

# Introduction

## General Outline

This book deals with *sample surveys* that can be conceptually divided into two broad categories. In *descriptive surveys*, certain, usually few, population characteristics need to be precisely and efficiently estimated. For example, in a business survey, the average salaries for different occupational groups are to be estimated on the basis of a sample of business establishments. *Statistical efficiency* of the sampling design is of great importance. *Stratification* and other means of using *auxiliary information*, such as the sizes of the establishments, can be beneficial in sampling and estimation with respect to efficiency. Inference in descriptive surveys concerns exclusively a fixed population, although superpopulation and other models are often used in the estimation.

*Analytical surveys*, on the other hand, are often multi-purpose so that a variety of subject matters are covered. In the construction of a sampling design for an analytical survey, a feasible overall balance between statistical efficiency and *cost efficiency* is sought. For example, in a survey where personal interviews are to be carried out, a sampling design can include several stages so that in the final stage all the members in a sample household are interviewed. While this kind of clustering decreases statistical efficiency, it often provides the most practical and economical method for data collection. Cost efficiency can be good, but gains from stratification and from the use of other auxiliary information can be of minor concern for statistical efficiency when dealing with many diverse variables. Although in analytical surveys descriptive goals can still be important, of interest are often, for example, differences of subpopulation means and proportions, or coefficients of logit and linear models, rather than totals or means for the fixed population as in descriptive surveys. Statistical testing and modelling therefore play more important roles in analytical surveys than in descriptive surveys.

Both descriptive and analytical surveys can be *complex*, e.g. involving a complex sampling design such as multi-stage stratified cluster sampling. Accounting for the sampling complexities is essential for reliable estimation and analysis in both types of surveys. This holds especially for the clustering effect, which involves *intra-cluster correlation* of the study variables. This affects variance estimation

and testing and modelling procedures. And if *unequal selection probabilities* of the population elements are used, appropriate weighting is necessary in order to attain estimators with desired statistical properties such as unbiasedness or consistency with respect to the sampling design. Moreover, *element weighting* may also be necessary for adjusting for *nonresponse*, and *imputation* for *missing variable values* may be needed, in both descriptive and analytical surveys.

Thus, there are many common features in the two types of complex surveys and often, in practice, no real difference exists between them. A survey primarily aimed at descriptive purposes can also involve features of an analytical survey and vice versa. However, making the conceptual separation can be informative, and is a prime intention behind the structuring of the material in this book.

## **Topics Covered**

To be useful, a book on methods for both design and analysis of complex surveys should cover topics on sampling, estimation, testing and modelling procedures. We have structured a survey process so that we first consider the principles and techniques for sample selection. The corresponding estimators for the unknown population parameters, and the related standard error estimators, are also examined so that estimation under a given sampling design can be manageable in practice, reliable and efficient. These topics are considered in the first part of the book (Chapters 2 and 3), mainly under the framework of descriptive surveys.

Estimation and analysis specific to analytical surveys is considered in the second part of the book (Chapters 5, 7 and 8). For complex analytical surveys, more sophisticated techniques of variance estimation are needed. Our main focus in such surveys, however, is on testing and modelling procedures. Testing procedures for one-way and two-way tables, and multivariate analysis (including methods for categorical data and logistic and linear regression) are selected because of their importance in survey analysis practice. Topics relevant to both descriptive and analytical surveys, concerning techniques for handling nonsampling errors such as *reweighting* and *imputation*, are placed between the two main parts of the book (Chapter 4). Chapter 6 discusses domain estimation also being relevant to both survey types although the main concern is in descriptive surveys.

Fully worked examples and case studies taken from real surveys on health and social sciences and from official statistics are used to illustrate the various methods. Finally (Chapter 9), additional case studies are presented covering a range of different topics such as travel surveys, business surveys, socioeconomic surveys and educational surveys. We use a total of seven different survey data sets in the examples and case studies. A summary of the survey data sets, with selected technical information, is given in Table 1.1. Three types of survey data are included in the table. The aggregate-level census data set (1) (source: Official Statistics) is used in Chapters 2 to 4 to illustrate sampling and estimation for descriptive surveys. The real survey data sets (2) (source: National Public Health

**Table 1.1** Real survey data sets used in examples and case studies.

| Name of survey  | Type of primary sampling unit<br>PSU | Number of strata, clusters and elements in the survey data set |                                     |                   |
|---|--------------------------------------|--|-------------------------------------|-------------------|
|   |                                      | Strata   | Clusters (PSU:s)                    | Elements          |
| <b>Census register data set</b>   |                                      |  |                                     |                   |
| (1) <i>Province'91</i> Population (data for one province)                             | Municipality                         | 2  | 8 regional groups of municipalities | 32 municipalities |
| <b>Real survey data sets adjusted for pedagogical use</b>                             |                                      |  |                                     |                   |
| (2) Mini-Finland Health Survey (data for males aged 30–64 years)                      | Municipality                         | 24   | 48 municipalities                   | 2699 persons      |
| (3) Occupational Health Care Survey (data for establishments with 10 workers or more) | Industrial establishment             | 5  | 250 industrial establishments       | 7841 employees    |
| <b>Real survey data sets used in case studies</b>                                     |                                      |  |                                     |                   |
| (4) Passenger Transport Survey  | Person                               | 25   | (Element-level sampling)            | 11711 persons     |
| (5) Wages Survey  | Business firm                        | 25   | 744 firms                           | 13 987 employees  |
| (6) Health Security Survey (data for one stratum)                                     | Household                            | 1  | 878 households                      | 2071 persons      |
| (7) PISA 2000 Survey (data for 7 countries)   | School                               | 7  | 1388 schools                        | 32101 pupils      |

