

245. There are three major basic outputs from a statistical survey:

- *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 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.

246. These basic outputs need to be packaged and should be made available in a user-acceptable form through appropriate distribution channels.

8.4.6.1. Survey reports

247. The results from a household survey (micro and macrodata) can be made available in different forms e.g. statistical tables and graphs, and on suitable media e.g. paper publications and electronic databases, and disseminated through suitable channels (e.g. through publishers, fax, and Internet).

248. Traditionally, macrodata are published and stored in statistical tables in printed publications. Though this tradition will continue to exist, it is likely that the production of such publications will move and will be based on electronic outputs. It is also likely that they will first be stored on and published through a database service. Thereafter, printed results will be produced automatically from the data and metadata contained in the database.

249. The traditional approach has usually included the following documents:

a. The preliminary report - to serve as a final test for the entire production process and to provide early results from the survey.
b. The main report – bearing most of the planned outputs.
c. The technical report - a survey methodological report.

The structure of the main survey report

250. Survey reports are basically of two different kinds:

- The most common report includes only two parts - A short introductory text followed by detailed tabulations.
- The second type is more analytical. It usually contains an extensive text part including tables and diagrams/charts, followed by detailed tabulations.

251. Irrespective of the kind, every report should meet the following demands:

a. Summary chapter including:
   - Main results illustrated by tables, diagrams/charts on an aggregated level.
   - This part should be user-oriented and be found easy to read by the users.
b. Introductory chapter including:
   - The main objectives of the survey
   - The main issues/questions
   - The methods used
   - The definitions
   - Quality of data
   - Interpretation of data
   - References to the technical report (if any).
c. Detailed tabulations.
d. The questionnaire (as an appendix).

252. The more analytical report will include a text part within its several chapters. Usually the objectives of the survey will serve as a structure where each main issue is discussed in a separate chapter. Furthermore, subgroups or defined target groups would be treated separately in one or more chapters. References to the underlying theory, the model (policy relevant), earlier findings in the field, population census data etc. will be used as background material, explanations and support of the findings.

The structure of the technical report

253. The technical report is often included in the introductory chapter of the Main Report. This is the convention in annual surveys. Surveys carried out for the first time or intermittently demand a detailed and separate technical report. The main items contained in the report could be as follows:

a. Introduction – some overview of the survey.
b. Methodology – the general planning of the survey, the use of sampling methods, etc.
c. Data processing – the whole organization of data processing for the survey, the coding schemes, etc.

8.4.6.2. Databases

254. The aim for a household survey program should eventually be a situation where the data archiving is based on a combination of micro-level and macro-level data. For a survey 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. This is where the systems design model discussed earlier – describing the objects types on which information has been collected and file structures used – can be particularly useful.

255. Some of the basic requirements for an effective dissemination database are that:

- The data files are edited and documented to high standards and are structured to facilitate analysis.
- The micro-level data files are supported by full description of the data (providing code books, marginal distributions, summaries, aggregates, etc., ideally in a machine-readable form).
- The dissemination database holds for each survey (or series of surveys) number tables.
- When figures are added to the dissemination database figures causing disclosure problems must be suppressed.
- Technical and computer support is provided where possible to data users.

256. Further, and perhaps more important, there should be retrieval functions in the system enabling the user to retrieve and analyze both time series and cross-section data.

257. In addition there should be interfaces to software packages for econometric modeling and multivariate analysis.

258. Examples of databases are *MASEDA and the LDB need to be pursued.*

8.4.6.3. The Internet and Web sites

259. Internet technology is composed of a number of functional features:

- *Electronic mail* (a common engine for sending electronic messages).
- *Websites* (the most widespread Internet facility that offers many different functional possibilities).
- *FTP (File Transfer Protocol) server* (a basic function to disseminate data files).
- *Browsers and mailing software* – software packages used on the user workstations to manage the Internet functionalities. They provide access to websites and FTP-servers, and enable receipt and sending of e-mails. MS explorer and Netscape are probably the most widely used browsers in the world today.

260. Web sites are, increasingly, becoming dissemination channels for statistical data. The Web site may offer a simple, comparatively cheap and efficient way to provide timely information to the core users of statistics as well as to a broader audience. Internet pages can be created as soon as the result from the survey becomes available. Very detailed information can be provided to users at very low extra cost. The content of the web site should consist of:

Fixed statistical tables and graphs with key figures. This kind of information is probably most requested by the not so advanced user, i.e. the general public and journalists.

