# Introduction

**Graham Kalton**
Westat
Rockville, Maryland
United States of America

1. When the data for a survey have been collected, they need to be prepared for analysis. This step has three important components. First, as will be discussed in chapter XV, decisions need to be made on how to format the data most effectively for analysis, taking account of the computing facilities available and the analysis software to be used. The survey analyses often involve two or more different units of analysis: in particular, households and persons are separate units of analysis in many household surveys. The data file therefore needs to be able to handle hierarchic structures efficiently; for example, it needs to cater for the facts that persons are nested within households and that the number of persons varies between households.

2. The second component of data preparation is data cleaning or editing. Inevitably, the survey responses will contain identifiable errors of various forms, for example, responses that are inconsistent with other responses or that fall outside the range of possibilities. These errors need to be resolved before analyses can start (see chap. XV for details on data cleaning).

3. An important task in data cleaning is to finalize the analytic status of each sampled unit. All of the units selected for the sample need to be placed in one of the following categories: respondent, eligible non-respondent, ineligible unit, or non-responding unit of unknown eligibility (see chap. VIII). A classification as respondent generally requires more than just the presence of a questionnaire for the sampled unit. Usually a minimum amount of acceptable data has to be collected for the unit to be so classified. The assignment of response status thus necessitates a review of the questionnaires. Note, however, that even though a unit is to be retained in the analysis as a respondent, there may well be some items for which acceptable answers have not been obtained. To cope with this problem, some form of imputation method may be used to assign values for the missing responses.

4. The analytic statuses of all sampled units are required for the last step of data preparation: the computation of survey weights. Survey weights are computed for each of the units of analysis. Since the starting point in the construction of weights is to determine the selection probabilities be all the sampled units, it is vitally important that careful records of the selection probabilities be kept in the sample selection process. The initial, or base, weights for sampled units are computed as the inverses of the units' selection probabilities. The base weights for respondents are then adjusted to compensate for the eligible non-respondents and for a proportion of the non-respondents with unknown eligibility status. A further adjustment is often applied to make the adjusted weighted sample distributions for certain key variables conform to known distributions of these variables available from an external source. The development of weights is described in chapters XV and XIX.

5. An important responsibility of data preparation is to ensure that the sampling information required for analysis is recorded on each respondent data record. Survey weights are needed for each responding unit of analysis in order that valid estimates of parameters of the survey population may be produced. Information on each responding unit's PSU and stratum is needed in order that sampling errors may be computed correctly for the survey estimates (see chap. XXI).

6. Two considerations distinguish analyses of survey data from the analyses described in standard statistical texts. One is the need to use survey weights in survey analyses in order to compensate for unequal selection probabilities, non-response, and non-coverage. Failure to use weights in the analyses may well result in distorted estimates of population values.

7. The second distinguishing consideration of survey analyses is the need to compute sampling errors for survey estimates in a way that takes account of the survey's complex sample design. The theory presented in standard statistical texts in effect assumes unrestricted sampling, whereas most household surveys employ stratified multistage sampling. In general, sampling errors for estimates from a stratified multistage sample are larger that those from an unrestricted sample of the same size, so that the application of the formulas in standard statistical texts will overstate the precision of the estimates (see chaps. VI, VII and XXI). This implies that standard statistical software packages produce invalid standard error estimates for survey estimates. Fortunately, however, there are now a sizeable number of survey analysis software packages that can be used to produce appropriate sampling error estimates from survey data obtained from complex sample designs. Chapter XXI contains a review of a number of these packages.

8. Much of the analysis conducted with government surveys is descriptive in nature. Often the results are reported in tabular form, with the table cells containing means, percentages or totals; sometimes, they are presented in graphical displays. In narrow statistical terms, the estimates involved are often very simple, the only issue being the need to make sure that the survey weights have been used. There are, however, important issues of definition and presentation to be considered. Careful attention needs to be given to defining the construct to be measured (for example, poverty: see chap. XVII), and to specifying the set of units for which it is to be measured, in suitable ways for the purpose in hand. Also, the results need to be presented in a fashion that clearly communicates what has been measured and for which set of units. Guidance on the presentation of simple descriptive estimates is given in chapter XVI.

9. Often, the construct to be measured can be defined in a relatively straightforward logical manner in terms of the survey responses. Sometimes, however, the construct is more complex and it may need to be measured by creating an index using multivariate statistical methods, such as cluster analysis and principal component analysis. Several examples are provided in chapter XVIII, including, for instance, one in which a "wealth" index was constructed using information on such variables as whether the household had electricity, the number of persons per sleeping room, and the principal type of drinking water.

10. Finally, it should be noted that, while the production of descriptive estimates remains the main form of survey analysis, there is increasing use of analytic techniques with survey data. These techniques are often applied to examine the relationships between variables and to explore

possible cause-effect relationships. The most common form of this type of analysis is one in which a statistical model is constructed to best predict a dependent variable in terms of a set of independent (or predictor) variables. If the dependent variable is a continuous one (for example, household income), then multiple linear regression methods may be used. If it is a categorical variable with a binary response (for example, whether the household has or does not have running water), then logistic regression methods may be used. These methods, and the effects of the complex sample design on them, are described in chapters XIX and XX. Chapter XIX also describes the use of multilevel modelling in a survey context and chapter XX also discusses the effect of complex sample designs on standard chi-square tests of the associations between categorical variables.