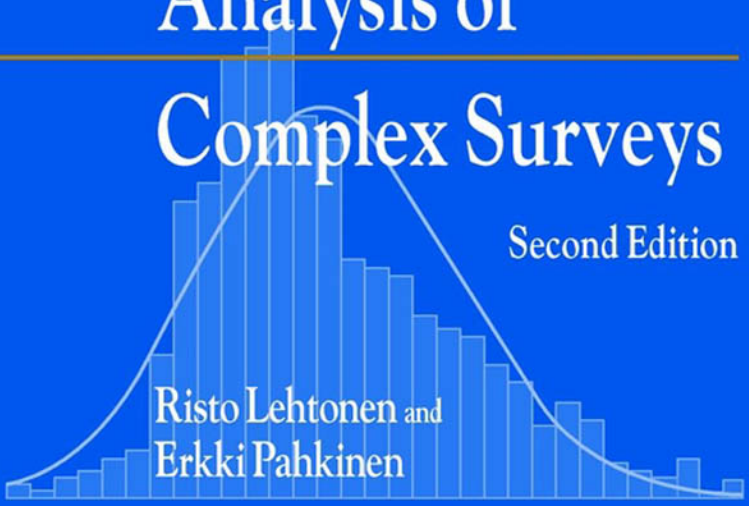
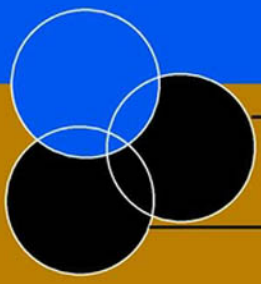
WILEY

Practical Methods for Design and Analysis of Complex Surveys

Second Edition

Risto Lehtonen and
Erkki Pahkinen





STATISTICS IN PRACTICE

Practical Methods for Design and Analysis of Complex Surveys

Second Edition

Risto Lehtonen and Erkki Pahkinen
Department of Mathematics and Statistics, University of Jyväskylä, Finland

John Wiley & Sons, Ltd

Practical Methods for Design and Analysis of Complex Surveys

Statistics in Practice

Founding Editor

Vic Barnett
Nottingham Trent University, UK

Statistics in Practice is an important international series of texts, which provide detailed coverage of statistical concepts, methods and worked case studies in specific fields of investigation and study.

With sound motivation and many worked practical examples, the books show in down-to-earth terms how to select and use an appropriate range of statistical techniques in a particular practical field within each title's special topic area.

The books provide statistical support for professionals and research workers across a range of employment fields and research environments. The subject areas covered include medicine and pharmaceutics; industry, finance and commerce; public services; the earth and environmental sciences, and so on.

The books also provide support to students studying statistical courses applied to the above areas. The demand for graduates to be equipped for the work environment has led to such courses becoming increasingly prevalent at universities and colleges.

It is our aim to present judiciously chosen and well-written workbooks to meet everyday practical needs. The feedback of views from readers will be most valuable to monitor the success of this aim.

A complete list of titles in this series appears at the end of the volume.

Practical Methods for Design and Analysis of Complex Surveys

Second Edition

Risto Lehtonen and Erkki Pahkinen
Department of Mathematics and Statistics, University of Jyväskylä, Finland

John Wiley & Sons, Ltd

John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester,
West Sussex PO19 8SQ, England

Telephone (+44) 1243 779777

Email (for orders and customer service enquiries): cs-books@wiley.co.uk
Visit our Home Page on www.wileyeurope.com or www.wiley.com

Other Wiley Editorial Offices

John Wiley & Sons Inc., 111 River Street, Hoboken, NJ 07030, USA

Jossey-Bass, 989 Market Street, San Francisco, CA 94103-1741, USA

Wiley-VCH Verlag GmbH, Boschstr. 12, D-69469 Weinheim, Germany

John Wiley & Sons Australia Ltd, 33 Park Road, Milton, Queensland 4064, Australia

John Wiley & Sons (Asia) Pte Ltd, 2 Clementi Loop #02-01, Jin Xing Distripark, Singapore 129809

John Wiley & Sons Canada Ltd, 22 Worcester Road, Etobicoke, Ontario, Canada M9W 1L1

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic books or in possible web extensions of books.

Library of Congress Cataloging-in-Publication Data

Lehtonen, Risto.
Practical methods for design and analysis of complex surveys / Risto Lehtonen and Erkki Pahkinen.—2nd ed.
p. cm.—(Statistics in practice)
Includes bibliographical references and index.
ISBN 0-470-84769-7 (alk. paper)
1. Sampling (Statistics) 2. Surveys—Methodology. I. Pahkinen, Erkki. II. Title. III. Statistics in practice (Chichester, England)

QA276.6.L46 2004
001.4′33—dc21

2003053783

British Library Cataloguing in Publication Data

A catalogue record for this book is available from the British Library

ISBN 0-470-84769-7

Typeset in 10/12pt Photina by Laserwords Private Limited, Chennai, India
Printed and bound in Great Britain by Biddles Ltd, Guildford, Surrey
This book is printed on acid-free paper responsibly manufactured from sustainable forestry in which at least two trees are planted for each one used for paper production.

Contents

Preface

Our main goals in updating the materials of *Practical Methods for Design and Analysis of Complex Surveys*, published in 1995, for a second edition have been well-focused extension of coverage, improved usability and meeting user feedback. As examples of extension, model-assisted estimation now covers a chapter on estimation for domains. The chapter on handling nonsampling errors has been completely re-written. More sophisticated estimation techniques have been included in analysis methods for complex surveys. We have extended the chapter of case studies. Practical methods for quality monitoring of survey processes are now illustrated. A stronger aspect of international comparison is introduced by a case study on a multinational educational survey. We believe that with these and other extensions and enhancements, the book meets a wider spectrum of user needs.

An important change has taken place in computational aspects since the previous edition. We have inserted the technical materials into a web extension of the book. The web extension is aimed to improve the practical applicability of methods and to provide tools for teaching and training. Examples and case studies can be worked out in an interactive environment and program codes, real data sets and other supporting materials can be downloaded. For us, this gives an option to flexibly update the technical materials when appropriate.

We greatly appreciate the support given by organizations when writing the manuscript. In particular, we would like to mention the Institute for Educational Research, University of Jyväskylä; Ministry of Transport and Communications, Finland; National Public Health Institute, Finland; the Social Insurance Institution of Finland; Statistics Finland and the University of Jyväskylä. Chief Statistical Analyst Antero Malin has produced materials for the case study on a multinational educational survey and Senior Consultant Virpi Pastinen for the case study on quality monitoring of survey processes. We are very grateful for these contributions.

Detailed comments given by Professor Carl-Erik Särndal on several parts of the book have been very valuable. Dr Juha Lappi has given helpful comments on a part of the book. Thanks are also due to Vesa Kiviniemi, a doctoral student in statistics, and Antti Pasanen, a graduate student in statistics, for their technical work in building the web extension and to Elina Nykyri, a graduate student in

statistics, who has assisted us in proofreading and similar final-phase tasks. We are thankful to anonymous referees for comments on our proposal for the second edition. Last but not the least, we are grateful to the staff of Wiley & Sons for their patience and flexibility.

Jyväskylä, September 2003

Risto Lehtonen Erkki Pahkinen

1

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

Practical Methods for Design and Analysis of Complex Surveys Risto Lehtonen and Erkki Pahkinen
 ISBN: 0-470-84769-7

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 PSU	Number of strata, clusters and elements in the survey data set: Strata	Clusters (PSU:s)	Elements
Census register data set				
(1) *Province'91* Population (data for one province)	Municipality	2	8 regional groups of municipalities	32 municipalities
Real survey data sets adjusted for pedagogical use				
(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
Real survey data sets used in case studies				
(4) Passenger Transport Survey	Person	25	(Element-level sampling)	11 711 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	32 101 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

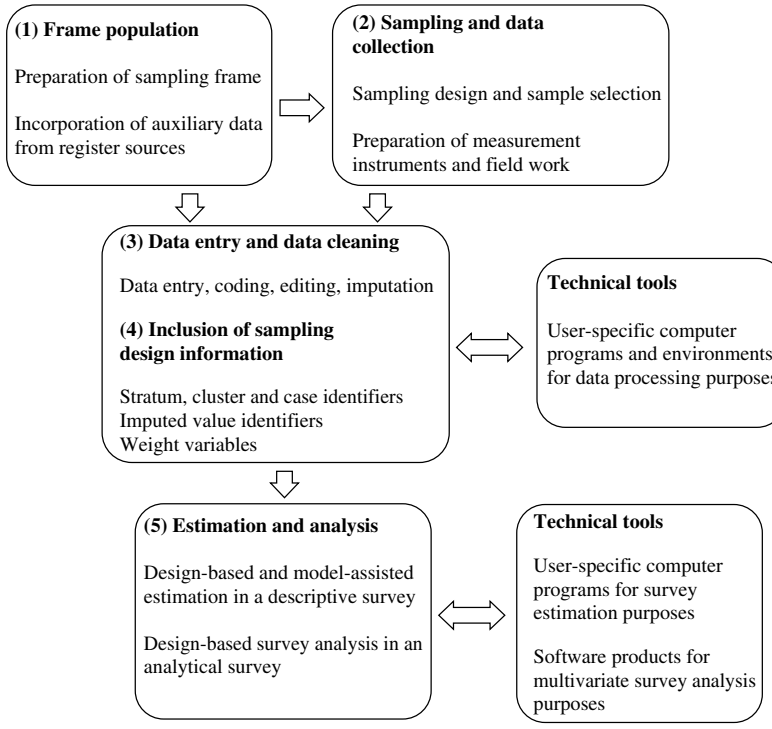


Figure 1.1 Flow chart for design-based estimation and analysis of complex survey data.

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.

2

Basic Sampling Techniques

Simple random sampling, systematic sampling and *sampling with probability proportional to size* are introduced as the basic sampling techniques in this chapter. We start with a discussion of sampling, and sampling errors, and estimation of a given sampling scheme. Definitions of some key concepts are given.

Sampling and Sampling Error

In survey sampling, a *fixed finite population* is under consideration, where the population elements are *labelled* so that each element can be identified. *Probability sampling* provides a flexible device for the selection of a random sample, or a *sample* for short, from such a fixed population. A key property of probability sampling is that for each population element a positive probability of selection is assigned; this probability need not be equal for all the elements. A specific sampling scheme is used in drawing the sample. The term *sampling scheme* refers to the collection of techniques or rules for the selection of the sample. The composition of the sample is thus randomized according to the probabilistic definition of the sampling scheme.

In principle, a large number of different samples could be drawn from a population using a particular sampling scheme. Depending on which specific population elements happen to be drawn, different numerical estimates are obtained from the sample for an unknown population parameter such as a *total*, i.e. the sum of the population values of a variable. *Sampling error* describes the variation of the estimates calculated from the possible samples. In the design of the sample-selection procedure for a specific survey, a sampling scheme is desired under which the sampling error would be as small as possible. In order to attain this goal, knowledge on the structure of the population can be helpful. Relationships between the sampling scheme and the structure of the population are considered for various specific sampling situations in this chapter and in Chapter 3. In this discussion, the *standard error* of an unbiased estimate is used as

Practical Methods for Design and Analysis of Complex Surveys Risto Lehtonen and Erkki Pahkinen
 ISBN: 0-470-84769-7

a measure of the sampling error, and the comparison of the sampling errors under various sampling schemes is carried out using the *design-effect* statistic.

Estimation from Selected Sample

When an actual sample is drawn using a specific sampling scheme, measurements are recorded from the sampled elements for some variable of interest, called a *study variable*. After data collection, statistical analyses can be carried out. For example, an estimate of the population total of the study variable and its estimated standard error are frequently calculated. In this chapter and the next, we examine practical methods for designing manageable sampling procedures and for carrying out proper estimation under a given sampling scheme. For this, let us first discuss various approaches concerning the role of the sampling scheme in the estimation process.

When a survey is analysed in practice, it is emphasized that the estimation should take into account the structure of the sampling scheme. To accomplish this, the analysis is carried out using the so-called *design-based approach*. An essential property of the design-based approach is that any of the complexities due to the sampling scheme can be properly accounted for in the estimation. These complexities can arise, for example, when elements have unequal selection probabilities; this will be discussed further in this chapter and Chapter 3. These features of a sampling scheme can be incorporated into the estimation in the design-based approach because a fixed finite population with labelled elements is being considered. By using the labels assigned to each element, appropriate *sampling design identifiers* can be included in the sample data set and used in the analysis. Making use of the sampling identifiers is examined in some detail in this and the next chapter, for estimation under various sampling schemes.

An analysis ignoring all the sampling complexities is used often in this book as a reference to the design-based analysis. Especially, a certain sampling situation, namely where elements are selected with equal probabilities and are replaced in the population after each draw is called *simple random sampling with replacement* and will occasionally be used as a reference design when comparing the efficiencies of more complex sampling schemes. In the design-based approach, it can sometimes be useful to assume that the finite population is a realization from some hypothetical *superpopulation*. This assumption together with appropriate *auxiliary information* can be used by postulating models for the estimation of parameters of the finite population under consideration. When auxiliary variables are incorporated in the estimation procedure by using a model, but the inference is still design-based, we call this the *design-based model-assisted approach*, or more simply the *model-assisted approach* (Särndal *et al*. 1992). This approach is introduced in the last part of Chapter 3 and applied further in Chapter 6.

Let us consider the design-based model-assisted approach more closely to show how a model assumption can be used to simplify the estimation for a

certain sampling scheme. Suppose that a shipping company wants to know the approximate total weight of the passengers on a ferry. This piece of information is important for future planning. Weighing all the passengers would be too expensive and time-consuming, thus sampling would be more appropriate in this context. Suppose, therefore, that every tenth passenger is weighed. This yields a sample data set of n passengers denoted by $y_1, \ldots, y_k, \ldots, y_n$. The researcher is faced with the problem of estimating the total weight of passengers using the *sample observation*, and moreover, of evaluating the precision of the estimate.

In estimating the total weight of passengers, the researcher notes that the sample was drawn from a specific finite population using a particular sampling scheme. Obviously, *systematic sampling* was used, and from the passenger register, the total number of passengers on board, N, would be known as an additional information. An *estimator* for the total weight is easily defined in the form $\hat{t} = N\overline{y}$, where $\overline{y} = \sum_{k=1}^{n} y_k/n$ is the sample mean of the n passenger weights. To assess the sampling error, the standard error of $\hat{t}$ should be estimated as the square root of the variance estimate $\hat{v}(\hat{t})$. To estimate $\hat{v}(\hat{t})$, the researcher uses the textbook variance estimator $\hat{v}_{srs}(\hat{t}) = N^2(1 - n/N)\hat{s}^2/n$, which is for *simple random sampling without replacement*, where $\hat{s}^2 = \sum_{k=1}^{n}(y_k - \overline{y})^2/(n-1)$ is the sample variance of the passenger weights.

The estimates obtained using the above formulae would usually be adequate for practical purposes. But it is instructive to progress further and examine the present estimation problem more closely. Actually, the researcher made a procedure-simplifying assumption when estimating the variance of $\hat{t}$ as an estimator from simple random sampling. In fact, the variance formula for systematic sampling would be more complex, because another design parameter, the *intra-class correlation* ρ_{int}, should be included. The two variance estimators are related by $\hat{v}_{sys}(\hat{t}) = \hat{v}_{srs}(\hat{t})[1 + (n-1)\hat{\rho}_{int}]$, where $\hat{v}_{sys}$ is the variance estimator under systematic sampling.

Unfortunately, the variance estimator $\hat{v}_{sys}(\hat{t})$ is not suitable for practical purposes, since only one element is drawn into the sample from each sampling interval. Therefore, an estimate $\hat{\rho}_{int}$ cannot be obtained from the selected sample without having auxiliary information on the order in which the passengers step on board, or without making a simplifying model assumption for the process of boarding.

The simplest model assumption would be that the passengers step on board in a completely random order. In this case the intra-class correlation would be zero. Then, the variance of $\hat{t}$ estimated from systematic sampling would coincide with that from simple random sampling. By using this simplifying model assumption, we thus implicitly make use of auxiliary information in the design-based analysis, in the form of a superpopulation assumption. For systematic sampling, the alternative ways of making use of auxiliary information, or a model assumption, are examined in Section 2.4. There, it will be shown that proper use of auxiliary information not only simplifies the estimation but can also make the estimation more efficient.

In this and the next chapter, five different sampling techniques are introduced and selected population parameters are estimated with corresponding standard errors under the design-based approach. It will become evident that it is essential to derive appropriate *element weights* w_k specific to each sampling scheme. In the example above, the weights would be equal to N/n for all passengers, i.e. the inverse of the probability of selecting a passenger in the sample. This weight derivation holds, for example, for both simple random and systematic sampling; for more complex schemes, the weights are not necessarily equal for all elements. The estimators and standard error estimators are derived for a given sampling scheme so that the correct weights are incorporated into the equations. Moreover, it will be pointed out to what extent, and how, auxiliary information available on the population can be used with a specific sampling scheme. In addition to the use of auxiliary information in sampling, such information will also be used for model-assisted estimators applied to a selected sample for reducing standard errors and to obtain estimates close to the corresponding population values. There, a new type of weight is derived called the *g weight* and denoted g_k. Its value depends on both the selected sample and the chosen model-assisted estimator.

2.1 BASIC DEFINITIONS

The formal framework and basic definitions are now given for Chapters 2 to 4, and the various sampling schemes are briefly described in relation to their use of auxiliary information.

Population and Variables

A *finite population* $\{u_1, \ldots, u_k, \ldots, u_N\}$ of N elements is considered with elements labelled from 1 to N. For simplicity, let the kth element of the population be represented by its label k, so that the finite population can be denoted by

$$U = \{1, \ldots, k, \ldots, N\}.$$

We denote by y the *study variable* with unknown population values $Y_1, \ldots, Y_k, \ldots, Y_N$. In some cases an additional study variable, x, and an auxiliary variable, z, are also used. The unknown population values of x are denoted by $X_1, \ldots, X_k, \ldots, X_N$. The *auxiliary variable* z represents additional information on the finite population and is usually assumed known for all the N population elements. The known population values of the auxiliary variable are denoted by $Z_1, \ldots, Z_k, \ldots, Z_N$.

Population Parameters

A *parameter* of the finite population U is a function of the population values Y_k of the study variable y; in some cases, the function includes population values X_k of

the study variable x. Typical parameters are the **total**, the **ratio** and the **median**. They are defined as follows:

$$\textit{Total } T = \sum_{k=1}^{N} Y_k = Y_1 + Y_2 + \cdots + Y_N$$

$$\textit{Ratio } R = T/T_x, \text{ where } T_x \text{ is the population total of the study variable } x$$

$$\textit{Median } M = F^{-1}(0.5), \text{ where } F \text{ is the population distribution function of } y.$$

The population total has been chosen because of its importance in survey sampling, most notably by descriptive surveys carried out by statistical agencies publishing *official statistics*. Much of the classical literature on survey sampling deals with the estimation of population totals. Because the population mean $\overline{Y}$ is a simple transformation of the total, i.e. $\overline{Y} = T/N$, the estimators presented below for totals are equally applicable to means with a few minor changes. Instead of the mean, the median is considered since it is often a more appropriate measure of location, as is the case for the demonstration data used later. The ratio is chosen as a more complicated parameter to estimate, and because it is frequently used in practice. Ratio-type estimators will be extensively used in the survey analyses considered in Chapters 5 to 9.

Sampling Design and Sample

The aim of a sample survey is to estimate the unknown population parameters T, R or M based on a sample from the population U. A *sample* is a subset of U. There are many different samples that could be drawn. We denote by S the set of all possible samples of size n ($n < N$) from the population U. The actual sample is denoted by $s = \{1, \ldots, k, \ldots, n\}$, so that s is one of the possible samples in the set S. To draw a sample from U a specific sample selection scheme is used. Under a sampling scheme it is possible to state the *selection probability* for a sample s. This probability is denoted as $p(s)$. Formally, the function $p(\cdot)$ is called a *sampling design*. The sampling design determines the statistical properties (expectation and sampling error) of random quantities such as the sample total, sample ratio and sample median calculated from the sample drawn under the actual sampling scheme. In what follows, we will use interchangeably the terms *sampling scheme* and *sampling design*, although somewhat different definitions have been given for these concepts in the literature. For the purpose of this book, the terms are taken to refer roughly to the way in which we draw a sample from the fixed population.

Under a fixed sampling design $p(\cdot)$, an *inclusion probability* is assigned for each population element to indicate the probability of inclusion of the element in the sample. For a population element k, the inclusion probability is denoted by π_k. It

is also called the *first-order* inclusion probability. Such inclusion probabilities will be used when we introduce the various sampling techniques.

A population element can appear more than once in a sample s if sampling involves *replacement* of the selected element in the population after each draw. Such a sampling design is of a *with-replacement*-type (WR). On the contrary, under *without-replacement*-type sampling (WOR), a population element can appear in a sample s only once. The with-replacement assumption simplifies the estimation under complex sampling designs and is often adopted, although in practice sampling is usually carried out under a without-replacement-type scheme. Obviously, the difference between with-replacement and without-replacement sampling becomes less important when the population size is large and the sample size is noticeably smaller than it.

The study variable y is measured for the elements belonging to the sample s. The n sample values of y are denoted by lower-case letters $y_1, \ldots, y_k, \ldots, y_n$. In some cases, as for the estimation of the ratio R, the data set also includes the measurements x_k, $k = 1, \ldots, n$, of a study variable x. We assume for simplicity that the measurements are free from measurement errors. In addition to the study variables, the data set should include appropriate information on the sampling design, i.e. the *design identifiers* such as stratum and cluster identifiers and a weight variable. An auxiliary variable z (or several such variables) are also often included in the data set. These variables are described in detail under each sampling technique to be introduced.

Estimator

An *estimator* of a population parameter refers to a specific computational formula or algorithm that is used to calculate the sample statistics for the selected sample. Estimators that are *unbiased* or *consistent* with respect to the sampling design are usually desired so that the expectation of an estimator equals, or approximates more closely, the population parameter, with increasing sample size n. The following three estimators will be considered:

$$\textit{Total}\ \hat{t} = \sum_{k=1}^{n} w_k y_k, \text{ where } w_k \text{ is the element weight}$$

$$\textit{Ratio}\ \hat{r} = \hat{t}/\hat{t}_x, \text{ where } \hat{t}_x \text{ is the estimated total of } x$$

$$\textit{Median}\ \hat{m} = \hat{F}^{-1}(0.5), \text{ where } \hat{F} \text{ is the estimated distribution function of } y$$

The observed numerical value obtained by using an estimator for the actual sample is called an *estimate*.

A combination of a sampling design $p(\cdot)$ and an estimator is a *strategy*. This concept will be used especially in the last part of Chapter 3 when discussing model-assisted estimation.