Download functions for tables and publications. An advanced user (for example planners and analysts) will require more detailed information and they would probably like data in a form, which permits them to incorporate data into their own work. Formats for downloading could be Microsoft's xls (Excel) and doc (Word).

List of publications and statistical publications in a web readable format. The list of all the publications with possibility to order the publications should be available at the web site. The publications/report could also be available in full text version on the web site.

Metadata. The availability of metadata should support and provide understanding and transparency of the content of data, definition of variables, classifications, description of statistical surveys providing statistical data, etc.

8.4.6.4. Other electronic media

261. Diskettes or compact disks (CD-ROM) are able to store copies of survey publications, more detailed cross-classifications, etc. They especially have the advantage of being cheaper than publications, less bulky for transport and storage, and when necessary the required pages can be printed by the user. Also the data can easily be re-analyzed given that it will be machine-readable form.

References and further reading

Anders W et al. (1996): Graphing Statistics and Data, SAGE, 1996.

Backlund S. (1996): Future Directions on IT Issues – Report from a mission to National Statistical Center (NSC), Vientiane, Lao PDR, September 9-16, 1996.

Chromy J.R./ Abeysasekara S. (2003): Analytical uses of survey data, United Nations Statistics Division, March 2003.

Chronholm P. and Edsfeldt ? (1996): Course and Seminar on Systems Design from – Report from a mission to Central Statistics (CSS), Pretoria, South Africa, November 19 – December 8, 1996.

Brogan D. (2003): Comparison of Data Analysis Software Suitable for Surveys in Developing Countries, United Nations Statistics Division, 11 March 2003.

Eurostat (2003): Metadata production systems within Europe, The Case of Statistical Office of the Republic of Slovenia.

Giles M. (1996): Turning Data into Information, A Manual for Social Analysis, The Australian Bureau of Statistics.

Glewwe B.I. (2003): An Overview of Questionnaire Design for Household Surveys in Developing Countries, United Nations Statistics Division, January 2003.

Graubard B.I. and Korn E.L. (2002): The use of Sampling Weights in the Analysis of Survey Data, United Nations Statistics Division, July 2002.

International Labour Office (1990): Survey of economically active population, employment, unemployment and underemployment. An ILO manual on concepts and methods, ILO, Geneva.

Jambwa, M., M. (?): Application of Database Technology in the African Context Harare, Central Statistics Office, Invited Paper, 46th session of ISI, Japan.

_____ (1989): Data processing at the Central Statistical Office – Lessons from recent history. Harare, Central Statistics Office.

Jambwa M., Parirenyatwa C., Rosen B., (1989): Data processing at the Central Statistical Office – Lessons from recent history. Harare, Central Statistics Office.

Lagerlof B. (1988): Development of Systems for National Household Surveys – SCB R&D Report – Statistics Sweden, 1988:4.

Lehtonen R. and Pahkine E.J. (1995): Report on short tem mission on for Design and Analysis of Complex Surveys. John Wiley & Sons.

Lundell, L-G. (1996): Information Systems Strategy for CSS – Report to Central Statistical Service (CSS), Pretoria, South Africa, March 12 – 29, 1996.

Munoz J. (2003): A Guideline for Data Management of Household Surveys, United Nations Statistics Division, May 2003.

Olofsson, P. O. (1985): Report on short term mission on A Labour Force Survey in Lesotho 1985/86, Proposals for survey design, Kingdom of Lesotho, Bureau of Statistics.

Olsson, U (1990): Approaches to Agricultural Statistics in Developing Countries – An Appraisal of ICO's Experiences. Statistics Sweden International Consulting Office.

_____ (1990): Applied Statistics Lecture Notes. Statistics Sweden International Consulting Office – Special Reports TAN 1990:1.

Petterson H. (2003): The Design of Master Sampling Frames and Master Samples for Sample Surveys in Developing Countries, United Nations Statistics Division, February 2003.

Pedro Luis do Nascimento Silva. (2002): Reporting and Compensating for Nonsampling Errors for Surveys in Brazil: Current Practice and Future Challenges, United Nations Statistics Division, ….

Puide A. (1994): Report on a mission to Takwimu, Dar es Salaam November 21 – December 21, 1994, Statistics Sweden International Consulting Office, TASTAT 1994:20 (January 1995).

Rauch L. (2001): Best Practices in Designing Websites for Dissemination of Statistics; United Nations Statistical Commission and Economic Commission for Europe, Geneva.