Institute) and (3) (source: Social Insurance Institution of Finland) are used in Chapters 5 to 8 for worked examples on domain estimation, variance estimation and multivariate modelling in complex analytical surveys. The real survey data sets (4) to (7) (sources: Ministry of Traffic and Communications; Statistics Finland; Social Insurance Institution of Finland; OECD's PISA International Database, respectively) are used in further case studies presented in Chapter 9.

To fully benefit the practical orientation of the book, the reader is encouraged to consult the web extension where the empirical examples and case studies are worked out in more detail. There, the accompanying program codes and datasets can be downloaded for further interactive training.

In Chapters 2 and 3, the basic and more advanced sampling techniques, namely, *simple random sampling*, *systematic sampling*, *sampling with probability proportional to size*, *stratified sampling* and *cluster sampling* are examined for the estimation of three different population parameters. These parameters are the *population total*, *ratio* and *median*. The estimators of these parameters provide examples of linear, nonlinear and robust estimators respectively. A small fixed population is used throughout to illustrate the estimation methods, where the main focus is on the derivation of appropriate *sampling weights* under each sampling technique. Special efforts are made in comparing the relative performances of the estimators (in terms of their standard errors) and the available information on the structure of the population is increasingly utilized. The use of such *auxiliary information* is considered for two purposes: the sampling design and the estimation of parameters for a given sampling design. The use of this information varies between different

sampling techniques, being minor in the basic techniques and more important and sophisticated in others, such as in stratified sampling and in cluster sampling. *Estimation using poststratification, ratio estimation and regression estimation* are considered in some detail under the framework of *model-assisted estimation*. The *design effect* is extensively used for efficiency comparisons. It is shown that proper use of auxiliary information can considerably increase the efficiency of estimation. Statistical properties of the total, ratio and median estimators, such as bias and consistency, are also examined by Monte Carlo simulation techniques. This treatment is extended in the web extension, where the behaviour of the estimators can be examined under various sampling designs.

In Chapter 5, we extend the variance estimation methodology of Chapters 2 and 3 by introducing additional (approximative) techniques for variance estimation. Subpopulation means and proportions are chosen to illustrate ratio-type estimators commonly used in analytical surveys. The *linearization method* and *sample reuse techniques* including *balanced half-samples*, *jackknife* and *bootstrap* are demonstrated for a two-stage stratified cluster sampling design taken from the Mini-Finland Health Survey. This survey is chosen because it represents an example of a realistic but manageable design. Approximation of variances and covariances of several ratio estimators is needed for testing and modelling procedures. Using the linearization method, various sampling complexities including clustering, stratification and weighting are accounted to obtain consistent variance and covariance estimates. These approximations are applied to the Occupational Health Care Survey sampling design, which is slightly more complex than that of the previous survey. Chapter 6 addresses the estimation of totals for domains, which are subpopulations constructed on regional or similar criteria. Design-based model-assisted techniques are introduced and illustrated using data from the Occupational Health Care Survey.

The analysis of complex survey data is considered in Chapters 7 and 8. For testing procedures of goodness of fit, homogeneity and independence hypotheses in one-way and two-way tables, we introduce two main approaches, the first of these using *Wald-type test statistics* and the second, *Rao–Scott-type adjustments* to standard Pearson and Neyman test statistics. The main aim in these test statistics is to adjust for the clustering effect. These testing procedures rely on the assumption of an asymptotic chi-square distribution of the test statistic with appropriate degrees of freedom; this assumption presupposes a large sample and especially a large number of sample clusters. For designs where only a small number of sample clusters are available, certain degrees-of-freedom corrections to the test statistics are derived, leading to *F*-distributed test statistics.

In Chapter 8, we turn to *multivariate survey analysis*, where a binary or a continuous response variable and a set of predictor variables are assumed. In the analysis of categorical data with logit and linear models, *generalized weighted least squares estimation* is used. Further, for logistic and linear regression in cases in which some of the predictors are continuous, we use the *pseudo-likelihood* and *generalized estimating equations (GEE) methods*. For proper analysis using either of

these methods, certain analysis options are suggested. Under the full design-based option, all the sampling complexities are properly accounted for, thus providing a generally valid approach for complex surveys. The options based on an assumption of simple random sampling are used as references when measuring the effects of weighting, stratification and clustering on estimation and test results. Using these options, multivariate analysis is further demonstrated in the additional case studies in Chapter 9.