Variance of Estimator

The estimates for a population parameter vary from sample to sample. This variation due to sampling describes the uncertainty of inference based on a particular sample. The sample-to-sample variation is measured by the variance $V_{p(s)}$ of an estimator. Because $V_{p(s)}$ depends on the sampling design, it is also called the *design variance*. Its value can be estimated from the actual sample by using an appropriate *variance estimator*, which will be denoted by $\hat{v}_{p(s)}$. The square root of a variance estimator is the *estimated standard error* (s.e) of an estimator.

Strictly speaking, the design variance is only appropriate for unbiased estimators; for biased estimators, a more general measure of sampling error called the *mean squared error*, MSE, should be used. The MSE can be expressed as the sum of the design variance and the squared bias of an estimator, where the bias is the deviation of the expected value of an estimator from the corresponding parameter. Generally, in survey estimation, unbiased or approximately unbiased estimators are preferred, so that the use of design variances can be justified. This holds also for *consistent estimators* whose bias decreases with increasing sample size.

Design Effect

Different sampling designs use different design variances of an estimator of a population parameter. A convenient way to evaluate a sampling design is to compare the design variance of an estimator to the design variance from a references sampling scheme of the same (expected) sample size. Usually, simple random sampling with or without replacement is chosen as the reference. For example, for an estimator $\hat{t}$ of the total T, the ratio of the two design variances, called the *design effect* and abbreviated to DEFF, is defined by

$$\mathrm{DEFF}_{p(s)}(\hat{t}) = \frac{V_{p(s)}(\hat{t})}{V_{srs}(\hat{t})},$$

where $p(\cdot)$ refers to the actual sampling design. Obviously, obtaining a DEFF requires the values of both design variances. These are rarely available in practice. However, in some instances we will calculate such figures. In practice, an estimate of the design effect is calculated using the corresponding variance estimators for the sample data set. An estimator of the design effect is thus

$$\mathrm{deff}_{p(s)}(\hat{t}) = \frac{\hat{v}_{p(s)}(\hat{t})}{\hat{v}_{srs}(\hat{t})}.$$

More generally, the design effect can be defined for a strategy $\{p(\cdot), \hat{t}^*\}$, where $p(\cdot)$ denotes the sampling design and $\hat{t}^*$ denotes a specified estimator for the total T:

$$\mathrm{DEFF}_{p(s)}(\hat{t}^*) = \frac{V_{p(s)}(\hat{t}^*)}{V_{srs}(N\overline{y})},$$

where $\overline{y} = \sum_{k=1}^{n} y_k/n$ is the sample mean of y. In this DEFF, $\hat{t}^*$ is a design-based or model-assisted estimator of T under $p(s)$ and $N\overline{y} = \hat{t}$ is a design-based estimator under simple random sampling, and $V_{p(s)}$ and V_{srs} are the corresponding variances. For example, the estimator $\hat{t}^*$ of the total can be a regression estimator (see Section 3.3).

As a rule, a sampling design is equally as efficient as SRS if DEFF is equal to one, more efficient if DEFF is less than one and less efficient if DEFF is greater than one. The efficiencies of different sampling designs or strategies will be compared using a design-effect statistic based on either of the definitions given above.

Use of Auxiliary Information in Sampling and Estimation

A *sampling frame*, i.e. a list or register of the population elements from which the sample is drawn, often includes additional information on the population elements. Auxiliary information can also be taken from other sources such as administrative registers and official statistics. This *auxiliary information* can be useful in the construction of the sampling design and in improving the efficiency of the estimation for the actual sample. To be useful, auxiliary information should be related to the variation of the study variable.

The use of auxiliary information in the *sample selection phase* is as follows.

Simple random sampling (SRS) The sample is drawn without using auxiliary information on the population. Therefore, a simple random sampling scheme (with or without replacement) provides a reference when assessing the gain from the use of auxiliary information in more complex designs or in improving the estimation.

Systematic sampling (SYS) Auxiliary information is used in the form of the list order of population elements in the sampling frame. For example, if the values of the study variable increase with the list order, then systematic sampling appears to be more efficient than simple random sampling. *Intra-class correlation*, an additional design parameter in the design variance of an estimator, provides a measure of the correlation between list order and the values of the study variable.

Sampling with probability proportional to size (PPS) An auxiliary variable z is assumed to be a measure of the size of a population element. Varying inclusion probabilities can be assigned using this auxiliary variable. The magnitude of sampling error depends on the relationship between the study variable y and the auxiliary variable z.

Stratified sampling (STR) The population is first divided into non-overlapping subpopulations called *strata*, and sampling is executed independently within each stratum. The total sampling error is the sum of the stratum-wise sampling errors.

If a large share of the total variation of the study variable is captured by the variation between the strata, then stratified sampling can be more efficient than simple random sampling.

Cluster sampling (CLU) The population is assumed to be readily divided into naturally formed subgroups called *clusters*. A sample of clusters is drawn from the population of clusters. If the clusters are internally homogeneous, which is usually the case, then cluster sampling (CLU) is less efficient than simple random sampling. The *intra-cluster correlation coefficient* is the important design parameter in cluster sampling and it measures the internal homogeneity of the clusters.

These five sampling techniques can be used to construct a manageable sampling design for a complex sample survey, either using a particular method or more usually a combination of methods. In all the schemes, excluding simple random sampling, auxiliary information on the elements of the population is required. For the selected sample, auxiliary information can be used in the *estimation phase*. The general framework is *model-assisted estimation*. The use of auxiliary information during the estimation phase is as follows:

Poststratification The selected sample is divided into non-overlapping poststrata according to a categorical auxiliary variable, and the estimation follows that of stratified sampling. The assisting model is of an ANOVA type. Efficiency can improve if the poststrata are internally homogeneous.

Ratio estimation The population total of a continuous auxiliary variable z is assumed known. The assisting model is of regression-type (without an intercept term). Efficiency can improve if the study variable y and the auxiliary variable z are correlated.

Regression estimation As in ratio estimation, the population total of an auxiliary variable z is assumed known. The assisting model is of regression-type (with an intercept term). Here, also, efficiency can improve if y and z are correlated.

Thus, auxiliary information can be used in the construction of the sampling design and, for a given sample, to improve the efficiency. As a rule, sampling error can be decreased by the proper use of auxiliary information. Thus, it is worthwhile to make an effort to collect this type of data.

Further Reading

The main topic of this book is design-based survey estimation and analysis, especially methods to account for sampling design complexities in estimation and analysis phases. An early textbook written in a similar spirit is Kish (1965), where the design effect statistic is introduced and used for a variety of practical

applications. A practical orientation is adopted in Lohr (1999) in which design-based, model-assisted and model-based estimation are illustrated with examples taken from real surveys. A more mathematically oriented book by Särndal *et al.* (1992) covers the important areas of survey sampling under a sound theoretical and mathematical framework. For additional references, the reader is advised to consult the web extension of this book.

2.2 THE *PROVINCE'91* POPULATION

In practical survey sampling, we are interested in finite populations, which are limited in size. Indeed, real populations are generally very large as will be seen later in this book when practical survey samples are analysed. In the case of real surveys, it is not easy to see how sampling error arises and how the properties of the estimators depend on it. For this reason, we have chosen a more restricted problem and a small finite population in order to demonstrate different sampling schemes and their influence on sampling error. For example, the parameters total, ratio and median of the target population can be calculated exactly and compared with their estimates computed from the appropriate sample. This allows a view of the whole target population. This finite population consists of only 32 population elements from which a sample of fixed size of 8 units is drawn. It is immediately obvious that there is an enormous gap between this demonstration survey and a real large-scale sample survey. But the demonstration data set can help clarify such important concepts and issues as how to determine the sampling distribution and how a sampling design affects estimators and their design variances.

To illustrate the main ideas, a small data set under the title *Province'91* has been taken from the official statistics of Finland. This data set will be used as a sampling frame in Chapters 2 to 4. Finland is divided into 14 provinces from which one has been selected for demonstration. This province comprises 32 municipalities and had a total population of 254 584 inhabitants on 31 December 1991. The data set is presented in Table 2.1.

The *Province'91* population contains three kinds of information categorized according to their purpose throughout the survey process. The first phase is sampling design in which identification variables, such as labels, and the ability to identify important subgroups of the population such as strata and clusters, are needed. Here, as the population of elements are municipalities, the name or register number serves as an identifier of a population element. The other two types of information define the study and the auxiliary variables.

In the official statistics of Finland, municipalities are listed in alphabetical order with urban municipalities in the first group and rural municipalities in the second group. This gives a natural order for a certain sampling technique called *systematic sampling* and further, allows the population of municipalities to be divided into non-overlapping subpopulations called *strata*. Another type of population subgroup is formed by combining four neighbouring municipalities in

Table 2.1 The *Province'91* population. Percentage unemployment (%UE) and totals of unemployed persons (UE91), labour force (LAB91), population in 1991 (POP91) and number of households (HOU85) by municipality in the province of Central Finland in 1985.

ID	LABEL	STR	CLU	%UE	UE91	LAB91	POP91	HOU85
	Urban			**12.67**	**8022**	**63 314**	**129 460**	**49 842**
1	Jyväskylä	1	1	12.20	4123	33 786	67 200	26 881
2	Jämsä	1	2	11.07	666	6016	12 907	4663
3	Jämsänkoski	1	2	13.83	528	3818	8118	3019
4	Keuruu	1	2	12.84	760	5919	12 707	4896
5	Saarijärvi	1	3	14.62	721	4930	10 774	3730
6	Suolahti	1	5	15.12	457	3022	6159	2389
7	Äänekoski	1	3	13.17	767	5823	11 595	4264
	Rural			**12.63**	**7076**	**56 011**	**125 124**	**41 911**
8	Hankasalmi	2	5	15.07	391	2594	6080	2179
9	Joutsa	2	6	9.38	194	2069	4594	1823
10	Jyväskylän mlk.	2	7	11.82	1623	13 727	29 349	9230
11	Kannonkoski	2	4	18.64	153	821	1919	726
12	Karstula	2	4	13.53	341	2521	5594	1868
13	Kinnula	2	8	13.92	129	927	2324	675
14	Kivijärvi	2	8	15.63	128	819	1972	634
15	Konginkangas	2	3	21.04	142	675	1636	556
16	Konnevesi	2	5	12.91	201	1557	3453	1215
17	Korpilahti	2	1	11.15	239	2144	5181	1793
18	Kuhmoinen	2	2	12.91	187	1448	3357	1463
19	Kyyjärvi	2	4	11.31	94	831	1977	672
20	Laukaa	2	5	12.11	874	7218	16 042	4952
21	Leivonmäki	2	6	10.65	61	573	1370	545
22	Luhanka	2	6	10.34	54	522	1153	435
23	Multia	2	7	11.24	119	1059	2375	925
24	Muurame	2	1	9.79	296	3024	6830	1853
25	Petäjävesi	2	7	15.08	262	1737	3800	1352
26	Pihtipudas	2	8	13.02	331	2543	5654	1946
27	Pylkönmäki	2	4	17.98	98	545	1266	473
28	Sumiainen	2	3	12.80	79	617	1426	485
29	Säynätsalo	2	1	10.28	166	1615	3628	1226
30	Toivakka	2	6	11.72	127	1084	2499	834
31	Uurainen	2	7	16.47	219	1330	3004	932
32	Viitasaari	2	8	14.16	568	4011	8641	3119
	Whole province			**12.65**	**15 098**	**119 325**	**254 584**	**91 753**

Sources: Statistics Finland: Population Census 1985. Statistics Finland (1992): Statistical Yearbook of Finland, Volume 87. Ministry of Labour of Finland (1991): Employment Service Statistics, November 30, 1991.

a cluster. Thus, the total number of clusters is eight. The identification variables STR (stratum) and CLU (cluster) correspond to the urban versus rural and neighbouring municipalities, respectively.

For the following calculations, the *total number of unemployed persons* on 30 November 1991, abbreviated as UE91, is taken as the study variable. Technically, the process is as follows: using a certain sampling technique, a fixed-size sample of eight municipalities is selected. From this observed sample, a design-based estimate of a parameter of UE91 is calculated, and its efficiency studied, by means of the design-effect statistic. For model-assisted estimation and for sampling proportional to size (PPS), an auxiliary variable from a Population Census (see Table 2.1, footnote) is selected. This is the *number of households*, abbreviated as HOU85. The reason for taking HOU85 as an auxiliary variable is that it is available from the population register and is highly correlated with the study variable UE91. The frequency histogram for UE91 is displayed in Figure 2.1. Since the distribution is skewed, the mean is not the most appropriate statistic for location and the median has been chosen for further analysis.

Three different types of population parameters are considered: total T, ratio R and median M. The total of UE91 is the number of unemployed persons. The population total is given by

$$T_{ue91} = \sum_{k=1}^{32} Y_k = 15\,098.$$

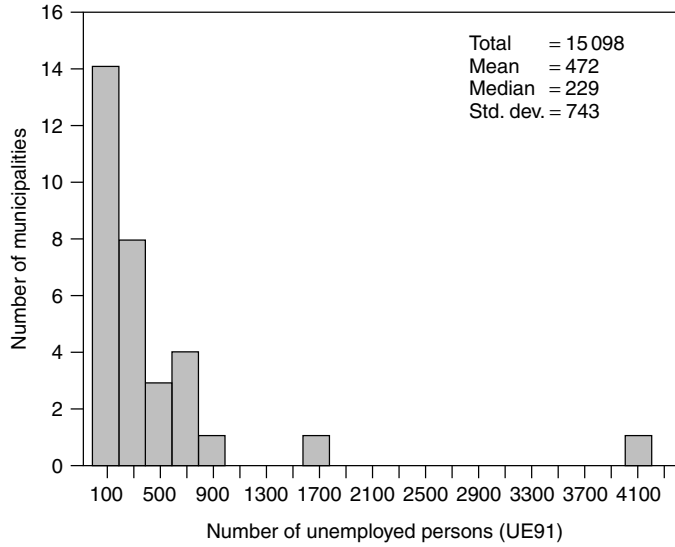


Figure 2.1 Frequency histogram for the number of unemployed persons in 1991 (the *Province'91* population; $N = 32$).

Another population total is the *size of labour force* LAB91, which can also be calculated from the figures in Table 2.1. This total is given as

$$T_{lab91} = \sum_{k=1}^{32} X_k = 119\,325.$$

Finally, *the total population size* in the *Province'91* population data is 254 584 inhabitants. The totals have long been the main parameter of interest in classical sampling theory, and official statistical agencies often produce survey estimates of population totals.

In what follows, the total T_{ue91} remains the target parameter that will be estimated under the various sampling techniques. It provides in a single figure the information on how many persons are unemployed in the province under consideration. Because an estimator $\hat{t}$ of the total is a *linear estimator* on the observations, its design variance and the corresponding variance estimator are simple and tractable.

Another interesting population parameter is the unemployment rate in this province. It can be given as the ratio of two totals

$$R = \frac{T_{ue91}}{T_{lab91}} = (15\,098/119\,325) = 0.1265.$$

A more practical expression of the ratio is to express it as an unemployment percentage given by $\%\mathrm{UE} = 100R = 100 \times 0.1265 = 12.65\%$.

Although the parameter R is simple, the design variance of an estimator $\hat{r}$ of the ratio can be complicated even if the sampling design is not complex. This is because the estimator of the ratio is of a *nonlinear* type and calls for approximations in the derivation of the design variance. In classical sampling theory, a ratio estimator refers to ratio estimation; this will be considered in Section 3.3.

The third parameter of interest is the median or 50th percentile of the distribution of municipalities according to the number of unemployed persons. It is obtained by first deriving the population cumulative distribution function (c.d.f.) given by

$$F(y) = \sum_{k=1}^{N} I(y_k \leq y)/N,$$

where $I(y_k \leq y) = 1$ if $y_k \leq y$ and zero otherwise. From the c.d.f. of UE91 the population median M is calculated as

$$M = F^{-1}(0.5) = 229.$$

Here, the median has been chosen instead of the mean since the distribution of the number of unemployed persons is very skewed; the mean is $\overline{Y} = 472$. The

median estimator $\hat{m}$ belongs to the family of *robust estimators*. These estimators are reasonably unaffected by extreme or outlying observations. However, the derivation of the design variance of the median estimator and the corresponding variance estimator can be cumbersome and requires approximations.

We have defined three population parameters: the total T, the ratio R and the median M. In the *Province'91* population these parameters have clear interpretations. The parameter T measures the total number of unemployed persons in the whole province and the parameter R, multiplied by 100, gives the province's unemployment percentage. The parameter M, the median, gives information on the location of the distribution of unemployed persons and is more appropriate than the mean because of the strongly skewed distribution of UE91.

In the following examples, we will take a sample of $n = 8$ elements from the *Province'91* population using five different sampling techniques. These are simple random sampling (SRS), systematic sampling (SYS), stratified sampling (STR), sampling proportional to size (PPS) and cluster sampling (CLU). Sampling causes sampling error, which varies according to the sampling design, but the computationally manageable size of the demonstration population will provide an opportunity to analyse the behaviour of the sampling distributions.

2.3 SIMPLE RANDOM SAMPLING AND DESIGN EFFECT

Simple random sampling can be regarded as the basic form of probability sampling applicable to situations where there is no previous information available on the population structure. This sampling technique ensures that each population element has an equal probability of selection, and thus the resulting sample constitutes a fair representation of the population.

Simple random sampling serves two functions. Firstly, it sets a baseline for comparing the relative efficiency of other sampling methods. Secondly, amongst the more advanced sampling techniques such as stratified sampling and cluster sampling, simple random sampling can be used as the final method for selecting the elementary or *primary sampling units* and for working out randomization.

Simple random sampling is seen in this section from the viewpoint that sampling a subset from a population always gives rise to sampling variation in computations. A parameter for a fixed and finite population, as for example in the *Province'91* population, the total number of unemployed in the province, is a fixed number ($T = 15\,098$), which is a constant. However, if a sample of 8 municipalities is selected out of this population of 32 municipalities, then naturally the sample estimate $\hat{t}$ of the total number of unemployed will vary among different samples depending on sample structure. This variation leads to uncertainty in statistical inference, and the way it comes into being is the reason for labelling it a *sampling error*.

However, in actual practice there is only one sample to be analysed. The random variation due to sampling needs to be kept under control in statistical inference,

and consequently one has to be familiar with the *sampling distributions* of the estimators of the unknown population parameters.

In the following, simple random sampling is introduced by looking at three sampling techniques: *Bernoulli sampling* (SRSBE), *simple random sampling with replacement* (SRSWR) and *simple random sampling without replacement* (SRSWOR). These sampling techniques have been illustrated by selecting an SRSWOR sample of eight elements from the *Province'91* population for further analysis. On this basis, sample estimates for three parameters are supplied: total T, ratio R and median M. The estimates are obtained by using survey estimation software, which produces point estimates and appropriate standard error and design-effect estimates.

Finally, the behaviour of the sampling error is examined by simulating 1000 Monte Carlo samples from the *Province'91* population and calculating the mean and variance of this sampling distribution. In the case of an unbiased estimator, the mean of the sampling distribution of the estimator should be equal to the parameter under consideration and the variance of the simulated distribution is expected to be close to the design variance of the estimator. A design variance can be calculated exactly in a fixed and known population as exemplified by the *Province'91* population. The examination of simple random sampling is concluded by presenting design-effect parameters and the corresponding estimates obtained from the actual sample.

Sample Selection

Simple random sampling can be executed by three specific selection techniques: Bernoulli sampling, simple random sampling with replacement and simple random sampling without replacement. In the first method, the sample size is not fixed in advance; in the two other methods it is fixed. Sample selection in both the Bernoulli and without-replacement types of random sampling can be conveniently carried out by a *list-sequential* procedure applied to a database. In the with-replacement type of selection, on the other hand, each separate instance of sampling has to be done by lottery or a *draw-sequential* procedure. All these techniques belong to the class of *equal-probability sampling designs* where the inclusion probabilities are $\pi_k = \pi$, i.e. a constant for all population elements.

Bernoulli sampling (SRSBE) The selection probability is set first, which in this case is the constant π with regard to all elements so that $0 < \pi < 1$. The value of the constant π is fixed so that the expected or mean sample size is $E(n_s) = N\pi$. In practice, the selection is done by appending two variables to the frame population register; let one variable be PI with the same value or a chosen π for each observation and the other variable EPSN takes a value drawn from a uniform distribution over the range (0, 1). The kth population element is included in the sample if EPSN $< \pi$. Following this procedure, all the population elements are treated sequentially. This method leads to a variation in sample size with the

expected value $E(n_s) = N\pi$ and the variance $V(n_s) = N(1-\pi)\pi$. This creates problems in variance estimation for small samples, but varying sample size is relatively unimportant in large samples. Note that Bernoulli sampling is a without-replacement-type sampling scheme.

Simple random sampling with replacement (SRSWR) Simple random sampling with replacement is based on a selection by lottery from the population by replacing the chosen element in the population after each draw. The probability of the selection of an element remains unchanged after each draw, and any two separately selected samples are independent of each other. This property also explains why this method is used as the default sampling technique in many theoretical statistical studies. Because the with-replacement assumption considerably simplifies the formulae for estimators, especially variance estimators, it is often adopted as an approximation when working with more complex sampling designs. An SRSWR design is often used also as a reference design in design-effect calculations.

Simple random sampling without replacement (SRSWOR) The most common simple random sampling method used in practice is that of simple random sampling without replacement. For simplicity, an abbreviation SRS for SRSWOR sampling is used in formulae. The probability of the selection of a single element is a constant, but this is related to how far the sampling has progressed, since the probability of selecting an element still present in the population increases with each draw. This causes difficulties in calculating the variance estimators; with-replacement sampling, dealt with earlier, is easier in this respect.

One of the possible SRSWOR samples of size 8 elements from the *Province'91* population is presented in Table 2.2. The sampling rate is $n/N = 0.25$. It is

Table 2.2 A simple random sample drawn without replacement ($n = 8$) from the *Province'91* population.

Element	Study variables	
LABEL	UE91	LAB91
Jyväskylä	4123	33786
Keuruu	760	5919
Saarijärvi	721	4930
Konginkangas	142	675
Kuhmoinen	187	1448
Pihtipudas	331	2543
Toivakka	127	1084
Uurainen	219	1330

Sampling rate $= 8/32 = 0.25$

noteworthy that this sample could have been produced by any of the three SRS methods, namely, Bernoulli, with replacement or without replacement. Even under complex designs, the assumption can be made that the actual sample would be a realization of one of these basic selection techniques. This being the case, simple random sampling without replacement can also be used as the reference in design-effect calculations when dealing with actual complex designs. The sample just drawn will now be subjected to design-based estimation.

Estimation

Statistical inference generalizes from the sample to the target population, by calculating point and interval estimates for parameters and, further, by performing tests of statistical hypotheses. For the *Province'91* population, the interest focuses on the population total T, the relative proportion $100R\%$ and the median M, with the calculations including point estimates and their standard error estimates reflecting sampling errors. In the case of simple random sampling, the design is not complex but can still be used to highlight the essential features when developing design-based estimators, design variances and the estimators for these variances.

When the corresponding estimates have been computed from the sample, the desired confidence intervals can be obtained. Moreover, a statistical test can be performed on the percentage of unemployed in the province. For example, we can test whether the percentage has remained the same since last year, i.e. $H_0: 100R\% = 100R_0\% = 9\%$.