Rosen B. and Sundgren B., (1991): Documentation for reuse of microdata from surveys carried out by Statistics Sweden. Statistics Sweden.

Rosen, B. (2002): Theory for Sample Surveys.

_____ (2002): Report on short term mission on Framework for the Master Sample, Kingdom of Lesotho, Bureau of Statistics, LESSSTAT 2002:7.

_____ (2002): Report on short term mission on Estimation procedure for Master Sample Surveys, Kingdom of Lesotho, Bureau of Statistics, LESSTAT 2002:?

Sundgren, B. (1984): Conceptual Design of Databases and Information Systems. P/ADB Report E19. Statistics Sweden.

_____ (1986): User-Oriented Systems Development at Statistics Sweden. U/ADB Report E24. Statistics Sweden.

_____ (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 Modeling 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, South Africa, July 9 – August 10, 1996.

Thiel, L. Olson. (2001): Report on short tem mission on Design and Developing A Web Site, Kingdom of Lesotho, Bureau of Statistics, LESSTAT: 2001:17.

World Bank (1991): The SDA Survey Instrument- An Instrument to capture Social Dimensions of Adjustment, Poverty and Social Policy Division Technical Department Africa Division.

United Nations (1982): National Household Survey Capability Programme. Survey Data Processing: A Review of Issues and Procedures. New York, UN.

_____ (1985): National Household Survey Capability Programme. Household Income an Expenditure Surveys: A technical study. New York, UN.

Yansaneh I.S (2003): An Overview of Sample Design Issues Household Surveys in Developing Countries, United Nations Statistics Division, 2003

Annex 1. Software options for different steps of survey data processing

Type of operation	Software options
Database management system	Sybase SQL server Access SAS Etc.
Data entry and editing	Visual Basic Access CSPro LSD 2000 Etc.
Data retrieval	SAS SPSS Access Excel Etc.
Tabulation, analysis, and presentation	MS-Word Excel SAS SPSS Etc.
Variance estimation	STATA SUDAAN SAS PC-CARP WESVAR Etc.

Annex 2. Documentation Structure for a survey and its production system

0 Documentation template 0.0 Documentation template 0.1 Survey name and identification, organization and staff responsible 0.2 Documentation modules and subsystems 0.3 Archived data material and published statistics 0.4 Reference to other relevant documentation	**1 Documentation template** 1.1 Domain of interest and target domain, verbal description 1.2 Target domain, formal description 1.2.1 Target objects, description and object graph 1.2.2 Target populations 1.2.3 Target variables 1.3 Survey output 1.3.1 Structured overview of the tabulation plan 1.3.2 Publications in printed form 1.3.3 Electronic Dissemination 1.3.4 Data base storage
2 Survey Plan 2.1 Frame procedure and observation objects 2.1.1 Overview 2.1.2 Frame and its links to objects 2.1.3 Frame production 2.1.4 Over-coverage and under-coverage 2.2 Sampling procedures 2.3 Data collection procedures 2.3.1 Observation objects, description and object graph 2.3.2 Data sources, including contact procedures 2.3.3 Observation variables and measurement instruments 2.3.4 Interruptions (including measures for over-overage) 2.3.3 Treatment of non-response 2.4 Planned data processing (coding, data entry, editing and correction) 2.5 Planned observation register/file 2.5.1 Overview 2.5.2 Object types, including derived object types 2.5.3 Object graph 2.5.4 Object/variable-matrices, including derived variables 2.5.5 Data set's descriptions 2.5.3 Derivation procedures (if complicated)	3 -

4 Statistical processing and presentation	**5 Data processing system**
4.1 Observation models	5.0 System summary and time frame
4.1.1 Sample	5.0.1 Verbal Description
4.1.2 Non-response	5.0.2 System flow
4.1.3 Measurement/observation	
4.1.4 Frame coverage	5.1 Description of subsystems
4.1.5 Total model	5.1.1 Overview
4.2 Population models	5.1.1.1 Verbal description
4.3 Computation formulas for estimation	5.1.1.2 System flow
4.3.1 Point estimations	5.1.2 Component descriptions
4.3.2 Estimates of sampling errors (variance estimations)	5.1.2.1 Data sets
4.3.3 Estimation/judgment of other quality characteristics	5.1.2.2 Processes
4.4 Analysis	5.1.2.3 Other components
4.5 Presentation and dissemination procedures	
6 The Log	