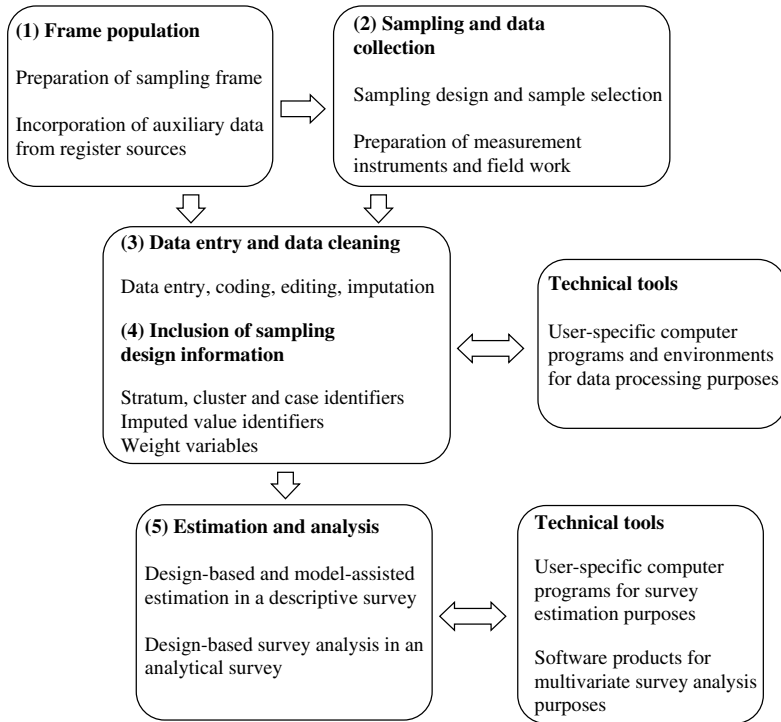
The *nuisance (or aggregated) approach*, where the clustering effects are regarded as disturbances to estimation and testing, is the main approach for the *design-based* analysis in this book. In this approach, the main aim is to eliminate these effects to obtain valid analysis results. In the alternative *disaggregated approach*, which also provides valid analyses, clustering effects are themselves of intrinsic interest. We demonstrate this approach for *multi-level modelling* of hierarchically structured data in the last of the additional case studies in Chapter 9.

## Computation

In the design of a survey, whether descriptive or analytical, the various phases of the so-called *total survey process* should be carefully worked out. Typically, a survey process starts with a problem-setting phase arising from an actual information need. An overall plan of the survey will be prepared, including sampling, measurement and analysis designs as phases in which statistical and survey methodologies are obviously needed. In the course of the implementation of the survey, the plan will be evaluated and made operational. Finally, the results will be disseminated. In the total survey process, a number of statistical operations relevant to this book can be identified. These are illustrated in Figure 1.1, where the necessary methodologies and technical tools are referred to.

A computerized frame population, prepared in phase (1), serves as a basis for the sample selection in phase (2). The frame population includes usually auxiliary information on all population elements. The auxiliary data can be taken from various sources, such as a population census and different administrative registers. These data are assumed to be merged on a micro level (this is often possible in practice e.g. by using the element identification keys that are unique in all the data sources). The collected data are cleaned in phase (3), where also selected auxiliary data from the frame population can be incorporated, to be used in estimation and analysis phases. In the data processing phase (4), the sampling design identifiers are included in the cleaned survey data set to be analysed in phase (5). Thus, the auxiliary data can be used in two phases: to construct an efficient sampling design, and to improve the efficiency for a given sample by model-assisted estimation techniques. Both of these phases are discussed extensively in this book. Usually in practice, user-specific computer programs are used in phases (1) to (4). In phase (5), both standard survey estimation and analysis software packages and user-specific solutions can be used.

To be manageable in practice, we have in the examples and case studies demonstrated the methodology and computational tools using commercially available



**Figure 1.1** Flow chart for design-based estimation and analysis of complex survey data.

software products for data processing and survey estimation and analysis. A more technical treatment of the methodologies and computational tools is included in the web extension of the book.

## Use of the Book

This book is primarily intended for researchers, sample survey designers and statistics consultants working on the planning, execution or analysis of descriptive or analytical sample surveys. We have aimed to supply such workers with an applied source covering in a compact form the relevant topics of recent methodology for the design and analysis of complex surveys. By using real data sets with computing instructions and computerized examples, the reader can also be led to a deeper understanding of the methodology. In this effort, the reader is encouraged to consult the web extension of the book. In the web environment, many of the empirical examples are extended and worked out in more detail. An option for further training is provided, including the possibility to download program codes and real data sets for interactive analysis in the user's personal computing environment.

The material in the book can also be used in university-level methodological courses. A first course in survey sampling can be based on Chapters 2 to 4 where

the students can also be guided to real sampling and estimation using the small population provided. A more advanced course can be based on Chapters 5 to 8. In both types of courses, the web extension can be used to support the teaching and learning. Also, useful data sets are supplied in the web extension for practising variance approximation, testing procedures and multivariate analysis in complex surveys. Chapter 4 might be included in a more advanced course. Chapter 6 might serve as material for a course on estimation for domains.