Let us introduce the formulae for the estimators $\hat{t}$, $\hat{r}$ and $\hat{m}$ of the total T, the ratio R and the median M, and the corresponding design variance and standard error estimators under simple random sampling without replacement. For the total T, we have an estimator $\hat{t}$ given in the standard form by

$$\hat{t} = N\overline{y} = N\sum_{k=1}^{n} y_k/n \tag{2.1}$$

or the sample mean $\overline{y}$ multiplied by the population size N. The estimator can be expressed as $\hat{t} = \sum_{k=1}^{n} w_k y_k = (N/n)\sum_{k=1}^{n} y_k$, where $w_k = N/n$. The constant N/n is the *sampling weight* and is the inverse of the sampling fraction n/N. Alternatively, an estimator for the total can be written by first defining the inclusion probability of a population element. Under SRSWOR, the *inclusion probability* of a population element k is $\pi_k = n/N$ or the same constant for every population element. On the basis of the inclusion probabilities, an estimator of the total can be expressed as a more general *Horvitz–Thompson*-type estimator:

$$\hat{t}_{HT} = \sum_{k=1}^{n} w_k y_k = \sum_{k=1}^{n} \frac{1}{\pi_k} y_k = \frac{N}{n}\sum_{k=1}^{n} y_k. \tag{2.2}$$

In this case, the estimators $\hat{t}$ and $\hat{t}_{HT}$ obviously coincide, because the inclusion probabilities $\pi_k = n/N$ are equal for each k. The *Horvitz–Thompson*-type estimator is often used, for example, with probability-proportional-to-size sampling where inclusion probabilities vary. The estimator has the statistical property of unbiasedness in relation to the sampling design.

The estimator of the ratio R is the ratio of the estimators of two totals or

$$\hat{r} = \hat{t}/\hat{t}_x, \tag{2.3}$$

where $\hat{t}_x$ denotes the total of the study variable x. Although both the estimators for totals are unbiased, the estimator $\hat{r}$ of a ratio nonetheless belongs to the class of *biased estimators*. Let us consider more closely the bias of $\hat{r}$.

The bias of $\hat{r}$ is related to the linear regression existing between the two variables, y and x, which takes the form $y = A + Bx$. If the intercept is $A = 0$, then the regression line goes through the origin, which means that the ratio Y_k/X_k is constant among the elements of the population. In this instance the ratio estimator $\hat{r}$ is unbiased, whereas if $A > 0$ the bias amounts to

$$\text{BIAS}(\hat{r}) = E(\hat{r}) - R \doteq V_{srs}(\overline{y})\frac{A}{\overline{Y}^2\overline{X}}, \tag{2.4}$$

where $V_{srs}(\overline{y})$ denotes the design variance of $\overline{y}$ under the SRSWOR design and $\overline{Y}$ and $\overline{X}$ are the population means of the study variables y and x.

The formula shows that if the constant A is large, the bias is also considerable. On the other hand, with increasing sample size the variance $V_{srs}(\overline{y})$ declines, leading to a reduced bias. Therefore $\hat{r}$ is a *consistent* estimator of R and can be considered more reliable as the sample size increases (see Figure 2.3 for finite-population consistency).

An estimator of the median M can be constructed by first estimating the cumulative distribution function of the study variable at the point y. The *Horvitz–Thompson*-type estimator of the c.d.f. is given by

$$\hat{F}(y) = \sum_{k=1}^{n} w_k I(y_k \le y)/\hat{N}, \tag{2.5}$$

where w_k denotes the weight for the kth sample element and $I(y_k \le y)$ is one if $y_k \le y$ and zero otherwise. The sum of the weights is $\hat{N} = \sum_{k=1}^{n} w_k$. The estimated c.d.f. is a step function that should first be smoothed to form an estimate $\hat{m}$ of the median M. The procedure is described only briefly. The smoothed distribution function is constructed by connecting the points $\hat{F}(y)$ with straight lines and the estimated quantiles, including the median, are computed from this. The procedure provides an unbiased estimator for the median. More details are given in Francisco and Fuller (1991).

To determine confidence intervals and test statistics, the design variances, or rather the estimators of these variances, are required for the estimators $\hat{t}$, $\hat{r}$ and $\hat{m}$. They are used to estimate the sampling error brought about by the random selection of a sample from the population. Here we derive those variance estimators that are suitable for the single-sample situation. The behaviour of sampling error in more general terms is taken up separately in the context of design variances and sampling distributions of estimators.

An unbiased estimator of the design variance $V_{srs}(\hat{t})$ (see equation (2.8)) of the estimator $\hat{t}$ of the total is given by

$$\hat{v}_{srs}(\hat{t}) = N^2\left(1-\frac{n}{N}\right)\sum_{k=1}^{n}(y_k-\overline{y})^2/n(n-1) = N^2\left(1-\frac{n}{N}\right)\hat{s}^2/n, \qquad (2.6)$$

where $\overline{y}=\sum_{k=1}^{n} y_k/n$ is the sample mean and $\hat{s}^2=\sum_{k=1}^{n}(y_k-\overline{y})^2/(n-1)$ is an estimator of the element variance S^2. The square root of the variance estimator is the *standard error* of the estimator $\hat{t}$ and is denoted by s.e $(\hat{t})$.

Variance estimators for the ratio $\hat{r}$ and the median $\hat{m}$ are considerably more complicated since both must be regarded as nonlinear estimators. The approximate variance estimator for the estimator $\hat{r}$ of the ratio is

$$\hat{v}_{srs}(\hat{r}) = \left(1-\frac{n}{N}\right)\left(\frac{1}{\overline{x}^2}\right)\sum_{k=1}^{n}(y_k-\hat{r}x_k)^2/n(n-1). \qquad (2.7)$$

In developing this variance estimator, the ratio estimator has been *linearized* with the *Taylor series expansion*, and therefore the above equation gives an approximate estimator of the design variance. This technique will be considered in more detail in Chapter 5. The variance estimator of $\hat{m}$ also requires use of the linearization method. This implies that the variance estimator of the median cannot be expected to be very stable, especially for small samples. The standard error for a median is determined as follows. A lower 0.975-level and an upper 0.025-level bounds for the smoothed cumulative distribution function are created. The standard error for the pth quantile is a quarter of the horizontal distance at level p between the upper and lower bounds of the smoothed distribution function.

Computation of Design-based Estimates

The computation of design-based estimates and their standard errors has been performed here and elsewhere in this book using the appropriate software, which accounts for the design complexities. The statistical analysis follows the steps presented in the flow chart in Figure 1.1. We assume that the data are cleaned such that the necessary data-processing operations have been completed successfully. This includes data entry, coding, editing and imputation and the derivation of the sampling weight.

For design-based estimation, the following *sampling-design identifiers* must be included in the data set to be analysed: *stratum identification variable, cluster identification variable* and *sampling weight variable*. It should be noted that in addition to complex designs, these identifiers can also be assigned for simple designs, for example, for a design involving only one stratum or a design without clustering (each single unit constitutes a cluster of its own). In addition to these variables, sampling rates must be supplied under without-replacement sampling. *User-specific computer programs* are often used to prepare the cleaned data set for analysis purposes.

In the analysis phase, the sampling identifiers are then supplied to the chosen *survey analysis software*. Of course, the use of the design information requires full awareness of the complexities of the actual sampling design. Use of the design information in estimation is illustrated under all the sampling techniques to be considered in this book. The output of a standard survey estimation software includes the point estimates and their estimated standard errors, coefficients of variation and design effects. These statistics are calculated by taking the sampling design into account. In addition, some useful sampling design information is usually included. Our first example is of design-based estimation under simple random sampling without replacement.

Example 2.1

Analysing an SRSWOR sample from the *Province'91* population. We produce the estimates of the total, the ratio and the median, and their standard error estimates, from the sample selected earlier under simple random sampling without replacement. First, the design identifiers are appended to the sampled data set. These include the stratum identifier STR, which in the case of a simple random sample is a constant for all sample elements, i.e. STR = 1. Next, we need to know whether an element belongs to a group of elements or a cluster. In element sampling, each element is a cluster of its own; therefore CLU equals the ID number of the observation. Finally, we enter the weight variable, which under the SRSWOR design is the inverse of the inclusion probability or $w_k = \pi_k^{-1} = (n/N)^{-1} = N/n$. It is used to weight the sample observations in the estimation of the total so that the weights sum to N. In general for the estimation of a total, the weight variable should be scaled such that the sum of the weights equals the population size. In this example, the population size is 32 municipalities ($N = 32$) and the selected sample includes eight municipalities ($n = 8$); therefore, the weight variable is given the value WGHT $= 32/8 = 4$.

As soon as these preliminary steps are completed, the data set should resemble Table 2.3. To make the table more readable, an alphanumeric variable LABEL has been included and the rest of the variables have been divided into two headlines: 'Sample design identifiers' and 'Study variables'.

It is important under without-replacement-type sampling to provide the sampling rate to account for the *finite-population correction* (f.p.c.) in the variance

Table 2.3 A simple random sample drawn without replacement from the *Province'91* population ($n = 8$) provided with the sample design identifiers.

Sample design identifiers			Element	Study variables	
STR	CLU	WGHT	LABEL	UE91	LAB91
1	1	4	Jyväskylä	4123	33786
1	4	4	Keuruu	760	5919
1	5	4	Saarijärvi	721	4930
1	15	4	Konginkangas	142	675
1	18	4	Kuhmoinen	187	1448
1	26	4	Pihtipudas	331	2543
1	30	4	Toivakka	127	1084
1	31	4	Uurainen	219	1330

Sampling rate $= n/N = 8/32 = 0.25$

estimators when dealing with small populations. In this example, the sampling rate is $8/32 = 0.25$, and thus the f.p.c. equals $(1 - n/N) = 0.75$.

Estimation results are displayed in Table 2.4. It includes the point estimates for $\hat{t}$, $\hat{r}$ and $\hat{m}$, and their estimated standard errors, coefficients of variation and design effects. The *coefficient of variation* is, for example, for the total c.v $(\hat{t}) =$ s.e $(\hat{t})/\hat{t}$. In this case, the deff estimates are equal to unity, since SRSWOR design is also the reference scheme. In addition to the estimates, the values of the corresponding population parameters T, R and M are supplied. For further details, the reader is advised to consult the web extension of the book.

The results of the estimation are interpreted as follows. The point estimate of the total number T of unemployed persons UE91 for the whole province is $\hat{t} = 26\,440$ and the corresponding standard error estimate is s.e $(\hat{t}) = 13\,282$. On the basis of these two estimates, and by using the standard normal distribution $N(0,1)$ as an approximate distribution for the estimated total, the following 95% confidence interval is obtained for the total number of unemployed persons in the province:

$$\hat{t} - 1.96 \times \text{s.e}(\hat{t}) < T < \hat{t} + 1.96 \times \text{s.e}(\hat{t})$$

i.e. $407 < T < 52\,472$, which is so wide as to lack any significance for administrative purposes. We shall see later how this confidence interval is affected by

Table 2.4 Estimates from a simple random sample drawn without replacement ($n = 8$); the *Province'91* population.

Statistic	Variables	Parameter	Estimate	s.e	c.v	deff
Total	UE91	15098	26440	13282	0.50	1.00
Ratio (%)	UE91, LAB91	12.65%	12.78%	0.41%	0.03	1.00
Median	UE91	229	226	150	0.66	1.00

selecting a more effective sampling scheme in such a way as to produce a smaller sampling error.

The estimate $\hat{r}$ for percentage unemployment in the province is 12.78%. Since the standard error estimate (s.e) of $\hat{r}$ is available, we can test statistically whether the current unemployment rate R is different from that estimated a year ago: it was then 9%, thus $H_0: R = R_0 = 0.09$. Using again the normal approximation we have

$$Z = \frac{\hat{r} - R_0}{\text{s.e}(\hat{r})} = \frac{0.1278 - 0.09}{0.0041} = 9.22^{***},$$

and we reject the H_0 hypothesis and conclude that the unemployment percentage of the province has changed significantly during the past year. The significance level is denoted as *** referring to the rejection probability, i.e. the p-value of the test which in this case is less than 0.001.

On the other hand, the point estimates for the ratio and the median are close to the corresponding parameters.

Next, we study the design variances and sampling distributions of the estimators $\hat{t}$, $\hat{r}$ and $\hat{m}$ in greater detail.

Design Variance and Sampling Distribution

Simple random sampling is convenient for demonstrating how different estimators and their variances behave under a certain sampling design and how the sampling error is influenced by the randomization. We examine this behaviour by first calculating the design variances of $\hat{t}$, $\hat{r}$ and $\hat{m}$, denoted by V_{srs}, under the SRSWOR design. These variances can be calculated for the small fixed population under consideration. However, the *design variance* does not contain all the information on the sampling error; derivation of the sampling distributions of the estimators allows closer examination of the behaviour of the estimators.

Sampling distributions of estimators are often derived by simulating a large number of samples from the population using the given sampling scheme. We have simulated by the *Monte Carlo* method a total of 1000 samples of size eight ($n = 8$) elements from the *Province'91* population under SRSWOR. From each of these samples, the estimates $\hat{t}$, $\hat{r}$ and $\hat{m}$ are calculated. The distribution of each estimator constitutes an experimental sampling distribution for that estimator, i.e. the total, the ratio and the median. These distributions provide information about the location and shape of the sampling distribution.

Design variance formulae and the corresponding observed values for the total, ratio and median estimators under SRSWOR using the *Province'91* population are:

Total T: A design variance for $\hat{t}$ is

$$V_{srs}(\hat{t}) = \frac{N^2}{n}\left(1 - \frac{n}{N}\right)\sum_{k=1}^{N}(Y_k - \overline{Y})^2/(N-1) = N^2\left(1 - \frac{n}{N}\right)S^2/n, \qquad (2.8)$$

where $\overline{Y} = \sum_{k=1}^{N} Y_k/N$ is the population mean and $S^2 = \sum_{k=1}^{N}(Y_k - \overline{Y})^2/(N-1)$ is the population variance. The observed design variance is

$$V_{srs}(\hat{t}) = \frac{32^2}{8}\left(1 - \frac{8}{32}\right)743.36^2 = 7283^2.$$

Ratio R: An approximate design variance for $\hat{r}$ is

$$V_{srs}(\hat{r}) \doteq \frac{1}{\overline{X}^2}\frac{1}{n}\left(1 - \frac{n}{N}\right)\sum_{k=1}^{N}(Y_k - R \times X_k)^2/(N-1), \tag{2.9}$$

which gives the observed value

$$V_{srs}(\hat{r}) = \frac{1}{3729^2}\frac{1}{8}\left(1 - \frac{8}{32}\right)315.91^2/(32-1) = 0.005^2.$$

Median M: There are several approximative variances available for the design variance of the median $\hat{m}$. One possibility is to approximate the variance from the cumulative distribution function as follows:

$$V_{srs}[\hat{F}(\hat{m})] = \frac{N-n}{N-1}\frac{1}{n}F(M)(1 - F(M)) \doteq \frac{1 - n/N}{n}0.25, \tag{2.10}$$

which is very simple because no unknowns are included. It gives

$$V_{srs}[\hat{F}(\hat{m})] \doteq \frac{1 - 0.25}{8}0.25 = 0.02,$$

which should be rescaled to obtain the design variance of $\hat{m}$ on the ordinary study variable scale. In the *Province'91* population, however, we use the approximate design variance from the Monte Carlo simulations (see Figure 2.2); hence we obtain

$$V_{srs}(\hat{m}) \doteq \hat{v}(\hat{m}_{mc}) = 107^2.$$

Note that the design variances are displayed in terms of squared standard errors to facilitate comparison with the standard error estimates (s.e) exhibited in Table 2.4. When comparing the design variance, or standard error of an estimator to the corresponding estimate from the actual sample, it can be seen that they differ owing to sample-to-sample variation. For example, the variance estimate for the total was $\hat{v}_{srs}(\hat{t}) = 13\,282^2$, and the corresponding design variance was calculated as $V_{srs}(\hat{t}) = 7283^2$. The sample estimate considerably overestimates the design variance in this case. For the ratio estimator these figures are $\hat{v}_{srs}(\hat{r}) = 0.004^2$ and $V_{srs}(\hat{r}) = 0.005^2$, which are quite close. Finally, for the median we have

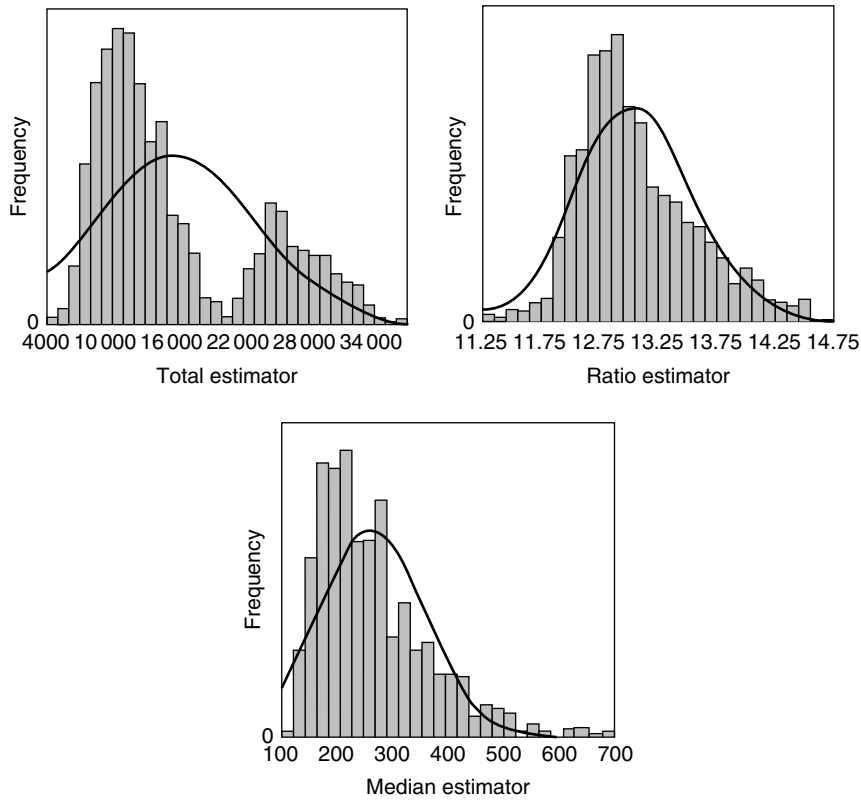


Figure 2.2 Sampling distributions of the estimators $\hat{t}$, $\hat{r}$ and $\hat{m}$ from 1000 Monte Carlo samples taken from the *Province'91* population under an SRSWOR design ($N = 32$, $n = 8$).

$\hat{v}_{srs}(\hat{m}) = 150^2$ and $V_{srs}(\hat{m}) = 107^2$; the sample estimate is again noticeably larger than the corresponding design variance.

For a closer examination of the behaviour of the estimators under simple random sampling without replacement, estimates for the total, ratio and median from Monte Carlo simulations are displayed as histograms in Figure 2.2. The mean of the distribution of a Monte Carlo estimator is expected to coincide with the corresponding population parameter, and the variance should approximate the design variance of the estimator.

The mean of the total estimates is $\hat{t}_{mc} = 15\,049$, which fits well with the corresponding parameter $T = 15\,098$. The variance of the total estimates is 7278^2, which is close to the design variance $V_{srs}(\hat{t}) = 7283^2$. In this respect the estimator $\hat{t}$ works well.

On closer examination, two peaks are noted in the histogram. The distribution does not seem bell-shaped when referred to the normal distribution, which can be

used as the reference (the values from the corresponding normal distribution are displayed as a solid curve in the figure). Great discrepancies are noted between the observed and theoretical distributions. This cautions us against basing our inferences on an assumption of a normal distribution. The causes are obvious. The sampling distribution of $\hat{t}$ strongly depends on the distribution of UE91 in the *Province'91* population, which is highly skewed in favour of one municipality (provincial capital), where one-third of the total population of the province lives (see Figure 2.1). The population and sample sizes are not large enough to meet the requirements of a normal approximation. Consequently, simple random sampling might not be an appropriate technique for the estimation of the total in this population.

The simulated distribution indicates that the estimator $\hat{r}$ for the ratio UE91/LAB91 works well. The mean of the ratio estimates is $\hat{r}_{mc} = 0.128$, which is almost equal to the population parameter $R = 0.1265$. The variance of the ratio estimates is 0.006^2 and coincides with the design variance, $V_{srs}(\hat{r}) = 0.005^2$. Moreover, the distribution is reasonably bell-shaped, indicating that the normal approximation is better motivated than that for the total estimator.

The median M was defined as the 50th percentile of the cumulative distribution function (c.d.f.) of the study variable y. Usually, the c.d.f. is unknown and the median should be approximated. The generally used procedure for a median estimate is to arrange the sample values in ascending order $y_{(1)} < \cdots < y_{(k)} < \cdots < y_{(n)}$ and to take the middle value as the median if the sample size is odd, otherwise the median is taken as the mean of the two middle values or $\hat{m} = \frac{1}{2}[y_{(n/2)} + y_{(n/2+1)}]$. This kind of an estimator of a median is often called *50% trimmed mean*.

For a symmetric population, the mean and median coincide. The *Province'91* population is heavily skewed, as can be seen in Figure 2.1, and therefore the difference between the population mean and median is as great as $\overline{Y} - M = 472 - 229 = 223$. We next investigate the effect of sample size on the behaviour of the estimators for a total and a ratio.

Finite Population Consistency and Sample Size

Statistical properties of two basic estimators, $\hat{t}$ (for a total) and $\hat{r}$ (for a ratio) are now examined in more detail by using simulation methods.

A method of estimation is called *unbiased* if the average value of the estimate, taken over all possible samples of given size n, is exactly equal to the true population value. Further, a method of estimation is called *consistent* if the estimate becomes exactly equal to the population value when $n = N$, that is, when the sample consists of the whole population (Cochran 1977, pp. 21–22). In Särndal *et al.* (1992, p. 168), this type of consistency is defined as *finite population consistency*. We examine the behaviour of total and ratio estimators by Monte Carlo methods by simulating 1000 samples with SRSWOR from the *Province'91* population.

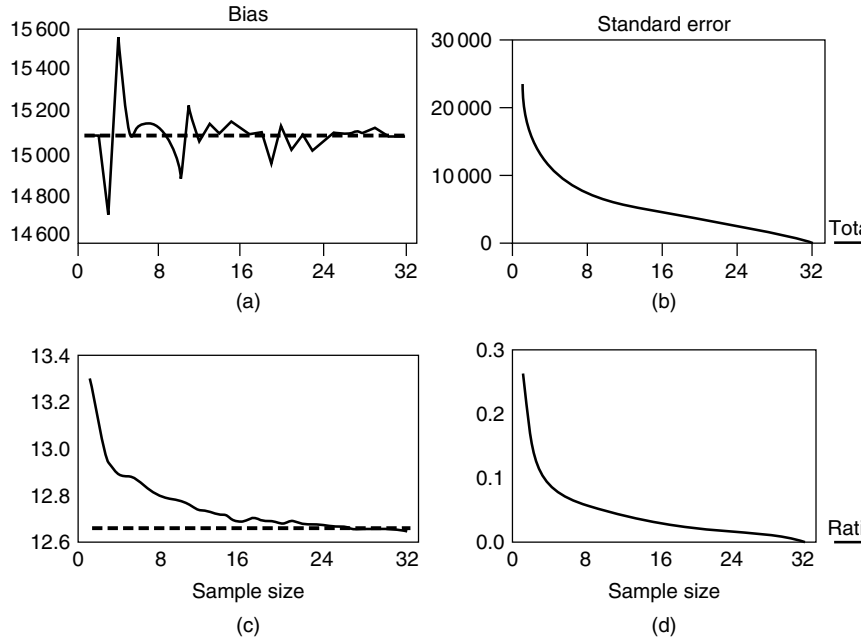


Figure 2.3 Bias, consistency and precision of the estimator $\hat{t}$ of a total and $\hat{r}$ of a ratio. Monte Carlo means $\hat{t}_{mc}$ and $\hat{r}_{mc}$ and the corresponding standard errors of simulated 1000 SRSWOR samples with different sample sizes drawn from the *Province'91* population.

Varying-size samples are selected; sample sizes vary from $n = 1$ to the population size $N = 32$. Results are presented in Figure 2.3.

The estimator $\hat{t} = N \times \sum_{k=1}^{n} y_k/n$ of the total T (= 15 098) of the study variable UE91 (number of unemployed) is unbiased, as Figure 2.3(a) indicates. As expected, the standard error s.e($\hat{t}$) decreases when the sample size increases, as can be seen from Figure 2.3(b). On the other hand, the estimator $\hat{r} = \sum_{k=1}^{n} y_k / \sum_{k=1}^{n} x_k$ of the ratio, where x refers to the study variable LAB91 (size of labour force), is somewhat biased for the population ratio R (= 0.1265), but is consistent (Figure 2.3(c)). Consistency is verified by a vanishing bias with increasing sample size. Also for the estimator of the ratio, the standard error estimate declines when the sample size increases (Figure 2.3(d)). We conclude that both estimators are consistent and, moreover, the estimator for the total is unbiased.

DEFF and Efficiency of Sampling Design

The design effect was previously defined as the ratio of two design variances where the numerator is the design variance of an estimator under the actual sampling

design and the denominator is the design variance of a simple random sample of the same number of elements. This definition was originally given by Kish (1965, p. 258) in which simple random sampling without replacement was taken as the reference. More formally, let the design variance of an estimator, e.g. for the total estimator $\hat{t}$, be $V_{p(s)}(\hat{t})$ under the actual design. The DEFF parameter is obtained as

$$\mathrm{DEFF}(\hat{t}) = \frac{V_{p(s)}(\hat{t})}{V_{srs}(\hat{t})}. \tag{2.11}$$

In the design effect (2.11), it is assumed that the estimator $\hat{t}$ applies to both the actual and reference designs. For more complex actual designs, the DEFF was, in Section 2.1, given also by a more general formula that allows a design-based estimator, denoted by $\hat{t}^*$, which differs from the SRSWOR counterpart $\hat{t}$. Moreover, in the Kish definition, SRSWOR acts as the reference. In practice, this definition is often interpreted more loosely. The reason for this is that simple random sampling either with or without replacement tends to lead to the same results if the target population is large and the sampling fraction n/N is small. This is generally the case with large-scale survey sampling. Variance estimators under SRSWR are algebraically simpler than those under SRSWOR, so SRSWR is in this respect more convenient as the reference design. This is also emphasized in software applications for survey analysis.

Obviously, if the actual sampling design is SRSWOR then DEFF $= 1$. And for simple random sampling with replacement (SRSWR), whose design variance for a total estimator $\hat{t}$ is $V_{srswr}(\hat{t}) = N^2(1 - 1/N)S^2/n$, the DEFF reduces to

$$\mathrm{DEFF}(\hat{t}) = \frac{V_{srswr}(\hat{t})}{V_{srs}(\hat{t})} = \frac{N^2\left(1 - \frac{1}{N}\right)\frac{S^2}{n}}{N^2\left(1 - \frac{n}{N}\right)\frac{S^2}{n}} = \frac{N-1}{N-n}.$$

This DEFF is always greater than one if $n \geq 2$, which implies that the SRSWR design is less efficient than the SRSWOR design. Thus, DEFF for SRSWR depends only on the population size N and sample size n. If the population is very large and the sampling rate n/N is negligible, then DEFF is close to one.

In practice, the design variance $V_{p(s)}$ and the corresponding SRSWOR (or SRSWR) reference variance of an estimator are estimated from the selected sample. Thus, the DEFF must be estimated from the sampled data and for obtaining an estimate deff, the estimates of the variances are used. In the next example, we calculate DEFF and deff figures for data from the *Province'91* population.

Example 2.2

A sample of size $n = 8$ is selected from the *Province'91* population ($N = 32$) by SRSWOR. This sample is now assumed to be a realization of SRSBE (Bernoulli

sampling) and SRSWR (simple random sampling with replacement). To compare the efficiencies of these sampling designs, we calculate DEFF parameters for the estimator $\hat{t}$ of the total number of unemployed UE91. From the population, we know that the standard deviation $S = 743$ and the mean $\overline{Y} = 472$. Thus,

$$\begin{aligned}
&\mathrm{DEFF}_{srs}(\hat{t}) = 1 \text{ (by definition)},\\
&\mathrm{DEFF}_{srswr}(\hat{t}) = \frac{N-1}{N-n} = \frac{32-1}{32-8} = 1.29 \text{ and}\\
&\mathrm{DEFF}_{srsbe}(\hat{t}) = 1 - \frac{1}{N} + \frac{\overline{Y}^2}{S^2} = 1 - \frac{1}{32} + \frac{472^2}{743^2} = 1.37.
\end{aligned}$$

The DEFF parameters show that both SRSWR and SRSBE are less efficient than the reference SRSWOR design. For SRSBE, increased variance is partly due to the random sample size.

We calculate the deff estimates from the selected sample presented in Table 2.3. The estimate for the population standard deviation is $\hat{s} = 1355.615$ and for the population mean $\overline{y} = 826.25$. Interpreting this sample as a realization of simple random sampling with replacement or Bernoulli sampling, the deff estimates are:

$$\begin{aligned}
&\mathrm{deff}_{srs}(\hat{t}) = 1 \text{ (by definition)},\\
&\mathrm{deff}_{srswr}(\hat{t}) = \frac{N-1}{N-n} = \frac{32-1}{32-8} = 1.29 \text{ and}\\
&\mathrm{deff}_{srsbe}(\hat{t}) = 1 - \frac{1}{N} + \frac{\overline{y}^2}{\hat{s}^2} = 1 - \frac{1}{32} + \frac{826.25^2}{1355.62^2} = 1.34
\end{aligned}$$

Of course, the deff estimate for SRSWR is the same as the parameter DEFF. Even for SRSBE sampling, the deff estimate is almost the same as the corresponding DEFF parameter.

Design variances and variance estimators of the total, ratio and median were considered under simple random sampling without replacement. For the linear estimator $\hat{t}$ of the total, an analytical design variance was derived, yielding a basically equal formula for the corresponding variance estimator. For the ratio $\hat{r}$ as a nonlinear estimator, an approximative design variance was derived by the linearization method; the variance estimator also mirrored the design variance. And for the design variance of the robust estimator, the median $\hat{m}$, alternative approximative estimators are available whose suitability, however, varies at least for small samples.

Summary

Simple random sampling was introduced in order to promote familiarity with the most important concepts of estimation under a specific sample-selection scheme.

The key statistical concepts appeared at three levels. At the first level are the unknown population parameters of the study variable, such as the total T, the ratio R and the median M, which are to be estimated from a selected sample. At the second level are the estimators of the population parameters, and the design variances of these estimators, including the design parameters and other characteristics of the sampling distribution of an estimator. The randomization produced by the sampling involves variation in the observed values of the estimators calculated from repeated samples from the population. The design variance is intended to capture this variation, which is also reflected in the sampling distribution of an estimator. It appeared that it is beneficial to be aware of the properties of the sampling distribution as a basis for appropriate point and interval estimation and for hypothesis testing. The efficiency of a sampling design is reflected in the design effect DEFF of an estimator.

In practice, only the sample actually drawn is available for the estimation. Thus, at the third level are the sample estimates of the population parameters, and the estimators of the design variances for obtaining standard error estimates and the corresponding confidence intervals. An important figure is the deff estimate calculated from the sample by using the estimated design variance and the respective variance estimate from the assumed simple random sample.

Covering all three levels, the properties of the estimators of the total, ratio and median were studied in detail for a simple random sample drawn without replacement from the *Province'91* population. The estimator $\hat{t}$ was for the total number T of unemployed persons UE91 in the province, the ratio estimator $\hat{r}$ was for the unemployment rate R in the province, and the median estimator $\hat{m}$ was for the average number M of unemployed persons per municipality. These estimators cover three important families of estimators, namely linear, nonlinear and robust estimators. In this case, all the DEFF figures and deff estimates were ones because SRSWOR was also the reference in the design-effect calculations. Under other sampling schemes, we will see in later chapters how efficiency varies according to both the estimator and the sampling design, and in many cases deff estimates differing from unity will be obtained.

Finally, note that SRS cannot be taken solely as a simple device for the demonstration of sampling error and other key concepts when discussing the basics of survey sampling, nor as the reference in efficiency comparisons. Simple random sampling can also be included as an inherent part of sampling designs in complex sample surveys; thus it is of practical value as well.

2.4 SYSTEMATIC SAMPLING AND INTRA-CLASS CORRELATION

Systematic sampling is one of the most frequently used sample selection techniques. A list of population elements or a computerized register serves as the selection frame from which every qth element can be systematically selected. For example,

many population registers are alphabetically ordered by family name. The first member is selected at random among the first q elements. The rest of the sample is selected by taking every qth element thereafter down to the end of the list. We have devoted a great deal of space to discussing estimation in a systematic sample, since it presents a good example of the complexities encountered when estimating under a design that involves a certain design parameter in the design variance of an estimator. Here the design parameter is the intra-class correlation coefficient ρ_{int}. A further complexity arises in the estimation of the design variance; as there is no known analytical variance estimator even for such a simple estimator as the total, we shall derive several approximate variance estimators. In choosing between them, further information on the structure of the target population would be helpful.

Systematic sampling may in some cases be more effective than simple random sampling. This will occur, for example, if there is a certain relationship between the ordering of the frame population and the values of the study variable. The most common cases are those where the population is already stratified or a trend exists that follows the population ordering, or there is a periodic trend; all these situations can also be reached by appropriate sorting procedures. Periodicity may be harmful in some cases, especially if harmonic variation coincides with the sampling interval. Good *a priori* knowledge of the structure of the population is thus beneficial to gaining efficient estimation.

Sample Selection

Let us suppose that a systematic sample of size n elements is desired from a fixed population of N elements. There are several ways of selecting the sample. The most common is to draw a single sample of size n with a *sampling interval* of $q = N/n$. Alternatively, two, or more generally m, replicated systematic samples can be taken, each of size n/m elements, the length of the sampling interval being $m \times q$. This method is suitable if variance estimation is to be carried out using so-called replication techniques.

Let us consider systematic sampling with *one random start*. The first task is to number the elements of the frame population consecutively by $1, 2, \ldots, q, q+1, \ldots, N-1, N$, where $q = N/n$ refers to the sampling interval. If q is not an integer, all sampling intervals can be defined as of equal length except one. The selection proceeds as follows. Select a random integer with an equal probability of $1/q$ between 1 and q. Let it be q_0. The sample will be composed of elements numbered $q_0, q_0 + q, q_0 + 2q, \ldots, q_0 + (n-1)q$, so that one member from each sampling interval is included.

Another selection with one random start can be executed by taking a random integer from the interval $[1, N]$. Let it be Q_0. Starting from Q_0, the selection proceeds forward and backward with steps of the length of the sampling interval q. The composition of the systematic sample will be $\ldots, Q_0 - 2q, Q_0 - q, Q_0, Q_0 +$

$q, Q_0 + 2q, \ldots$. Alternatively, a systematic sample can be drawn by treating the observations in the frame as a closed loop. Beginning from the random start Q_0 the selection proceeds successively by drawing elements $Q_0 + q$, $Q_0 + 2q, \ldots$, till the end of the frame, and then the selection continues from the beginning of the frame. The loop will be closed when n elements have been drawn. These random start methods lead to the selection of a systematic sample size of n elements, and the methods are equivalent with respect to the estimation.

In replicated systematic sampling, *multiple random starts* are used. The intended sample size n is first allocated to the m subsamples so that the sampling interval for each subsample of equal size n/q is $m \times q$. For every subsample, an integer for random start is chosen without replacement from the first sampling interval, and the selection is performed according to the first of the methods introduced above. This procedure gives a set of equal-sized replicate systematic samples comprising n distinct elements in the combined sample.

In systematic sampling, the number of different samples is quite small. If the sampling interval is $q = N/n$, there will be q separate systematic samples in total. Thus, the selection probability for a sample s is $p(s) = 1/q$. When one element from each sampling interval is included, the inclusion probability for the kth population element is $\pi_k = 1/q = n/N$, which is the same as the selection probability. The inclusion probability is also equal to that under simple random sampling without replacement. So systematic sampling is also an equal-selection-probability design of without-replacement type.

Estimation

The ease of selection of a systematic sample does not continue into the estimation phase. Point estimates for total T, ratio R and median M are still easily calculated using the corresponding estimators from simple random sampling. But it is not possible to estimate the design variance analytically from the selected sample; approximations have to be used for this purpose. This is the consequence of only one population member being drawn from each sampling interval. Thus, no information is available in the sample on the variation within a sampling interval required to analytically estimate the variance. The problem can be illustrated in the estimation of total T using the estimator

$$\hat{t} = N \sum_{k=1}^{n} y_k / n, \tag{2.12}$$

which is the same as equation (2.1) for SRSWOR. Under systematic sampling, the design variance of $\hat{t}$ is given by

$$V_{sys}(\hat{t}) = N^2 \sum_{j=1}^{q} (\overline{Y}_j - \overline{Y})^2 / q, \tag{2.13}$$

where $\overline{Y}_j$ is the mean of the jth systematic sample and $\overline{Y}$ is the population mean. The variation depends on the extent to which the q sample-specific means $\overline{Y}_j$ vary around the overall mean $\overline{Y}$. If each sample closely mirrors the composition of the population, the design variance would be small and thus the estimation of the total would be efficient. But if the sample-specific means vary, a large design variance would be obtained. The situation can be illustrated by a decomposition of the total variation between and within the systematic samples. This will be discussed further under intra-class correlation.

In practice, only one systematic sample is selected and the design variance is approximated by using one of the alternative, but more or less biased, variance estimators $\hat{v}_{sys}(\hat{t})$. The choice of the approximate variance estimator should be based either on auxiliary information available in the frame population or the use of certain methodological solutions such as sample reuse or selection of replicated systematic samples. Five approximative variance estimators are introduced in equations (2.14) to (2.18).

1. Randomly ordered population It is often natural to assume that the values of the study variable are in random order in the frame population. If this model is correct, the variance estimator of simple random sampling without replacement, given by

$$\hat{v}_{1.sys}(\hat{t}) \doteq \hat{v}_{srs}(\hat{t}) = N^2 \left(1 - \frac{n}{N}\right) \hat{s}^2/n, \tag{2.14}$$

is unbiased under the actual systematic sample. Although seldom exactly correct, this model seems to be realistic, for example, for population registers if the persons appear alphabetically by name within it.

2. Implicitly stratified population The population elements are sorted according to the values of a variable. For example, in a population register, persons can be listed according to sex so that females occur first followed by males. This kind of stratification is called *implicit stratification*. The corresponding approximate variance estimator is based on successive differences $a_i = y_i - y_{i-1}$ and is given by

$$\hat{v}_{2.sys}(\hat{t}) \doteq N^2 \left(1 - \frac{n}{N}\right) (1/n) \sum_{i=2}^{n} a_i^2/2(n-1). \tag{2.15}$$

Alternatively, it is possible to make direct use of the variance estimator of stratified random sampling with proportional allocation by using equation (2.6) from SRSWOR in each implicit stratum; hence we get an estimator denoted by $\hat{v}_{2.str}(\hat{t})$ to be introduced in Section 3.1.

3. Autocorrelated population This possibility arises under the superpopulation mechanism, which is assumed to generate a correlation ρ_q between each pair of elements of the population that are q units apart. This correlation is similar to the

autocorrelation familiar from the analysis of time-series. It is expected that this correlation is positive; if not, some of the other approximations should be used. The autocorrelation coefficient can be estimated from the selected sample and used as a correction factor for the variance estimator $\hat{v}_{srs}$ as follows:

$$\hat{v}_{3.sys}(\hat{t}) \doteq N^2 \left(1 - \frac{n}{N}\right) (\hat{s}^2/n)[1 + 2/\log(\hat{\rho}_q) + 2/(\hat{\rho}_q^{-1} - 1)], \tag{2.16}$$

where $0 < \hat{\rho}_q < 1$ is the estimated value of the autocorrelation. When the autocorrelation is greater than zero, the term in brackets is less than one and decreases towards zero with increasing $\hat{\rho}_q$. Thus, strong autocorrelation increases the efficiency.

4. Sample reuse The parent sample is split into two or more equally sized distinct systematic subsamples. The design variance is estimated from the observed variation between the m subsamples as follows:

$$\hat{v}_{4.sys}(\hat{t}) \doteq N^2 \left(1 - \frac{n}{N}\right) \sum_{l=1}^{m} (\overline{y}_l - \overline{\overline{y}})^2/m(m-1), \tag{2.17}$$

where $\overline{\overline{y}} = \sum_{l=1}^{m} \overline{y}_l/m$ is the mean of the m subsample means. In place of $\overline{\overline{y}}$, the estimate $\overline{y}$ can be used in (2.17). Other sample reuse methods such as bootstrap, jackknife and balanced half-samples are other possible candidates for variance estimation. Sample reuse methods will be discussed in more detail in Chapter 5.

5. Replicated systematic sample This method resembles the one above where the parent sample is split into two or more subsamples, but here this is done before the sample selection. Selection is performed by drawing without replacement two or more replicated systematic subsamples. The variation between the m subsamples gives an opportunity to estimate the design variance. The formula for the approximate variance is the same as that for the previous method, i.e.

$$\hat{v}_{5.sys}(\hat{t}) = \hat{v}_{4.sys}(\hat{t}). \tag{2.18}$$

All the five variance estimators are approximate and thus their statistical properties depend on the validity of the respective model assumption or on the success of the splitting of parent samples. In the real world there is, of course, no assurance of this. We can, however, evaluate the validity of these variance estimators for the *Province'91* population, since it is possible to calculate the value of the design variance V_{sys} and, therefore, also the intra-class correlation ρ_{int} as the design parameter.

Example 2.3

Variance approximations under systematic sampling from the *Province'91* population. A systematic sample of 8 municipalities ($n = 8$) from the total of 32

municipalities in the *Province'91* population can be selected in two alternative ways:

1. The province is divided into eight sampling intervals, each containing four municipalities. A single sample is selected, including, for example, the first municipality from each sampling interval. Thus, the sample size will be eight elements.
2. The province is divided into four sampling intervals, each containing eight municipalities. Two parallel systematic samples are selected without replacement, one of which includes, for example, the first municipality of each sampling interval and the other, the fifth. The sample is thus composed of two distinct replicated systematic samples of four municipalities, and the total sample size is again eight municipalities.

Both the methods are assumed to produce in this case, the same actual sample. The sampled data is displayed in Table 2.5. Recall from Table 2.1 that the implicit stratification is based on the ordering of the municipalities in the municipality register: densely populated towns are given first, followed by rural municipalities. Systematic sampling through such a frame register selects municipalities from each stratum in the same proportion that they are found in the stratum. The result of this sampling is the same as stratified sampling using proportional allocation. Stratified sampling will be discussed in more detail in Section 3.1.

All the five approximate variance estimators have been calculated on the basis of the sampled data set. To compute the variance estimate under the stratification assumption, the stratum identifiers receive the value STR = 1 if the municipality is a town, or STR = 2 for a rural municipality. Similarly, as under simple random sampling, the cluster identifier (CLU) receives the corresponding element-identification value. In proportionally stratified sampling the element

Table 2.5 A systematic sample from the *Province'91* population (sample design identifiers are given for implicit stratification).

Sample design identifiers			Element	Study variables	
STR	CLU	WGHT	LABEL	UE91	LAB91
1	1	4	Jyväskylä	4123	33786
1	5	4	Saarijärvi	721	4930
2	9	4	Joutsa	194	2069
2	13	4	Kinnula	129	927
2	17	4	Korpilahti	239	2144
2	21	4	Leivonmäki	61	573
2	25	4	Petäjävesi	262	1737
2	29	4	Säynätsalo	166	1615

Sampling rates: Stratum 1 = 0.25. Stratum 2 = 0.25

weights are constants or, as here, the weight equals WGHT = 4 as under simple random sampling. The sampling rate is given for each stratum separately, but even then it is the same figure, 0.25.

The estimation results under implicit stratification are displayed in Table 2.6 in addition to the values of the corresponding parameters. The point estimates $\hat{t}$, $\hat{r}$ and $\hat{m}$ are equal to those obtained under an SRSWOR design, but the variance estimates differ. Here, the variance estimator $\hat{v}_{2.str}(\hat{t})$ is used. The deff estimates for the total and the median are considerably smaller than one. Thus the use of implicit stratification in variance approximation under systematic sampling makes these estimates more precise when compared to variance estimators calculated under simple random sampling without replacement. The deff estimate of the ratio, however, is greater than one, indicating that no gain was reached from implicit stratification.

Let us consider more closely the variance approximations for the total $\hat{t}$. The point estimate for the total T of course remains the same under all the approximations and is $\hat{t} = 23\,580$. There are two variance estimators under the stratification assumption: the one ($\hat{v}_{2.str}$) based on implicit stratification and the other, $\hat{v}_{2.sys}$, based on successive differences. Put together, the following approximate variance estimates are obtained:

$$\hat{v}_{1.sys}(\hat{t}) \doteq N^2 \left(1 - \frac{n}{N}\right) \hat{s}^2/n = 13\,549^2 \qquad \text{deff} = 1.00$$

$$\hat{v}_{2.sys}(\hat{t}) \doteq N^2 \left(1 - \frac{n}{N}\right) (1/n) \sum_{i=2}^{n} a_i^2/2(n-1) = 13\,220^2 \qquad \text{deff} = 0.95$$

$$\hat{v}_{2.str}(\hat{t}) \doteq \sum_{h=1}^{2} \hat{v}(\hat{t}_h) = 11\,802^2 \qquad \text{deff} = 0.76$$

$$\hat{v}_{3.sys}(\hat{t}) \doteq N^2 \left(1 - \frac{n}{N}\right) (\hat{s}^2/n)[1 + 2/\log(\hat{\rho}_q) + 2/(\hat{\rho}_q^{-1} - 1)] = 8224^2$$

$$\text{deff} = 0.35$$

$$\hat{v}_{4.sys}(\hat{t}) = \hat{v}_{5.sys}(\hat{t}) \doteq N^2 \left(1 - \frac{n}{N}\right) \sum_{l=1}^{m} (\overline{y}_l - \overline{\overline{y}})^2/m(m-1) = 12\,959^2$$

$$\text{deff} = 0.87.$$

Table 2.6 Estimates from a systematic sample drawn from the *Province'91* population using implicit stratification.

Statistic	Variables	Parameter	Estimate	s.e	c.v	deff
Total	UE91	15 098	23 580	11 802	0.50	0.76
Ratio (%)	UE91, LAB91	12.65%	12.34%	0.33%	0.03	1.29
Median	UE91	229	198	27	0.14	0.21

Of the approximate variance estimates, the value of $\hat{v}_{1.sys}$, being based on an assumption of SRSWOR, is the largest. The others fall more or less below it. This could indicate that, in this case, systematic sampling is more efficient than simple random sampling. The most efficient approximation method turns out to be autocorrelative modelling, which gave the value deff = 0.35. This model is based on the assumption of an autocorrelated superpopulation, of which the fixed population constitutes one realization. The design effect turns out to be DEFF = 0.55, confirming the result.

The results on variance estimation can be evaluated by studying the properties of the intra-class correlation coefficient ρ_{int}, which is the single design parameter under systematic sampling, and the efficiency of this sampling scheme. Moreover, it is illustrated how the sorting order in the frame register is related to the value of the intra-class correlation coefficient.

Intra-class Correlation

Systematic sampling is our first example of a design where a design parameter exists. This parameter, called the *intra-class correlation coefficient* ρ_{int}, will be included in the design variance V_{sys} of an estimator. The magnitude of the intra-class correlation, and consequently its effect on variance estimates, depends partly on the selected sampling interval and partly on whether there is a successive system of ordering the study variable's values in the population frame. Under systematic sampling, the design variance of $\hat{t}$ was given in (2.13) as $V_{sys}(\hat{t}) = N^2 \sum_{j=1}^{q} (\overline{Y}_j - \overline{Y})^2/q$. The design variance can also be written as

$$V_{sys}(\hat{t}) = \sum_{j=1}^{q} (N\overline{Y}_j - N\overline{Y})^2/(N/n) = N \times \sum_{j=1}^{q} n \times (\overline{Y}_j - \overline{Y})^2. \tag{2.19}$$

Let us analyse the design variance (2.19) in more detail. First we decompose population variance into the variation between the systematic samples and the variation within the systematic samples, as in standard one-way analysis of variance. In ANOVA terms, we have

$$SST = SSW + SSB, \tag{2.20}$$

where *SST* represents the total sum of squares, *SSW* the within sum of squares and *SSB* the between sum of squares. The decomposition (2.20) can be written as

$$\sum_{k=1}^{N} (Y_k - \overline{Y})^2 = \sum_{j=1}^{q} \sum_{k=1}^{n} (Y_{jk} - \overline{Y}_j)^2 + \sum_{j=1}^{q} n(\overline{Y}_j - \overline{Y})^2. \tag{2.21}$$

Thus, an alternative form for design variance is $V_{sys}(\hat{t}) = N \times SSB$.

By using the decomposition of the total sum of squares (2.20), the intra-class correlation is defined as

$$\rho_{int} = 1 - \frac{n}{n-1} \times \frac{SSW}{SST}. \tag{2.22}$$

If the variance between the means is zero, or $SSB = 0$, then the intra-class correlation reaches its minimum $-1/(n-1)$ and, correspondingly, where $SSW = 0$ it reaches its maximum, or $\rho_{int} = 1$.

Further, we can write the variance of the total estimator in the form

$$V_{sys}(\hat{t}) = N^2 \left(1 - \frac{n}{N}\right) \frac{S^2}{n}[1 + (n-1)\rho_{int}], \tag{2.23}$$

or alternatively as the product of the SRSWOR design variance times a correction factor including the intra-class correlation coefficient as a correction factor

$$V_{sys}(\hat{t}) = V_{srs}(\hat{t}) \times [1 + (n-1)\rho_{int}].$$

Hence, the design effect is

$$\text{DEFF}_{sys}(\hat{t}) = \frac{V_{sys}(\hat{t})}{V_{srs}(\hat{t})} \doteq 1 + (n-1)\rho_{int}. \tag{2.24}$$

Systematic sampling compared with simple random sampling with replacement is

1. more efficient, if $-1/(n-1) < \rho_{int} < 0$,
2. equally efficient, if $\rho_{int} = 0$, or
3. less efficient, if $0 < \rho_{int} < 1$.

This can be interpreted to mean that the more heterogeneous the sampling intervals (i.e. negative intra-class correlation), the more efficient systematic sampling will be. Therefore, in systematic sampling there is a connection between the design parameter ρ_{int} and the sorting order of the frame population, a fact that can be successfully utilized in practice.

Example 2.4

Intra-class correlation (ρ_{int}) in the *Province'91* population. We will now calculate the intra-class correlation under systematic sampling from the *Province'91* population, where the total of UE91 is to be estimated. The intra-class correlation is calculated for systematic sampling involving a single systematic sample of eight (8) elements. The decomposition of the total sum of squares (2.21) is given in Table 2.7.

Hence, the intra-class correlation coefficient is

$$\rho_{int} = 1 - \frac{n}{n-1}\frac{SSW}{SST} = 1 - \frac{8}{8-1} \times \frac{162.14 \times 10^5}{171.32 \times 10^5} = -0.082.$$

Table 2.7 Population ANOVA Table; Systematic sampling $q = 4$ and $n = 8$.

Source of variation	df	Sum of squares	MSE
Between samples	3	$SSB = 9.18 \times 10^5$	$MSB = 3.06 \times 10^5$
Within samples	28	$SSW = 162.14 \times 10^5$	$MSW = 5.79 \times 10^5$
Total	31	$SST = 171.32 \times 10^5$	$S^2 = 5.53 \times 10^5 = 743^2$

Because the intra-class correlation is negative, systematic sampling will be more efficient in this case than simple random sampling without replacement. Thus, the design effect is

$$\mathrm{DEFF}_{sys}(\hat{t}) \doteq 1 + (n-1)\rho_{int} = 1 + (8-1) \times (-0.082) = 0.426,$$

which shows that systematic sampling is very efficient in this case.

Next, we examine in more detail the efficiency of systematic sampling under different model assumptions or assumptions on the sort order of the population, considered earlier for a given sample. We now use the corresponding design variances.

Example 2.5

Implicit stratification and DEFF. In the *Province'91* population, the urban municipalities in the province occur first, followed by the rural municipalities, both in alphabetic order. Thus, the order of the list involves two implicit strata. In the first stratum, there are the urban municipalities, which are relatively large in terms of population and, thus, also in terms of the number of unemployed. Consequently, there will be a slightly declining trend with the order of ID numbers. The corresponding scatterplot (Figure 2.4) shows the dependence of the study variable UE91 on the sort order of the elements in the population.

The dependence of the values of UE91 on the list order has certain implications for selecting a proper variance estimator.

1. The dispersion figure clearly shows that the successive order is not random, and thus it is not fair to consider this sample as a simple random sample. We found this out earlier when calculating $\mathrm{DEFF}_{sys}(\hat{t}) = 0.554 < 1$. Thus, the SRSWOR design variance $V_{srs}(= 7283^2)$ would distinctly overestimate the design variance $V_{sys}(= 5420^2)$.

2. The population is ordered successively by stratum in the register. The following stratum sizes and means of UE91 can be calculated for the implicit strata:

Stratum	ID	Size	Mean
1. Urban	1–7	7	1146
2. Rural	8–32	25	283
Whole population	1–32	32	472

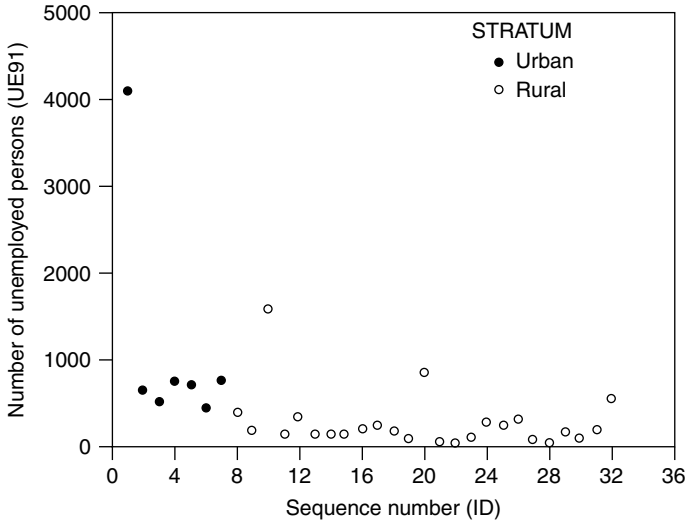


Figure 2.4 Plot of UE91 versus sequence number (ID) for the *Province'91* population. Implicit stratification to two strata is indicated.

Systematic sampling reveals these implicit strata and draws a sample that corresponds to a proportionally stratified sample (STR). If the stratum weights are known, the sample can be analysed as a poststratified sample, as considered in Section 3.3.

The design effect under stratified sampling would be

$$\mathrm{DEFF}_{sys,str}(\hat{t}) = 6251^2/7283^2 = 0.737,$$

hence this stratification makes estimation efficient. Hence, this approximation also overestimates the true design variance.

3. A linear trend exists between the study variable and identification number that can be modelled by a simple linear regression

$$Y_k = 1070.72 - 36.30 \times \mathrm{ID}_k.$$

The squared multiple correlation coefficient for this model is $R^2 = 0.21$. Using this regression model as auxiliary information in the actual estimation, we could use regression estimation (see Section 3.3). For example, the design effect under regression estimation would be

$$\mathrm{DEFF}_{srs,reg}(\overline{y}) \doteq 1 - R^2 = 1 - 0.21 = 0.79,$$

which falls in the interval $0.554 < 0.79 < 1$, where 0.554 is the exact DEFF for $\hat{t}$ under systematic sampling.

4. The listing order of the municipalities also includes autocorrelative dependence between the successive municipalities. Using the sampling interval $q = 4$ as the lag, the coefficient of autocorrelation turns out to be $\rho_4 = 0.09085$, so that the design effect under this autocorrelation would be

$$\mathrm{DEFF}_{srs,autocor}(\overline{y}) \doteq 4405^2/7283^2 = 0.366,$$

which is very close to the exact design effect 0.426 under systematic sampling. In the case of an autocorrelated situation, the only disadvantage appears if the frame population contains harmonic variation with a period corresponding to the sampling interval. This was not the situation here.

5. Pre-sorting of the register and efficiency of systematic sampling. Frame registers are usually presented as computer databases that can be sorted by desired variables. A sorting procedure affects the contents of the sampling intervals, but is not so damaging to the efficiency of estimation as might be expected. For example, the *Province'91* population was sorted by the number of unemployed in decreasing order in order to achieve a monotonic trend. Further, the internal order of the sampling intervals was alternated so that the number of unemployed was decreasing in every second sampling interval and increasing at every other interval. In this way, we achieved an optimal order of the frame population with respect to systematic sampling. The corresponding design variance is $V_{sys,opt}(\hat{t}) = 2348^2$ and $\mathrm{DEFF} = 0.104$, which indicates that the advantage of sorting is substantial in this case. Nonetheless, sorting to achieve certain implicit stratification is often used in large-scale surveys.

Summary

Systematic sampling is easy to accomplish from a computerized frame register and therefore it is very commonly used in practice. The problem, however, is the estimation of the design variance of an estimator under systematic sampling. One solution is to use auxiliary information already available in the frame population. If reasonable, it can be assumed that the population elements are in completely random order in the register and then the estimators under simple random sampling can be used. However, if certain structures such as implicit stratification, trend or periodicity of the study variable is present in the register, it is more efficient to use this information in the estimation, by using the corresponding approximative variance estimator. In our case, the estimates obtained using these approximative estimators were closer to the exact design variance than those produced by the estimator from SRSWOR, because a certain structure was present in the population. Particularly when working with a large systematic sample, it is worth trying out techniques based on the reuse of the selected sample, leading to other approximative variance estimators. Wolter (1985) offers a more comprehensive study of variance estimation under systematic sampling; he points

out that it is worthwhile to try alternative variance estimators in order to select the most appropriate for the situation at hand.

We have dealt rather broadly with systematic sampling because of its popularity in practice, and because it involves an interesting design parameter, i.e. intra-class correlation. The design parameter is not essential as such, but has a particular effect on variance estimation, and thus on the specification of sampling error, confidence limits and sizes of tests. Consequently, the main lines of approximative variance estimation were provided and supplemented by an excursion to model-assisted estimation.

2.5 SELECTION WITH PROBABILITY PROPORTIONAL TO SIZE

Situations can be met where the population contains a number of elements that have an extremely large value for the study variable. This is often the case in business surveys. A suitable sampling technique in such a case, especially for the estimation of a total, is one in which the inclusion probability depends on the size of the population element. Reduction in variance can then be expected if the size measure and the study variable are closely related. Because this sampling technique is based on inclusion probabilities proportional to relative sizes of the population elements, it is called *sampling with probability proportional to size* (PPS).

In PPS sampling, inclusion probabilities will vary according to the relative sizes of the elements. The size of a population element is measured by an auxiliary positive-valued variable z. It is assumed that the value Z_k of the auxiliary variable is known for each population element k, since the relative size equals the quotient $p_k = Z_k/T_z$, where T_z is the population total of the auxiliary variable or more precisely $T_z = \sum_{k=1}^{N} Z_k$. Commonly used size measures are variables that physically measure the size of a population element. In business surveys, for example, the number of employees in a business firm is a convenient measure of size, and in a school survey the total number of pupils in a school is also a good size measure.

The auxiliary variable z is selected such that its own variability resembles that of the study variable y. More precisely, a size measure z is sought whose ratio to the value of the study variable is, as close as possible, a constant. This is because the efficiency under PPS depends on the extent that the ratio Y_k/Z_k remains a constant C, for all the population elements. If the ratio remains nearly a constant, then the design variance of an estimator will be small.

In PPS sampling, the inclusion probabilities π_k are proportional to the relative sizes $p_k = Z_k/T_z$ of the elements, and the individual weighting of the sampled elements is based on the inverse values of these relative sizes. It is possible to draw a PPS sample either without or with replacement. Calculation of the inclusion probabilities is easier to manage under with-replacement-type sampling. Obtaining these probabilities can be complicated in without-replacement-type

PPS sampling because when the first element is sampled, the relative size of the remaining $(N-1)$ elements is changed and then new inclusion probabilities should be calculated. Various techniques have been developed to overcome this difficulty, and PPS sampling can be very efficient, especially for the estimation of the total, if a good size measure is available.

Sample Selection

A number of sampling schemes have been proposed for selecting a sample with probability proportional to size. The starting point is knowledge of the values of the *auxiliary variable z* for each population element so that probabilities of selection can be calculated. The *inclusion probability* π_k for a population element k is proportional to the relative size Z_k/T_z. For example, in the trivial case of simple random sampling with replacement, the relative sizes are $p_k = 1/N$ for each k. The quantity $1/N$ is also called the *single-draw selection probability* of a population element k. The inclusion probability of an element for a sample of size n would be $\pi_k = n \times p_k = n/N$. But in PPS sampling, the inclusion probabilities π_k vary and, thus, it is not an equal-probability sampling design in contrast to simple random sampling and systematic sampling.

In practice, the selection of a PPS sample can be based on the relative sizes of the population elements or, alternatively, on the cumulative sum of size measures. The cumulative total for the kth element is

$$G_k = \sum_{j=1}^{k} Z_j, \quad k = 1, \ldots, N, \quad G_N = T_z.$$

The natural numbers $[1, G_1]$ are associated with the first population element, and the numbers $[G_1 + 1, G_2]$ with the second element; generally, the kth element receives the numbers belonging to the interval $[G_{k-1} + 1, G_k]$. The sample selection process is based on these figures.

We consider five specific selection schemes for PPS sampling. These are *Poisson sampling*, which resembles Bernoulli sampling, the *cumulative total method* with replacement or without replacement, *systematic sampling with unequal probabilities* and the *Rao–Hartley–Cochran* method (RHC method; Rao *et al.* 1962). Of these, the cumulative total method with replacement and systematic sampling with unequal probabilities are considered in more detail. In the examples, the variable HOU85 measures the size of a population element. It is register-based and gives the number of households in each population municipality.

Poisson sampling This sampling scheme uses a list-sequential selection procedure. First the inclusion probabilities $\pi_k = n \times Z_k/T_z$ are calculated. Then, let $\varepsilon_1, \ldots, \varepsilon_k, \ldots, \varepsilon_N$ be independent random numbers drawn from the uniform (0,1) distribution. If $\varepsilon_k < \pi_k$, then the element k is selected. This procedure is applied to all population elements $k = 1, \ldots, N$, in turn.

Obviously, under Poisson sampling, the sample size is not fixed in advance but is a random variable. The expectation of the sample size is $E(n_s) = \sum_{k=1}^{N} \pi_k$. Poisson sampling is sometimes used in business surveys for sample coordination purposes (see Ohlsson, 1998).

PPS sampling with replacement (PPSWR) Sample selection with replacement has its own value in the evaluation of the statistical properties of estimators, since the corresponding design variance formulae are tractable. PPS sampling with replacement is rather like simple random sampling with replacement. The difference between these two methods is due to the way that selection numbers are assigned to population elements. In simple random sampling, a single number from the set of natural numbers $1, \ldots, k, \ldots, N$ is assigned to a population element. In PPS sampling, on the other hand, a corresponding interval from the set of numbers $1, \ldots, G_k, \ldots, G_N$ is assigned to an element, where G_k are cumulative totals.

PPS sampling with replacement is performed by first producing a single random number from the interval $[1, G_N]$. This number is then compared to the numbers associated with the population elements. An element whose selection interval includes this random number will be drawn. The single-draw selection probability of an element is thus $p_k = Z_k/T_z$. The procedure is repeated until the desired number n of draws are completed. Over all the draws, the inclusion probability of element k in the sample is $\pi_k = n \times p_k$. It should be noted that under with-replacement sampling the same population element may be selected several times. This is especially true for those population elements whose size is large, because their selection probabilities will also be large.

PPS sampling without replacement (PPSWOR) When selecting without replacement, a new problem arises concerning the computation of inclusion probabilities. With the selection of the first element, the *single-draw probability* is exactly $\pi_k = p_k = Z_k/T_z$. When the first sample element has been selected, the single-draw selection probability changes because the total T_z of the remaining $N - 1$ elements in the population decreases. Particularly for large samples, the calculation of inclusion probabilities becomes tedious. For this reason, numerous alternative without-replacement sample selection techniques have been developed to overcome this difficulty. For example, the population can be divided into a number of non-overlapping subpopulations or strata. Then, two elements are drawn without replacement from each stratum, as in the methods by Brewer (1963) and Murthy (1957). Alternatively, more than two units can be drawn from each stratum, as in Sampford's method (1967). We will discuss in greater detail two methods that enable the selection of a PPS sample of size two or more elements without replacement.

Systematic PPS sampling (PPSSYS) This method is the easiest to operate under without-replacement-type selection with probability proportional to size. In this

method, the properties of systematic sampling and sampling proportional to size are combined into a single sampling scheme. In ordinary systematic sampling, the sampling interval is determined by the quotient $q = N/n$. In systematic PPS sampling, the sampling interval is given by $q = T_z/n$. As in the ordinary one-random-start systematic sampling, we first select a random number from the closed interval [1, q]. Let it be q_0. The n selection numbers for inclusion in the sample are hence

$$q_0, \quad q_0 + q, \quad q_0 + 2q, \quad q_0 + 3q, \ldots, q_0 + (n-1)q.$$

The population element identified for the sample from each selection is the first unit in the list for which the cumulative size G_k is greater than or equal to the selection number. Given this method, the inclusion probability of the kth element in the sample is again $\pi_k = n \times p_k$.

PPS under the Rao–Hartley–Cochran method (RHC method) The population is first divided into n subpopulations $N_1, N_2, \ldots, N_g, \ldots, N_n$ using the size measure z so that in subpopulation g the sum T_g of the size measure will be close to T_z/n. There can be varying numbers of elements in the subgroups. Next, one element is drawn from each subpopulation with selection probabilities proportional to size so that for an element k the selection probability is $p_k = \mathrm{Z}_k/T_g$. The RHC method is easily managed and suitable for various PPS sampling situations.

Estimation

Estimation should be considered separately under the with-replacement and without-replacement options. Under with-replacement sampling, the single-draw selection probability of an element remains constant (i.e. equal to the relative size p_k of the element). But under without-replacement sampling, the selection probabilities of the remaining population elements change after each draw and this causes difficulties, especially in variance estimation. To introduce the basic principles of estimation under PPS sampling, we shall consider the with-replacement case only. And as an approximation, PPSSYS, which will be extensively used in the examples, is also simplified to the with-replacement case.

To construct the estimators, the relative size p_k of population element k is required; using the size measure Z_k the relative size is

$$p_k = \frac{\mathrm{Z}_k}{\sum_{k=1}^{N} \mathrm{Z}_k} = \frac{\mathrm{Z}_k}{T_z}.$$

The quantity p_k is also the single-draw selection probability for the kth element. The inclusion probability π_k of the element k in an n-element sample is, in turn, written as

$$\pi_k = n \times p_k = n \times \frac{\mathrm{Z}_k}{T_z}.$$

The inclusion probabilities should fulfil the requirement $\pi_k \leq 1$. In the trivial case of $n = 1$, this holds true for each population element. When $n > 1$ and some population values Z_k are exceptionally large, the inclusion probabilities for some of these elements may be greater than one, $n \times Z_k / \sum_{k=1}^{N} Z_k > 1$. This conflict can be encountered in practice but fortunately it is solvable. One possibility is to set $\pi_k = 1$ for all those values of k for which $nZ_k > \sum_{k=1}^{N} Z_k$, i.e. to take these elements with certainty. In practice, single-element strata are formed from these elements. For the remaining elements, π_k is set proportional to the size measure. For example, if only one of the population elements, say the element k', is overly large in this sense, set $\pi_{k'} = 1$, and the inclusion probabilities of the $N - 1$ remaining population elements are

$$\pi_k = (n-1)\frac{Z_k}{\sum_{k=1}^{N} Z_k - Z_{k'}}, k \neq k',$$

which assures that the condition $\pi_k \leq 1$ holds. An application of this is shown in Example 2.8.

The two well-known estimators of the total for PPS samples, namely the *Horvitz–Thompson* or the HT estimator, and the *Hansen–Hurwitz* or the HH estimator, are essentially based on these probability quantities. Let us derive these estimators of the total T. Under PPS sampling without replacement, an unbiased HT estimator of T (Horvitz and Thompson, 1952) is given by

$$\hat{t}_{HT} = \sum_{k=1}^{n} \frac{y_k}{\pi_k}, \tag{2.25}$$

where π_k denotes the inclusion probability. For a with-replacement PPS scheme the corresponding HH estimator (Hansen and Hurwitz, 1943) is given by

$$\hat{t}_{HH} = \frac{1}{n}\sum_{k=1}^{n} \frac{y_k}{p_k} = \frac{1}{n}(\hat{t}_1 + \cdots + \hat{t}_k + \cdots + \hat{t}_n), \tag{2.26}$$

where each $\hat{t}_k = y_k/p_k$ estimates the total T. An estimator $\hat{r}$ of the ratio R can be derived as a ratio of two HT estimators, or as a ratio of two HH estimators. Further, in the estimation of the median M, the empirical cumulative distribution function is constructed with the inverse inclusion probabilities $1/\pi_k$ as the element weights.

The with-replacement assumption also simplifies the estimation of the design variances. For the estimator $\hat{t}_{HH}$ of the total, the design variance under PPS with replacement is

$$V_{ppswr}(\hat{t}_{HH}) = \frac{N^2}{n}\sum_{k=1}^{N} p_k \left(\frac{Y_k}{Np_k} - \overline{Y}\right)^2 = \frac{1}{n}\sum_{k=1}^{N} p_k (T_k - T)^2, \tag{2.27}$$

where $T_k = Y_k/p_k$. From (2.27) it can be inferred that if Y_k is strictly proportional to Z_k such that $Y_k/Z_k = C$ holds for each k, then the design variance would be zero—an ideal case rarely met in practice. An unbiased estimator of the variance is given by

$$\hat{v}_{ppswr}(\hat{t}_{HH}) = \frac{N^2}{n(n-1)} \sum_{k=1}^{n} \left(\frac{y_k}{Np_k} - \overline{y} \right)^2 = \frac{1}{n(n-1)} \sum_{k=1}^{n} (\hat{t}_k - \hat{t}_{HH})^2 \qquad (2.28)$$

where $\overline{Y}$ and $\overline{y}$ are the population mean and sample mean of the study variable y, respectively.

We use this variance estimator as an approximation under systematic PPS sampling. Approximative variance estimators can also be derived for the without-replacement case and for the Rao–Hartley–Cochran method, but we omit the details here and refer the reader to Wolter (1985).

Example 2.6

Estimation under systematic PPS sampling. A sample of eight ($n = 8$) municipalities is drawn with PPSSYS from the *Province'91* population such that the number of households HOU85 is used as the size measure z. The cumulative sum over the population is $T_z = 91\,753$, and under PPSSYS the sampling interval would be $q = 91\,753/8 = 11\,469$.

The largest single element 'Jyväskylä' has the value 26 881 for the variable HOU85, which is more than twice the sampling interval. Therefore, the element 'Jyväskylä' would be drawn twice, and the remaining 6 elements would be drawn from the remaining population elements (31). Such a situation is commonly managed in the following way. An element that has a size measure larger than the selection interval is drawn with certainty (but only once). For such a certainty element, the weight and the inclusion probability are one by definition. In this case, therefore, we first put 'Jyväskylä' in the first stratum and take it with certainty, and then draw 7 elements from the remaining 31 population elements from the second stratum by systematic PPS sampling. This results in the following sample of eight ($n = 8$) municipalities. Note that the sample is sorted by the size measure HOU85 in Table 2.8.

It is important for the estimation under a systematic PPS design to construct a proper weight variable. For a population element k, the weight w_k is calculated using the formula

$$w_k = \frac{1}{p_k \times n} = 91\,753/(Z_k \times n),$$

where Z_k is the value of HOU85 for element k. However, in this case 'Jyväskylä' is an element drawn with certainty, whose weight gets the value one. The element

Table 2.8 A systematic PPS sample ($n = 8$) from the *Province'91* population.

Sample design identifiers			Element	Size measure	Study variables	
STR	CLU	WGHT	LABEL	HOU85	UE91	LAB91
1	1	1.000	Jyväskylä	26 881	4123	33 786
2	10	1.004	Jyväsk.mlk.	9230	1623	13 727
2	4	1.893	Keuruu	4896	760	5919
2	7	2.173	Äänekoski	4264	767	5823
2	32	2.971	Viitasaari	3119	568	4011
2	26	4.762	Pihtipudas	1946	331	2543
2	18	6.335	Kuhmoinen	1463	187	1448
2	13	13.730	Kinnula	675	129	927

Sampling rate: (not used here)

weights of the remaining seven municipalities in stratum two are calculated by

$$w_k = \frac{1}{p_k \times n} = (91\,753 - 26\,881)/(\mathrm{Z}_k \times 7).$$

In the estimation, the other required design identifiers are the stratum identifier STR, which is one for the certainty element and two for the remaining elements. The element identifier is used for CLU, because each element is taken to be a separate cluster. In addition, the finite-population correction $\left(1 - \sum_{k=1}^{n} p_k\right)$ could also be used to make sampling resemble the without-replacement type. The estimates in Table 2.9 are produced for the total $\hat{t}_{HT}$, ratio $\hat{r}_{HT}$ and median $\hat{m}_{HT}$ of UE91. For comparison, the values of the corresponding parameters *T, R* and *M* are also displayed.

As expected, PPSSYS is very efficient for the estimation of the total. The design-effect estimate for $\hat{t}_{HT}$ is close to zero (deff = 0.004). This results from the strong linear correlation of the size measure HOU85 and the study variable UE91, and is also due to the linearity of the estimator itself. For the estimator $\hat{r}_{HT}$ of the ratio, which is a nonlinear estimator, PPSSYS is still quite efficient but much less so, however, than for the total. And for the robust estimator $\hat{m}_{HT}$ for the median, the design is slightly more efficient than simple random sampling. This is in part caused

Table 2.9 Estimates under a PPSSYS design ($n = 8$); the *Province'91* population.

Statistic	Variables	Parameter	Estimate	s.e	c.v	deff
Total	UE91	15 098	15 077	521	0.03	0.0035
Ratio (%)	UE91, LAB91	12.65%	12.85%	0.2%	0.02	0.1854
Median	UE91	229	134	188	1.401	0.92

by the property of PPS sampling that the larger elements tend to be drawn, and these represent the margin rather than the middle part of the distribution of UE91.

Efficiency of PPS Sampling

We discuss the efficiency of PPS sampling in more detail for the estimation of the total T. It can be shown that the PPS design variance $V_{pps}(\hat{t}_{HT})$ of the estimator $\hat{t}_{HT}$ is related to the finite-population regression

$$Y_k = A + BZ_k + E_k$$

of the size measure z and the study variable y where $E_k, k = 1, \ldots, N$, is the residual term.

The relationship between the residual sum of squares and the population variance is given by

$$\frac{1}{N-1} \times \sum_{k=1}^{N} (Y_k - A - BZ_k)^2 \approx S^2(1 - \rho_{yz}^2),$$

where S^2 is the population variance of y and ρ_{yz}^2 is the squared correlation coefficient of the variables y and z. The residual variation is small if the correlation is close to ± 1. Actually, this variance coincides with that considered later under regression estimation. The efficiency of PPS sampling should thus be examined under the above regression model, but strong correlation ρ_{yz} alone does not guarantee efficient estimation, as will become evident.

A simple condition for the efficiency of PPS sampling can be looked for by comparing the variances of the total estimators from SRSWR and PPSWR. It can been shown that

$$V_{srswr}(\hat{t}) - V_{ppswr}(\hat{t}_{HT}) = N^2 \mathrm{Cov}(z, y^2/z)/n.$$

Thus, PPS sampling is more efficient than SRS if the correlation of the variable pair $(z, y^2/z)$ is positive. On the other hand, it was previously noted that most efficient PPS sampling occurs if the ratio Y_k/Z_k is a constant, say C for each population element. Then the design variance $V_{ppswr}(\hat{t}_{HT})$ attains its minimum, zero. If we insert $C = Y_k/Z_k$ in the previous covariance term, it is noted that $\mathrm{Cov}(z, y^2/z)$ reduces to the covariance of z and y. Thus, the correlation of z and y^2/z is equal to that of the original variables z and y in this case. We conclude that a necessary condition of PPS sampling being more efficient than SRSWR is that the study variable y and the auxiliary variable z are positively correlated in the population. But for a sufficient condition, the ratio Y_k/Z_k should remain constant over the population. These two conditions will be examined more closely in the next example.

Example 2.7

Efficiency of PPS sampling in the *Province'91* population. To evaluate the efficiency of PPS sampling two conditions should be examined. These are the stability of the ratio Y_k/Z_k across the population and the regression fit $\hat{Y}_k = 26.657 + 0.155 \times Z_k$, which, for good efficiency, should intercept the y-axis near the origin. For these purposes, two scatterplots from the *Province'91* population are displayed and appropriate coefficients are calculated.

The variation of the ratio Y_k/Z_k in the population is displayed in Figure 2.5. PPS sampling is efficient if the ratio is close to a constant over the population, as is the case here. It can be seen that the towns in the leftmost part (ID ≤ 7) are the largest, and especially among these, the ratio Y_k/Z_k is nearly a constant. Under PPS sampling the largest elements tend to be drawn, which means efficient estimation of the total. The same property also holds for the ratio Y_k/X_k if the ratio Y_k/Z_k and the ratio X_k/Z_k are constants.

The correlation of y and z is $\rho_{yz} = 0.997$ (see Figure 2.6). Strong correlation, however, is not sufficient for efficient estimation in a PPS sample. Let us consider the extreme case where this correlation is perfect, i.e. the regression $Y_k = A + B \times Z_k$ holds exactly. Using the usual interpretation of regression coefficients, it can be shown that if A is large, i.e. the regression line intercepts the y-axis far from the origin, then SRSWR is more efficient than PPS. In the *Province'91* population,

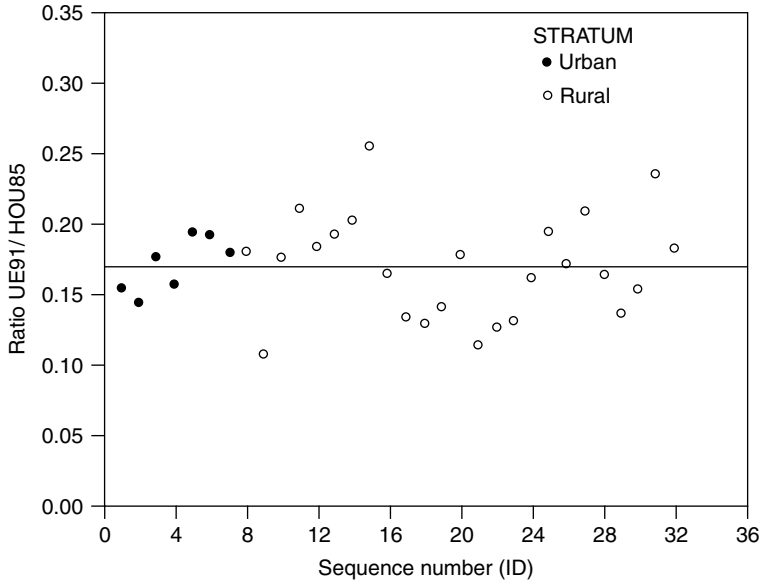


Figure 2.5 Scatterplot of the ratio UE91/HOU85 against sequence number (ID); the *Province'91* population.

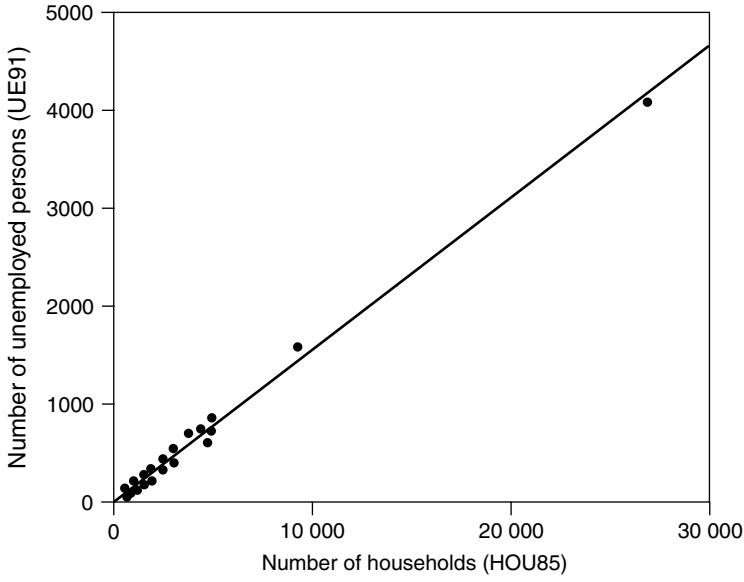


Figure 2.6 Scatterplot of UE91 against HOU85; the *Province'91* population.

the number of households HOU85 explains 99% of the variation in the number of unemployed UE91 and, moreover, the coefficient A is approximately zero, as can be seen from Figure 2.6.

Summary

Sampling with PPS provides a practical technique when sampling from populations with large variation in the values of the study variable, and often gives a considerable gain in efficiency. The efficiency of PPS sampling depends upon two things. First, efficiency varies considerably according to the type of parameter to be estimated; here these were the total, the ratio and the median. The estimation of the total appeared to be the most efficient. Under PPS sampling an auxiliary size measure (z) must be available and for efficient estimation the size measure should be strongly related to the study variable y. A condition for this is that the variable pair (z, y^2/z) is positively correlated. In the *Province'91* population this condition was satisfied, but this alone cannot guarantee efficient estimation. The ratio Y_k/Z_k must also remain constant over the population. Because this condition was satisfied in the *Province'91* population, PPS provided efficient estimation of the total. The reader who is more interested in PPS sampling is recommended to consult books by Brewer and Hanif (1983) or Hedayat and Sinha (1991, Chapter 5).

3

Further Use of Auxiliary Information

Auxiliary information recorded from the population elements can be successfully used to design a manageable and efficient sampling design and, after sample selection, to further improve the efficiency of estimators. We previously employed auxiliary information in systematic sampling (SYS) to select an appropriate variance estimator under various assumptions about the listing order of the population frame. In probability proportional sampling (PPS), auxiliary information was used in the sampling phase; an appropriate choice of an auxiliary size measure tended to considerably improve efficiency. In Sections 3.1 and 3.2, auxiliary information will be used for *stratified sampling* (STR) and *cluster sampling* (CLU). In both these techniques, auxiliary information is used to design the sampling scheme; under stratified sampling, the primary goal is to improve the efficiency, whilst in cluster sampling, the practical aspects of sampling and data collection are the main motivation for the use of auxiliary information.

Auxiliary information can be used to improve the efficiency of estimation under the sample already drawn, independent of the sampling design used. A categorical auxiliary variable could be used for poststratification, i.e. stratification of the sample after selection. If a continuous auxiliary variable is available that is strongly correlated with the study variable, it is possible to improve the efficiency by using ratio estimation or regression estimation. In these methods, auxiliary information is incorporated into the estimation procedure using statistical models. These *model-assisted techniques* are introduced in Section 3.3. The use of these techniques can considerably improve the accuracy of estimates, i.e. produce estimates that are close to the corresponding population values and, in addition, decrease the design variances of the estimators. This is demonstrated in the web extension of the book.

Auxiliary Information in Stratified Sampling

In *stratified sampling*, the target population is divided into non-overlapping subpopulations called *strata*. These are conceptually regarded as separate populations

Practical Methods for Design and Analysis of Complex Surveys Risto Lehtonen and Erkki Pahkinen
 ISBN: 0-470-84769-7

in which sampling can be performed independently. To carry out stratification, appropriate auxiliary information is required in the sampling frame. Regional, demographic and socioeconomic variables are often used as the stratifying auxiliary variables. The efficiency can benefit from stratification, because the strata are usually formed such that similar population elements, with respect to the expected variation in the values of the study variable, are collected together within a stratum. Hence, the within-stratum variation is small.

Information for the stratification can sometimes be inherent in the population. For example, strata are clearly identified if a country is divided into regional administrative areas that are non-overlapping. Separate sampling from each area guarantees the proper representation of different parts of the country in the sample. Auxiliary information of such an administrative type can be used in designing the sampling. Stratification can also be used in estimation for population subgroups or domains of interest. Important domains are then defined as separate strata, which allows the allocation of a desired sample size for each of them (see Chapter 6). Moreover, for example, regional comparisons or comparisons between the strata can also be conducted. Thus, in addition to functioning as a tool for creating internally homogeneous subpopulations, stratification can also serve as a classifying variable in the estimation and testing procedures.

Auxiliary Information in Cluster Sampling

Instead of drawing the sample directly from the element population, in *cluster sampling* a sample is drawn from the population of naturally occurring subgroups called *clusters*. Subgroups often used in practice are, for example, clusters of employees in establishments, clusters of pupils in schools and clusters of people in households. For sampling purposes, a frame of the population clusters is needed; however, it is not necessary to have a complete frame covering all the population elements, but only those elements from the sampled clusters. Recognizing the structure of the population reveals the existence of the primary sampling units. Educational surveys in which the primary sampling unit is usually a school, and a sample of schools is first drawn from a register of schools, are good examples of the use of such a structure. Moreover, the population clusters can be stratified before sample selection. Auxiliary information in cluster sampling therefore concerns not only the grouping of the population elements into clusters but also the properties of the clusters needed if stratification is desired.

In forming clusters of population elements, groups of elements are collected together, which often tend to be cluster-wise similar in the various respects relevant to the survey. This intra-cluster homogeneity tends to decrease the efficiency of estimation. However, cluster sampling can be cost-effective due to reduced fieldwork costs. Intra-cluster homogeneity involves a certain design parameter called *intra-cluster correlation*. There are two main approaches that take proper account of the intra-cluster correlation necessary for valid estimation.

Firstly, intra-cluster correlation can be taken as a nuisance effect in the estimation, with the aim being to remove this disturbance effect from the estimation and testing results. Alternatively, the clustering can be regarded as a structural phenomenon of the population to be modelled. The population is thus seen as having a hierarchical or multi-level structure. In educational surveys, for example, the first level of the structure contains the schools, the second the teaching groups, and the third or lowest level the pupils. Pupils' measured achievements are conditioned by this hierarchical structure. Modelling methods using the multi-level structure share this approach and also presuppose that the corresponding information exists in the data set. The nuisance approach and the multi-level approach are discussed in Chapter 8 and in Section 9.4, respectively.

Auxiliary Information in the Estimation Phase

Auxiliary information can be used to improve the efficiency of a given sample, by using *model-assisted estimation techniques* discussed in Section 3.3. In model-assisted estimation, the auxiliary data are incorporated in estimation by using statistical models. In *poststratification*, a linear analysis of variance or ANOVA model is assumed, and the auxiliary data consists of population cell and marginal frequencies of one or several categorical variables. *Ratio estimation* uses a linear regression model where the intercept is excluded, and the auxiliary data consists of the population totals of one or several continuous variables, which can come from a source such as official statistics. In *regression estimation*, a standard linear regression model is used to incorporate the auxiliary data in the estimation procedure. The methods are special cases of *generalized regression (GREG) estimators*. In all these methods, estimation can be more effective than that from just simple random sampling (SRS) if there is a relation between the study variable and auxiliary variable, such as a strong correlation.

3.1 STRATIFIED SAMPLING

Stratification of the population into non-overlapping subpopulations is another popular technique where auxiliary information can be used to improve efficiency. Such auxiliary information is often available in registers or databases that provide sampling frames. Typical variables used in stratification are regional (e.g. county), demographic (sex, age group) and socioeconomic (e.g. income group) variables gathered in a census. To fully benefit from the gains in efficiency of stratified sampling, it is important not only to be careful when selecting stratification variables but also to appropriately allocate the total sample to the strata.

There are several reasons for the popularity of stratified sampling:

1. For administrative reasons, many frame populations are readily divided into natural subpopulations that can be used in stratification.

2. Stratification allows for flexible stratum-wise use of auxiliary information for both sampling and estimation.
3. Stratification can enhance the precision of estimates if each stratum is homogeneous.
4. Stratification can guarantee representation of small subpopulations or domains in the sample if desired.

Estimation and Design Effect

In stratified sampling, auxiliary information is used to divide the population into H non-overlapping subpopulations of size $N_1, N_2, \ldots, N_h, \ldots, N_H$ elements such that their sum is equal to N. A sample is selected independently from each stratum, where the stratum sample sizes are $n_1, \ldots, n_h, \ldots, n_H$ elements respectively. In stratified sampling, the estimators are usually weighted sums of individual stratum estimators where the weights are stratum weights $W_h = N_h/N$. The strata can thus be regarded as mutually independent subpopulations. An estimator $\hat{t}$ for a population total T, is given by

$$\hat{t} = N \sum_{h=1}^{H} W_h \overline{y}_h = \sum_{h=1}^{H} \hat{t}_h = \hat{t}_1 + \cdots + \hat{t}_h + \cdots + \hat{t}_H, \tag{3.1}$$

where $\hat{t}_h = N_h \overline{y}_h$ is the total estimator in stratum h and $\overline{y}_h = \sum_{k=1}^{n_h} y_k/n_h$. If all the stratum totals are unbiased estimates, then the estimator of the population total is also unbiased. Because the samples are drawn independently from each stratum, the design variance $V_{str}(\hat{t}_{str})$ of $\hat{t}$ is simply the sum of stratum variances $V(\hat{t}_h)$. For example, if simple random sampling without replacement is used in each stratum, the design variance of the estimator $\hat{t}$ is

$$V_{str}(\hat{t}) = \sum_{h=1}^{H} V_{srs}(\hat{t}_h), \tag{3.2}$$

whose unbiased estimator is correspondingly

$$\hat{v}_{str}(\hat{t}) = \sum_{h=1}^{H} \hat{v}_{srs}(\hat{t}_h). \tag{3.3}$$

The design effect (DEFF) of $\hat{t}$ depends heavily on the proportion of the total variation given by the division into between- and within-stratum variance components. From the variance equation (3.2), it can be inferred that to benefit from a small design variance, internally homogeneous strata, which have small within-stratum variances, should be constructed. The efficiency is also affected by the allocation scheme, since the individual stratum variances depend on the respective stratum sample sizes. Let us consider the calculation of DEFF with the estimation of the total T using stratified sampling with *proportional allocation*

where stratum sample sizes are $n_h = n \times W_h$ and $n = \sum_{h=1}^{H} n_h$. If the elements are selected with simple random sampling without replacement (SRSWOR) within each stratum, the estimator $\hat{t}$ is unbiased for T and

$$V_{str}(\hat{t}) = N^2(1 - n/N) \sum_{h=1}^{H} W_h S_h^2/n$$

is the design variance of $\hat{t}$, where S_h^2 is the variance of y in stratum h. Alternatively, the SRSWOR variance $V_{srs}(\hat{t}) = N^2(1 - n/N)S^2/n$ of $\hat{t} = (N/n) \sum_{k=1}^{n} y_k$ can be written in terms of stratified sampling as follows. Assuming large n, we get

$$V_{srs}(\hat{t}) \doteq N^2(1 - n/N) \left[\sum_{h=1}^{H} W_h S_h^2 + \sum_{h=1}^{H} W_h (\overline{Y}_h - \overline{Y})^2 \right] /n,$$

where $\overline{Y}_h$ is the population mean in stratum h, and the first term in brackets measures the within-stratum variation and the squared differences $(\overline{Y}_h - \overline{Y})^2$ measure the variation of the stratum means around the population mean $\overline{Y}$, i.e. the between-stratum variation. The total variance is thus split into within-stratum and between-stratum variance components. Therefore, the DEFF of $\hat{t}$ is given by

$$\text{DEFF}_{str}(\hat{t}) \doteq \frac{\sum_{h=1}^{H} W_h S_h^2}{\sum_{h=1}^{H} W_h [S_h^2 + (\overline{Y}_h - \overline{Y})^2]}, \tag{3.4}$$

or by analogy with analysis of variance:

$$\text{DEFF}_{str}(\hat{t}) \doteq \frac{\text{within-stratum variance}}{\text{total variance}} = \frac{\text{MSW}}{S^2},$$

where total variance = within-stratum variance + between-stratum variance.

Example 3.1

Next, we will calculate the parameter $\text{DEFF}_{str,pro}(\hat{t})$ for stratified simple random sampling (STRSRS) with proportional allocation on the *Province'91* population. The population consists of two strata: stratum 1 for towns ($N_1 = 7$) and stratum 2 for rural municipalities ($N_2 = 25$). Using these two strata as levels of a factor in an ANOVA setting, we get a decomposition of the total variation of the study variable UE91 as presented in Table 3.1. Inserting in (3.4) the within-stratum variance component $\text{MSW} = 4.35 \times 10^5$ and the total variance $S^2 = 5.53 \times 10^5$ gives

$$\text{DEFF}_{str,pro}(\hat{t}) \doteq \frac{4.35}{5.53} = 0.79,$$

Table 3.1 Population ANOVA table for stratified SRSWOR sampling with $H = 2$ strata and $N_1 = 7$ and $N_2 = 25$.

Source of variation	df	Sum of squares	Mean square
Between strata	1	$\text{SSB} = 40.73 \times 10^5$	$\text{MSB} = 40.73 \times 10^5$
Within strata	30	$\text{SSW} = 130.60 \times 10^5$	$\text{MSW} = 4.35 \times 10^5$
Total	31	$\text{SST} = 171.32 \times 10^5$	$S^2 = 5.53 \times 10^5 = 743^2$

which is an approximation to the exact DEFF parameter calculated as $\text{DEFF}_{str,pro}(\hat{t}) = 0.84$.

Proportional allocation provides a simple allocation method. Stratified SRSWOR sampling with proportional allocation appears to be more efficient than the SRSWOR design. In the following, we will consider other allocation schemes that can be more efficient. This can be achieved by more effectively accounting for stratum-wise variances.

Allocation of Sample

Allocation provides a tool for determining the number of sample units to be taken from each stratum under the constraint that the total number of units to be sampled is n. The modest target is to find an allocation scheme which enables efficient estimation, under the rather restricted situation of a descriptive survey with one study variable. It should be noted, however, that in a large-scale analytical survey it is impossible to reach global optimality for the allocation with a stratified sampling design, because, generally, numerous study variables are present.

Optimality of the allocation depends on the stratum sizes and, more generally, on the share of the total variance of the study variable to the between-stratum and within-stratum variances. Of the many methods of allocation suggested in the literature, *optimal* or *Neyman allocation* and *power* or *Bankier allocation* will be considered, in addition to *proportional allocation*.

1. Proportional allocation This is the simplest allocation scheme and is widely used in practice. It presupposes a knowledge of the stratum sizes only, since the sampling fraction n_h/N_h is constant for each stratum. The number of sample elements n_h in stratum h is given by

$$n_{h,pro} = n \times \frac{N_h}{N} = n \times W_h,$$

where W_h is the stratum weight.

Proportional allocation guarantees an equal share of the sample in all the strata, but can produce less efficient estimates than generally expected.

As the sampling fraction is a constant n/N in each stratum, the inclusion probability of any population element k is also a constant $\pi_k = \pi = n/N$. The scheme therefore provides an equal-probability sampling design equivalent to that of SRSWOR. This property simplifies the estimation because then

$$\hat{t} = N \sum_{h=1}^{H} \sum_{k=1}^{n_h} y_{hk}/n,$$

so the within-stratum means need not be calculated. For this reason, a proportionally allocated sample has the property of *self-weighting*. This property is not present in the other allocation schemes where the inclusion probabilities vary between strata.

2. Optimal or Neyman allocation This can be used if S_h, the standard deviations for individual strata of the study variable, are known. The number of sample units n_h in stratum h under optimal allocation is given by

$$n_{h,opt} = n \frac{N_h S_h}{\sum_{h=1}^{H} N_h S_h}.$$

In practice, S_h is rarely known, but from experience gained in past surveys, close approximations to the true standard deviations may be made. In optimal allocation, a stratum which is large or has a large within-stratum variance, has more sampling units than a smaller or more internally homogeneous stratum. This type of allocation provides the most efficient estimates under stratified sampling.

3. Power allocation This is suggested for surveys in which there are numerous small strata and that also have the need for precise estimates at each stratum level. For example, under power allocation the n_h that are required to efficiently estimate stratum totals are given by

$$n_{h,pow} = n \frac{(T_{hz})^a \mathrm{C.V}_{hy}}{\sum_{h=1}^{H} (T_{hz})^a \mathrm{C.V}_{hy}},$$

where T_{hz} is the stratum total of an auxiliary variable z and $\mathrm{C.V}_{hy}$ is the coefficient of variation (C.V) of y in stratum h. The constant a is called the power of allocation and in practice, a suitable choice of a may be $\frac{1}{2}$ or $\frac{1}{3}$. This choice can be viewed as a compromise between the Neyman allocation and an allocation that leads to approximately constant precision for all strata.

Example 3.2

Different allocation schemes under stratified simple random sampling in the *Province'91* population. The population is first divided into two strata, one urban

Table 3.2 Stratum-level parameters for the variable UE91 from the *Province'91* population.

Statistic	Stratum 1	Stratum 2	All
Mean	1146	283	472
Total	8022	7076	15 098
Standard deviation	1318	331	743
Coefficient of variation	1.150	1.170	1.572
Stratum size	7	25	32

and the other rural. Of all the municipalities, seven ($N_1 = 7$) are towns and the remainder ($N_2 = 25$) are rural districts. A stratified simple random sample of eight ($n = 8$) municipalities is drawn, and the appropriate stratum sample sizes are calculated under (a) proportional, (b) optimal and (c) power allocation schemes. Certain background information for the strata is displayed in Table 3.2.

From Table 3.2, n_h for each stratum under various allocation schemes can be calculated.

(1) Proportional allocation:

$$n_{h,pro} = n\frac{N_h}{N} = \begin{cases} n_1 = (8)\dfrac{7}{32} = 1.75 \\ n_2 = (8)\dfrac{25}{32} = 6.25 \end{cases}$$

(2) Optimal allocation:

$$n_{h,opt} = n\frac{N_h S_h}{\sum_{h=1}^{H} N_h S_h} = \begin{cases} n_1 = (8)9226/(9226 + 8275) = 4.22 \\ n_2 = (8)8275/(9226 + 8275) = 3.78 \end{cases}$$

(3) Power allocation (approximate) with $a = 0$:

$$n_{h,a=0} = n \times \frac{\mathrm{C.V}_{hy}}{\sum_{h=1}^{H} \mathrm{C.V}_{hy}} = \begin{cases} n_1 = 8 \times \dfrac{1.150}{1.150 + 1.170} = 3.97 \\ n_2 = 8 \times \dfrac{1.170}{1.150 + 1.170} = 4.03 \end{cases}$$

(3') Power allocation (exact) with $a = 0$ and a stratum-specific coefficient c_h:

$$n_{h,a=0} = n \times \frac{\mathrm{C.V}_{hy}}{\sum_{h=1}^{H} \mathrm{C.V}_{hy}} \times c_h = \begin{cases} n_1 = 8 \times \dfrac{1.150}{1.150 + 1.170} \times 0.81 = 3.22 \\ n_2 = 8 \times \dfrac{1.170}{1.150 + 1.170} \times 1.19 = 4.78 \end{cases}.$$

These calculations lead to the following results. With proportional allocation, the individual stratum sample sizes are $n_1 = 2$ and $n_2 = 6$, whilst with the optimal

Table 3.3 Stratum sample sizes and coefficients of variation under different allocation schemes. Estimation of a total from an STRSRS sample ($n = 8$). *Province'91* population.

	Sample size		Stratum		Population	
Allocation	n_1	n_2	C.V($\hat{t}_1$)	C.V($\hat{t}_2$)	C.V($\hat{t}$)	DEFF
Optimal	4	4	0.38	0.54	0.32	0.44
Power (exact)	3	5	0.50	0.47	0.35	0.51
Proportional	2	6	0.68	0.42	0.42	0.74

and approximate power allocation $n_1 = n_2 = 4$. Note that the so-called *equal allocation*, in which the sample sizes in each stratum are equal ($n_h = n/H$), also gives $n_1 = n_2 = 4$. The efficiency of optimal allocation and power allocation over proportional allocation can be inferred from the corresponding DEFF values, which are 0.44 (for optimal allocation), 0.51 (for power allocation) and 0.74 (for proportional allocation).

In addition, exact power allocation has been calculated, because in the case of small populations (such as the *Province'91* population), the assumption of low sampling rate per stratum is not valid. Exact power allocation with $a = 0$ gives the following sample size per stratum: $n_1 = 3$ and $n_2 = 5$.

The allocation schemes can be compared by calculating coefficient of variation C.V($\hat{t}$) or relative standard error for the total sample and for each stratum. The results are shown in Table 3.3. On the population level, as expected, optimal allocation gives the most precise estimate C.V($\hat{t}$) = 0.32. But for equal precision on the stratum level, exact power allocation gives the best estimate because the c.v of $\hat{t}$ is about 0.5 in both strata. Proportional allocation gives poor precision on the population level and the difference between the two stratum-level coefficients of variation is substantial in this case.

As mentioned earlier, coinciding domains with strata before the sample allocation would give considerable gains in precision if power allocation (approximate or exact) were used. Domain estimation is considered in more detail in Chapter 6.

Sample Selection

Sample selection is carried out independently in each stratum, which provides an opportunity to use different selection schemes in different strata. However, for convenience the same selection scheme is often used. In STR sampling, the total population should first be stratified and then a random sample selected in each stratum. Simple random sampling, SYS or PPS sampling can be applied to individual strata.

Inclusion probabilities depend on stratum-wise sample selection methods. For example, using an SRSWOR design in all strata gives $\pi_{hk} = n_h/N_h$, where n_h is the stratum sample size and N_h is the total number of population elements

in stratum h. If PPS sampling is applied, then the inclusion probability is $\pi_{hk} = n_h \times (Z_{hk}/T_{hz})$, where T_{hz} is the stratum total $T_{hz} = \sum_{k=1}^{N_h} z_{hk}$ of a size measure z. The inclusion probabilities are needed to define appropriate sampling weights. Let us next consider stratified sampling with optimal allocation from the *Province'91* population.

Example 3.3

Stratified simple random sampling from the *Province'91* population using optimal allocation. The demonstration population is divided into two strata—rural and urban municipalities. The allocation scheme is the optimal method, which leads to equal stratum sample sizes $n_1 = n_2 = 4$, when the population total T is estimated, as previously shown. Under this allocation, a stratified simple random sample is selected (Table 3.4). Once the sample is drawn, the relevant design identifiers should be added to the data set as new variables (STR, CLU and WGHT) and used in the estimation procedure. The three estimation problems are considered as before. The estimator $\hat{t}_{str}$ of the total number of unemployed persons UE91 demonstrates clearly how stratification decreases the standard error in this case (Table 3.4). A similar effect is also noted for the ratio estimator $\hat{r}$ of the unemployment rate UE91/LAB91. For the third estimator $\hat{m}$, the median of the population distribution of UE91, no gain is achieved by using the stratification and optimal allocation.

The stratum identifier has the value STR $= 1$ for a town and STR $= 2$ for a rural municipality. Cluster identifier CLU refers to the groups of elements; here, each cluster contains a single element and the ID number of each municipality is chosen as the cluster identifier. The weight variable has to be calculated for each stratum separately from the stratum size and stratum sample-size figures. The weight,

Table 3.4 An optimally allocated stratified simple random sample from the *Province'91* population.

Sample design identifiers			Element	Study variables	
STR	CLU	WGHT	LABEL	UE91	LAB91
1	1	1.75	Jyväskylä	4123	33786
1	2	1.75	Jämsä	666	6016
1	4	1.75	Keuruu	760	5919
1	6	1.75	Suolahti	457	3022
2	21	6.25	Leivonmäki	61	573
2	25	6.25	Petäjävesi	262	1737
2	26	6.25	Pihtipudas	331	2543
2	27	6.25	Pylkönmäki	98	545

Sampling rates: Stratum $1 = 4/7 = 0.57$. Stratum $2 = 4/25 = 0.16$.

Table 3.5 Estimates from an optimally allocated stratified simple random sample ($n = 8$); the *Province'91* population.

Statistic	Variables	Parameter	Estimate	s.e	c.v	deff
Total	UE91	15 098	15 211	4286	0.28	0.21
Ratio (%)	UE91, LAB91	12.65%	12.78%	0.3%	0.02	0.38
Median	UE91	229	177	64	0.36	0.19

WGHT, for the first stratum is $w_{1k} = N_1/n_1 = 7/4 = 1.75$ and for the second is $w_{2k} = N_2/n_2 = 25/4 = 6.25$. In addition, for simple random sampling without replacement, sampling rates for each stratum are needed, given by $4/7 = 0.57$ for the first stratum and $4/25 = 0.16$ for the second.

The estimation results with the values of the corresponding population parameters are shown in Table 3.5. The point estimates $\hat{t}$ and $\hat{r}$ for the total and the ratio are close to the values of the population parameters T and R. However, for the median, the estimate $\hat{m} = 177$ deviates considerably from the true median $M = 229$. The optimally allocated stratified SRSWOR design seems to be very efficient for the estimation of the total and the ratio in this case, with design-effect estimates $\text{deff}(\hat{t}) = 0.21$ and $\text{deff}(\hat{r}) = 0.38$. However, the estimation of the median is more efficient than under the unstratified SRSWOR design, because $\text{deff}(\hat{m}) = 0.19$ is considerably less than one.

Finally, the stratum-wise precision or c.v is calculated for the total estimate of variable UE91. The estimates of totals are $\hat{t}_1 = 10\,507$ for the first stratum and $\hat{t}_2 = 4700$ for the second stratum. Corresponding standard error estimates are s.e $(\hat{t}_1) = 4015$ and s.e $(\hat{t}_2) = 1481$. Then the c.v estimates are c.v $(\hat{t}_1) = 0.38$ and c.v $(\hat{t}_2) = 0.32$ which are about the same size.

Summary

A small population split into two strata was considered with various allocation schemes. In estimating the total, ratio and median, stratified sampling with optimal allocation produced deff estimates that indicated gain in efficiency for the total and ratio estimators; however, the estimated median had a deff estimate greater than one. Generally, the overall gain of precision attained in stratified sampling depends on the stratification scheme and on the allocation of the sample between the strata. At stratum level, precision can be affected by a suitable allocation scheme; this is important especially if estimates are to be calculated for separate strata.

Stratification provides a powerful tool for improving the efficiency and, being suitable for various sampling situations, it is commonly used in practice. In addition to element sampling, stratified sampling is often present in sampling designs for complex surveys where the population of clusters is stratified.

3.2 CLUSTER SAMPLING

In complex surveys, naturally formed groups of population elements such as households, villages, city blocks or schools are often used for sampling and data collection. For example, a household can be chosen as the unit of data collection in an interview survey. In addition to the original person-level population, there is the additional population of households. Assuming that a suitable frame is available, a sample of households is drawn for the interviewing of the sample household members. This is an example of *one-stage cluster sampling*. If a household population frame is not available but a block-level frame is, a sample from the register of blocks can be drawn, and a sample of households can then be drawn from the sampled blocks by using lists of dwelling units prepared from only the selected blocks. This is an example of *two-stage cluster sampling*.

Cluster sampling in social and business surveys is motivated by the need for practical, economic and sometimes also administrative efficiency. An important advantage of cluster sampling is that a sampling frame at the element level is not needed. The only requirements are for cluster-level sampling frames and frames for subsampling elements from the sampled clusters. Cluster-level frames are often easily accessible, for example, for establishments, schools, blocks or block-like units etc. Moreover, these existing structures provide the opportunity to include important structural information as part of the analysis. For instance, in an educational survey it is practical to use the information that pupils are clustered within schools and further clustered as classes or teaching groups within schools. Schools can be taken as the population of clusters from which a sample of schools is first drawn and then a further sample of teaching groups can be drawn from those schools that have been sampled. If all the pupils in the sampled teaching groups are measured, then the design belongs to the class of two-stage cluster-sampling designs. And in addition to sample selection and data collection, the multi-level structure can be used in the analysis, for example, for examining differences between schools.

Thus, in *multi-stage sampling*, a subsample is drawn from the sampled clusters at each stage except the last. At this stage, all the elements from the sampled clusters can be taken in an element-level sample, or a subsample of the elements can be drawn. One- and two-stage cluster sampling are discussed in this chapter and demonstrated using the *Province'91* population. A more general setting for cluster sampling, also covering stratification of populations of clusters, will be demonstrated by various real surveys in Chapters 5 to 9.

The economic motivation for cluster sampling is the low cost of data collection per sample element. This is especially true for populations that have a large regional spread. Using cluster sampling, the travelling costs of interviewers can be substantially reduced as the workload for an interviewer can be regionally planned. The *cost efficiency* of cluster sampling can therefore be high. But there are also certain drawbacks of cluster sampling that concern statistical efficiency. If each cluster closely mirrors the population structure, we

would attain efficient sampling such that standard errors of estimates would not exceed those of simple random sampling. However, in practice, clusters tend to be internally homogeneous, and this homogeneity increases standard errors and thus decreases *statistical efficiency*. We shall consider this more closely by considering intra-cluster correlation. This concept will be used extensively in later chapters when analysing real data sets from cluster sampling using two approaches: by taking the intra-cluster correlation as a nuisance effect and by multi-level modelling methods.

Cost Efficiency in Cluster Sampling

Let us first use a simple case to illustrate the cost efficiency of cluster sampling relative to SRS without replacement. The cost efficiency of cluster sampling can be assessed by a simple cost function

$$C_{clu} = c_1(m) + c_2(m \times B),$$

where

$C_{p(s)}$	is the total sampling costs,
c_1	is the sampling cost for a cluster,
c_2	is the sampling cost for an element in a cluster,
B	is the number of elements in a cluster (equal-sized clusters),
m	is the number of sample clusters,
$n = m \times B$	is the element sample size.

Under SRSWOR the cost function is

$$C_{srs} = c_1 n + c_2 n,$$

where n is the element sample size.

The constraint of equal total sampling costs $C = C_{clu} = C_{srs}$ requires the following sample sizes for SRSWOR and CLU sampling:

$$n_{srs} = \frac{C}{c_1 + c_2}$$
$$n_{clu} = \frac{C}{(1/B)c_1 + c_2},$$

indicating that with a fixed sampling cost more population elements can be measured using cluster sampling than using SRSWOR. Moreover, standard errors decrease inversely with square root of sample size, which in part compensates for the counter-effect of intra-cluster homogeneity upon standard errors. This implies that the DEFF cannot serve as a single measure of the total efficiency of cluster sampling, so cost efficiency should also be taken into account.

Example 3.4

Cost efficiency under cluster sampling. The budget of a nationwide survey based on computer-assisted personal interviews (CAPI) includes a grant of EUR 15 000 to cover sampling and data-collection costs. Costs per interview are EUR 30 and average travelling expenses per interview are EUR 35. By first assuming that the sample is drawn by SRSWOR, the sample size under fixed total costs is

$$n_{srs} = \frac{15\,000}{35 + 30} = 231.$$

Next, assuming that the population can be split into clusters each consisting of five people ($B = 5$), the sample size is

$$n_{clu} = \frac{15\,000}{35/5 + 30} = 405$$

Cluster sampling nearly doubles the available sample size relative to SRSWOR, since the costs of a single journey will cover five interviews instead of one.

One-stage Cluster Sampling

Let us introduce the principles of cluster sampling under the simplest design of this sort, namely, one-stage cluster sampling. In one-stage cluster sampling, it is assumed that the N population elements are clustered into M groups, i.e. clusters. Making the somewhat unrealistic assumption of equal-sized clusters, each cluster is taken to consist of B elements. In a more general case, it is assumed that the population is clustered such that the size of cluster i is B_i elements. In both cases, a sample of m clusters is drawn from the population of M clusters, and all the elements of the sampled clusters are taken into the element-level sample. Remember, there is only a single sampling stage, namely that of the clusters, and therefore this design is known as *one-stage cluster sampling*.

The sample of m clusters is drawn from the population of clusters using a specific element-sampling technique such as SRS, SYS or PPS sampling. Because standard element-sampling schemes can be used in one-stage cluster sampling, the selection techniques previously described are readily available. The only difference is that a cluster, i.e. a group of population elements, constitutes the sampling unit instead of a single element of the population. Moreover, if the selection of the clusters is done with equal inclusion probabilities, for example, using SRSWOR or SYS, then the inclusion probabilities for the population elements are also equal, and this is independent of cluster sizes being equal.

In the simple case of equal-sized clusters, the element sample size is fixed and is $n = m \times B$. If the cluster sizes vary, as is often the case in practice, the sample size $n = \sum_{i=1}^{m} B_i$ cannot be fixed in advance and depends upon which clusters happen

to be drawn in the sample. The expected element-level sample size $(m/M) \times N$ and the actual sample size n can differ considerably if the variation in cluster sizes is large. This inconvenience can usually be controlled using an appropriate sampling scheme. For example, if the sizes of the population clusters are (even roughly) known as auxiliary information, the clusters can be stratified by size, making it possible to approximately control the element sample size, n.

We introduce the basics of the estimation under one-stage cluster sampling in the case in which M unequal-sized clusters are present with cluster sizes B_i, and SRSWOR is used to sample the m clusters; we call this one-stage CLU design. Equally sized clusters where $B_i = B$ is a special case of this. The element-level population size is thus given by $N = \sum_{i=1}^{M} B_i$ elements. Our aim is to estimate the population total T. For this, formulae from simple random sampling in Section 2.3 can be used and applied to the cluster totals. Certain alternative estimators are also given.

Let the value of the study variable be denoted Y_{ik}, $i = 1, \ldots, M$ in the population, and in the sample y_{ik}, $i = 1, \ldots, m$, and in both instances $k = 1, \ldots, B_i$. The cluster-wise totals T_i in the population are

$$T_i = \sum_{k=1}^{B_i} Y_{ik} = B_i \overline{Y}_i, \qquad i = 1, \ldots, M,$$

where $\overline{Y}_i$ is the mean per element in population cluster i, whose sample estimator is $\overline{y}_i = \sum_{k=1}^{B_i} y_{ik}/B_i$, $i = 1, \ldots, m$.

An unbiased estimator of the population total $T = \sum_{i=1}^{M} T_i$ is given by

$$\hat{t} = (M/m) \sum_{i=1}^{m} B_i \overline{y}_i. \tag{3.5}$$

The design variance $V_{clu-1}(\hat{t})$ of $\hat{t}$ and its unbiased estimator $\hat{v}_{clu-1}(\hat{t})$ can be derived from the corresponding SRSWOR equations, because the only source of variation is that of the cluster totals T_i around the overall mean per cluster $\overline{T}_M = \sum_{i=1}^{M} T_i/M$.

The design variance of $\hat{t}$ is given by

$$V_{clu-1}(\hat{t}) = M^2(1 - m/M) \sum_{i=1}^{M} (T_i - \overline{T}_M)^2/m(M-1). \tag{3.6}$$

An unbiased estimator of the design variance is

$$\hat{v}_{clu-1}(\hat{t}) = M^2(1 - m/M) \sum_{i=1}^{m} (B_i \overline{y}_i - \hat{\overline{T}}_m)^2/m(m-1), \tag{3.7}$$

where $\hat{T}_m = \sum_{i=1}^{m} B_i \overline{y}_i/m$ is an estimator of the mean per cluster $\overline{T}_M$.

It can be inferred from (3.6) that if the cluster sizes B_i are equal or nearly so and if the cluster means $\overline{Y}_i$ vary little, then the cluster totals $T_i = B_i\overline{Y}_i$ will also vary little and so a small design variance will be obtained. On the other hand, if the variation in the cluster sizes is large, the cluster totals will vary greatly and the design variance becomes large, showing inefficient estimation. However, the efficiency can be improved using a ratio estimator where the cluster sizes B_i are used as an auxiliary size measure z. We can then have an estimator for the total given by

$$\hat{t}_{rat} = N\frac{\sum_{i=1}^{m} T_i}{\sum_{i=1}^{m} B_i} = N \times \overline{y}, \tag{3.8}$$

where $\overline{y} = \sum_{i=1}^{m} T_i / \sum_{i=1}^{m} B_i$ is the sample mean per element, which is an estimator of the population mean per element $\overline{Y} = T/(M \times B)$. This ratio estimator is a special case of the ratio estimator considered later in Section 3.3. Assuming a large number of sample clusters, an approximate design variance of $\hat{t}_{rat}$ is

$$V_{clu-1}(\hat{t}_{rat}) \doteq M^2(1 - m/M)\sum_{i=1}^{M} B_i^2(\overline{Y}_i - \overline{Y})^2/m(M-1). \tag{3.9}$$

The variation in the cluster means per element $\overline{Y}_i$ around the population mean per element $\overline{Y}$ can usually be expected to be smaller than that of the cluster totals T_i around the mean per cluster $\overline{T}_M = \sum_{i=1}^{M} T_i/M$. If so, the estimation will be more efficient. Hence, an estimator of the design variance is

$$\hat{v}_{clu-1}(\hat{t}_{rat}) = M^2(1 - m/M)\sum_{i=1}^{m} B_i^2(\overline{y}_i - \overline{y})^2/m(m-1). \tag{3.10}$$

A similar effect on the efficiency can be expected when using PPS sampling for the clusters if we know their sizes B_i in advance. Then, one can use the corresponding PPS estimators from Section 2.5.

It is also possible to base the estimation of the total T on the mean $\overline{y}_m$ of the cluster means $\overline{y}_i$ given by

$$\overline{y}_m = \sum_{i=1}^{m} \overline{y}_i/m,$$

which is an estimator of the population mean $\overline{Y}_M = \sum_{i=1}^{M} \overline{Y}_i/M$ of the cluster means. If the clusters are equal-sized, i.e. if $B_i = B$, then the resulting estimator

$$\hat{t}_m = N\overline{y}_m = N\sum_{i=1}^{m} \overline{y}_i/m \tag{3.11}$$

is unbiased for T and equal to $\hat{t}$ given in (3.5), and $\hat{t}_{rat}$ given in (3.8). But the $\hat{t}_m$ can be biased and even inconsistent under unequal-sized clusters. This can be seen by looking more closely at the bias. The bias is given by

$$\text{BIAS}(\hat{t}_m) = -\sum_{i=1}^{M}(B_i - \overline{B})(\overline{Y}_i - \overline{Y}_M),$$

where $\overline{B}$ is the average cluster size. The equation for the bias indicates that the estimator $\hat{t}_m$ is unbiased if the cluster sizes B_i do not correlate with the cluster means $\overline{Y}_i$, which is the case when the cluster sizes are equal. Therefore, if $\hat{t}_m$ is intended to be used, the relation of the cluster sizes and cluster means should be examined carefully.

Under equal-sized clusters, the design variance of $\hat{t}_m$ can also be written as

$$V_{clu-1}(\hat{t}_m) = (M \times B)^2(1 - m/M)S_b^2/m, \tag{3.12}$$

where the between-cluster variance S_b^2 can be derived from the cluster means $\overline{Y}_i$ and their mean $\overline{Y}_M$ by

$$S_b^2 = \sum_{i=1}^{M}(\overline{Y}_i - \overline{Y}_M)^2/(M-1).$$

Because of equality in cluster sizes, $\hat{t}$ and $\overline{Y}$ can be used in place of $\hat{t}_m$ and $\overline{Y}_M$ in (3.12) and in S_b^2.

We shall next study the efficiency of one-stage cluster sampling by inspecting the DEFF of a total estimator under one-stage CLU design in the simple case in which the clusters are assumed to be equal-sized.

Example 3.5

Efficiency of one-stage cluster sampling from the *Province'91* population. We consider the efficiency of one-stage cluster sampling in the estimation of the total number T of unemployed persons (UE91) by calculating the DEFF of an estimator of T. Clusters are formed by combining groups of four neighbouring municipalities into eight clusters. The $N = 32$ municipalities of the province are divided into $M = 8$ equal-sized clusters so that $B_i = B = 4$. It should be noticed that in real surveys the cluster sizes are usually unequal and, moreover, the number of population clusters is noticeably larger than here; therefore, the calculations should be taken hypothetically with the aim of illustrating the principles of the estimation. Table 3.6 presents the cluster means $\overline{Y}_i$ and totals $\overline{T}_i$ of UE91 in all the population clusters.

The sum of cluster totals T_i is equal to the population total $T = 15\,098$. The population mean per element and the mean of the cluster means $\overline{Y}_M$ are both 472 because of the equality of the cluster sizes. Let the sample size be $m = 2$ clusters,

Table 3.6 Cluster means and totals in the *Province'91* population, where each regional cluster includes four neighbouring municipalities.

Cluster identifiers			Mean and total of UE91 for population clusters	
STR	CLU	Elements (municipalities included)	Mean $\overline{Y}_i$	Total T_i
1	1	Jyväskylä, Korpilahti, Muurame, Säynätsalo	1206	4824
1	2	Jämsä, Jämsänkoski, Keuruu, Kuhmoinen	535	2141
1	3	Saarijärvi, Konginkangas, Äänekoski, Sumiainen	427	1709
1	4	Kannonkoski, Karstula, Kyyjärvi, Pylkönmäki	172	686
1	5	Suolahti, Hankasalmi, Konnevesi, Laukaa	481	1923
1	6	Joutsa, Leivonmäki, Luhanka, Toivakka	109	436
1	7	Jyväskylä mlk., Multia, Petäjävesi, Uurainen	556	2223
1	8	Kinnula, Kivijärvi, Pihtipudas, Viitasaari	289	1156

Sum of cluster totals $T = 15\,098$
Mean per cluster $\overline{T}_M = 1887$
Mean per element $\overline{Y} = 472$
Mean of cluster means $\overline{Y}_M = 472$

then the element sample size is $n = m \times B = 2{\times}4 = 8$. Because the cluster sizes are equal, the total estimators $\hat{t}$, $\hat{t}_{rat}$ and $\hat{t}_m$ would provide the same estimates, and any of the corresponding design variances could be used. To evaluate the efficiency, we calculate the design variance of $\hat{t}$ using equation (3.16).

First, the between-cluster variance is obtained as

$$S_b^2 = \frac{1}{(8-1)} \sum_{i=1}^{8} (\overline{Y}_i - 472)^2 = 340^2,$$

giving the design variance

$$V_{clu-1}(\hat{t}) = (8{\times}4)^2(1 - 2/8)S_b^2/2 = 32^2{\times}3/4{\times}340^2/2 = 6663^2.$$

The between-cluster variance $S_b^2 = 340^2$ will also be used in two-stage cluster sampling. Hence, the design effect of the total estimator $\hat{t}$ is

$$\mathrm{DEFF}_{clu-1}(\hat{t}) = \frac{V_{clu-1}(\hat{t})}{V_{srs}(\hat{t})} = \frac{6663^2}{7283^2} = 0.84.$$

The one-stage cluster sampling design appears to be slightly more efficient than the SRSWOR design in this case. However, under complex surveys, due to positive intra-cluster correlation, cluster sampling usually tends to be less efficient than SRSWOR when measured by the estimated design effects, as shown

in later chapters. The unexpected result here can be partly explained by the method of forming the clusters on an administrative basis, which produces relatively internally heterogeneous clusters with respect to the variation of UE91. If the clusters were formed by some other criteria, for example, on a travel-to-work area basis, different results might be obtained because unemployment may be more homogeneous in such areas than in the regionally neighbouring municipalities.

In the next example in which a one-stage cluster sample is drawn from the *Province'91* population, it appears that, based on the estimated variances, the efficiency can be worse than that of SRSWOR. This result, however, is crucially dependent on the composition of the sample in this case because only two clusters will be drawn from the small and heterogeneous population of clusters.

Example 3.6

Analysing a one-stage CLU sample drawn from the *Province'91* population. The *Province'91* population is divided on a regional basis into eight ($M = 8$) clusters, each comprising four ($B = 4$) neighbouring municipalities. Eight municipalities are required in the sample, so the element sample size is $n = 8$. Because the clusters are equal-sized, the cluster-level sample size is $m = 2$. The sample of clusters is drawn by simple random sampling without replacement. As a result, the clusters 2 and 8 were drawn and we obtained the sample of eight municipalities from the population of clusters as shown in Table 3.7.

The sample identifiers required for the analysis of the data set are the following three variables: STR is the stratum identifier, which in this case is a constant because the population of clusters is not stratified, i.e. there is only one stratum. The

Table 3.7 A one-stage CLU sample of two clusters from the *Province'91* population (sample clusters are in bold).

Cluster identifiers			Mean and total of UE91 for sampled clusters	
STR	CLU	Elements (municipalities included)	Mean $\overline{Y}_i$	Total T_i
1	1	Jyväskylä, Korpilahti, Muurame, Säynätsalo	···	···
1	**2**	**Jämsä, Jämsänkoski, Keuruu, Kuhmoinen**	**535.25**	**2141**
1	3	Saarijärvi, Konginkangas, Äänekoski, Sumiainen	···	···
1	4	Kannonkoski, Karstula, Kyyjärvi, Pylkönmäki	···	···
1	5	Suolahti, Hankasalmi, Konnevesi, Laukaa	···	···
1	6	Joutsa, Leivonmäki, Luhanka, Toivakka	···	···
1	7	Jyväskylä mlk., Multia, Petäjävesi, Uurainen	···	···
1	**8**	**Kinnula, Kivijärvi, Pihtipudas, Viitasaari**	**289.00**	**1156**

Sampling rate (clusters) $m/M = 2/8 = 0.25$.
···Nonsampled cluster.

Table 3.8 Estimates from a one-stage CLU sample ($n = 8$); the *Province'91* population.

Statistic	Variables	Parameter	Estimate	s.e	c.v	deff
Total	UE91	15 098	13 188	3412	0.26	1.92
Ratio (%)	UE91, LAB91	12.65%	12.93%	0.6%	0.04	1.44
Median	UE91	229	337	132	0.39	1.29

cluster identification (2 or 8) is given by the variable CLU; and the weight variable is a constant WGHT = 4, i.e. the cluster size. The finite-population correction at the cluster level is $(1 - 0.25) = 0.75$, and so the sampling rate is 0.25.

Estimation results for the total $\hat{t}$, ratio $\hat{r}$ and median $\hat{m}$, and the values of the corresponding parameters T, R and M are displayed in Table 3.8. From there it can be seen that one-stage cluster sampling appears to be inefficient for all three estimators. The deff estimates are noticeably greater than one ($1.29 \leq \text{deff} \leq 1.92$). Moreover, for this actual sample, the estimated deff $(\hat{t}) = 1.92$ differs noticeably from the corresponding parameter DEFF $(\hat{t}) = 0.84$. This is due to the small number of sample clusters, which causes instability in the estimated design variances. The variance estimates depend heavily on which clusters happen to be drawn; thus, by selecting two clusters other than those just drawn, deff estimates noticeably less than one could be obtained. The problem of instability will be discussed in more detail in Chapter 5.

Two-stage Cluster Sampling

Subsampling from the sampled clusters is common when working with large clusters. This offers better possibilities, for instance, for the control of the element-level sample size n, when the cluster sizes vary. Moreover, with subsampling, the number of sample clusters can be increased when compared to one-stage cluster sampling for a fixed-element sample size, which can increase efficiency. A practical motivation is the availability of sampling frames that are only required for subsampling from the sampled clusters.

In *two-stage cluster sampling*, a sample of clusters is drawn from the population of clusters, i.e. primary sampling units (PSUs) at the first stage of sampling, using the standard element-sampling techniques such as SRSWOR, SYS or PPS. Moreover, the population of clusters can be stratified by using available auxiliary information. The simplest stratified two-stage cluster-sampling design in which exactly two clusters are drawn from each stratum is often used in practice, offering the possibility of using a large number of strata and thereby increasing efficiency. At the second stage, an element-level sample is drawn from the sampled clusters again using standard element-sampling techniques. In practice, the cluster sizes in the population, and the cluster sample sizes, usually vary. Moreover, the inclusion probabilities can vary at each stage of sampling. But a sample with

a constant overall sampling fraction can be obtained by an appropriate choice of the sampling fractions and selection techniques at each stage of sampling. This kind of multi-stage design is called an *epsem design* (equal probability of selection method).

In one-stage cluster sampling, all the elements in the sampled clusters make up the element-level sample and, thus, the only variation due to sampling is between-cluster variation. But in two-stage cluster sampling, an additional source of variation arises due to subsampling, namely, the variation within the clusters, and this also contributes to the total variation.

For illustrating the basics of two-stage cluster sampling, we assume SRS without replacement at both stages of sampling and equality of the cluster sizes in the M population clusters, i.e. $B_i = B$ for all i. The element-level population size is thus $N = M \times B$. Moreover, let us further assume that the element-level sample sizes are also equal for simplicity, i.e. $b_i = b$ in all the m sample clusters; the sample size is thus $n = m \times b$. Cluster sampling under these assumptions results in equal inclusion probabilities for the population clusters, and they are also equal for the population elements, which provides an epsem sample. We can see by writing the sampling fractions m/M for the first stage and b/B for the second stage, giving a constant overall sampling fraction $(m/M) \times (b/B) = n/N$.

The main interest in the estimation is usually concentrated on the second stage, i.e. element-level parameters. Let us consider the estimation of the population total $T = \sum_{i=1}^{M} T_i$, where $T_i = B \times \overline{Y}_i$ is the population total in cluster i and $\overline{Y}_i = \sum_{k=1}^{B} Y_{ik}/B$ is the mean per element in cluster i as previously. An unbiased estimator of the total T is

$$\hat{t} = (M \times B) \sum_{i=1}^{m} \overline{y}_i / m, \tag{3.13}$$

where $\overline{y}_i = \sum_{k=1}^{b} y_{ik}/b$ is the mean per element in sample cluster i. In the derivation of the design variance for $\hat{t}$, a decomposition of the total variance into the between-cluster variance and within-cluster variance components can be used. The design variance for the estimator $\hat{t}$ is the weighted sum of the *between-cluster* variance S_b^2 and *within-cluster* variance S_w^2:

$$V_{clu-II}(\hat{t}) = (M \times B)^2 \left[\left(1 - \frac{m}{M}\right) \frac{S_b^2}{m} + \left(1 - \frac{b}{B}\right) \frac{S_w^2}{mb} \right], \tag{3.14}$$

with

$$S_b^2 = \frac{1}{(M-1)} \sum_{i=1}^{M} (\overline{Y}_i - \overline{Y})^2,$$

$$S_w^2 = \frac{1}{M(B-1)} \sum_{i=1}^{M} \sum_{k=1}^{B} (Y_{ik} - \overline{Y}_i)^2,$$

and $\overline{Y} = T/(M \times B)$ is the overall population mean per element. The between-cluster variance term is due to the first-stage sampling of the clusters and is similar to one-stage cluster sampling and the additional within-cluster variation is due to the subsampling. In one-stage cluster sampling, the within-cluster variance component is zero because all the B elements were taken from the sampled clusters, i.e. $b = B$.

Estimators of the variance terms S_b^2 and S_w^2 are obtained by inserting the sample counterparts in place of the population values. We hence obtain

$$\hat{s}_b^2 = \frac{1}{(m-1)} \sum_{i=1}^{m} (\overline{y}_i - \overline{y})^2,$$

$$\hat{s}_w^2 = \frac{1}{m(b-1)} \sum_{i=1}^{m} \sum_{k=1}^{b} (y_{ik} - \overline{y}_i)^2,$$

where $\overline{y} = \sum_{i=1}^{m} \overline{y}_i/m$ is the sample mean per element. The estimator of the design variance $\hat{t}$ is then given by

$$\hat{v}_{clu-II}(\hat{t}) = (M \times B)^2 \left[\left(1 - \frac{m}{M}\right) \frac{\hat{s}_b^2}{m} + \left(1 - \frac{b}{B}\right) \frac{m}{M} \frac{\hat{s}_w^2}{mb} \right]. \tag{3.15}$$

From (3.15), it can be inferred that if the first-stage sampling fraction m/M is small, then the second component in the variance estimator becomes negligible. Then, a variance estimator based on only the between-cluster variation can be used as a slightly negatively biased approximation of the design variance of $\hat{t}$, which has the convenient property that it is only computed from cluster-level quantities. Further, if m/M is small, the first-stage finite-population correction would be close to one and thus can be omitted, leading to a with-replacement-type variance estimator. This kind of variance approximation will be extensively used when discussing survey analysis in later chapters. Alternatively, if the fraction m/M is not negligible, the within-variance component can contribute substantially to the variance estimate.

In practice, the cluster sizes B_i and the sample sizes from the sampled clusters b_i usually vary, and moreover, the population of clusters can be stratified. Appropriate estimators for the total and the design variance of the total estimator should be used to properly account for the stratification and the variation in the cluster sample sizes. For the total, a ratio-type estimator or an estimator based on PPS sampling of the clusters can be used with the cluster sizes as the auxiliary size measure. The estimation of the design variance of a ratio-type estimator under two-stage stratified cluster sampling will be discussed in Chapter 5. There, various approximate variance estimators are introduced.

The inconvenient effect of variation in cluster sizes can be controlled using PPS sampling of the clusters. Let us suppose that an epsem sample is desired with a

fixed size of n elements. This can be attained by drawing a constant number $b_i = b$ of elements from each of the m unequally sized sample clusters when the clusters are selected with PPS with inclusion probabilities proportional to the cluster sizes B_i, as can be inferred from the following formula:

$$\frac{n}{N} = \frac{m \times B_i}{\sum_{i=1}^{M} B_i} \times \frac{b}{B_i},$$

where m is the desired number of sample clusters and $b = n/N \times \sum_{i=1}^{m} B_i/m$.

In the next example, we evaluate the efficiency of two-stage CLU design in the simple situation of equal-sized clusters, based on the calculation of the DEFF. Comparison is made with the one-stage cluster design.

Example 3.7

Efficiency of two-stage cluster sampling from the *Province'91* population. The number of clusters consisting of neighbouring municipalities is 8, so that $M = 8$, and each cluster comprises $B = 4$ municipalities. We compare the efficiency of one- and two-stage CLU designs in the estimation of the total T. Both designs involve equal clustering at the first stage. In one-stage cluster sampling, two clusters ($m = 2$) were drawn, and all the four municipalities from the sampled clusters were taken into the element-level sample. The sample size was thus $n = m \times B = 2{\times}4 = 8$ municipalities. In two-stage cluster sampling, we take $m = 4$ clusters in the first-stage sample, and we draw $b = 2$ municipalities from each sampled cluster at the second stage. The element-level sample size is then also $m \times b = 4{\times}2 = 8$ municipalities.

Under the one-stage CLU design, the design variance was calculated as $V_{clu}(\hat{t}) = 6663^2$, and the design effect was $\text{DEFF}(\hat{t}) = 0.84$. Under the two-stage CLU design, we must first calculate the between-cluster and within-cluster variance components. The between-cluster variance in Example 3.3 was calculated as $S_b^2 = 340^2$. The within-cluster variance is

$$S_w^2 = \frac{1}{8(4-1)} \sum_{i=1}^{8} \sum_{k=1}^{4} (Y_{ik} - \overline{Y}_i)^2 = 660^2.$$

The design variance of $\hat{t}$ is thus

$$V_{clu-II}(\hat{t}) = (8{\times}4)^2 \left[\left(1 - \frac{4}{8}\right) \frac{340^2}{4} + \left(1 - \frac{2}{4}\right) \frac{660^2}{4{\times}2} \right] = 6532^2$$

and the DEFF of $\hat{t}$ for the two-stage design is

$$\text{DEFF}_{clu-II}(\hat{t}) = 6532^2/7283^2 = 0.80.$$

When compared to the one-stage CLU design, the two-stage design is slightly more efficient. This is in part due to the property of the two-stage design that, with a given n, more first-stage units can be drawn than for the one-stage design. In this case, the number of sample clusters is doubled, which decreases the first-stage variance component. Of the total variance, 35% is contributed by the first stage (between-cluster) and 65% by the second stage (within-cluster). Thus, the within-cluster contribution dominates, which is in part due to the relative heterogeneity of the clusters. It should, however, be noticed that in the *Province'91* population, the population of clusters is small and so are the cluster sample size and the sample size in subsampling. Therefore, these calculations should be taken as a hypothetical example, because in a real survey the corresponding figures are larger, the clusters tend to be relatively homogeneous and a major share of the design variance is often due to between-cluster variation.

The next example demonstrates computational results based on a sample drawn from the *Province'91* population using the two-stage CLU design. The efficiency is studied on the basis of estimated design variances. The efficiency is also compared with that of the one-stage CLU sample from Example 3.6.

Example 3.8

Analysing a two-stage CLU sample drawn from the *Province'91* population. In the first stage, the clusters numbered 2, 3, 4 and 7 were drawn. In the second stage, two municipalities were drawn from each sample cluster. The population of clusters and the two-stage CLU sample is displayed in Table 3.9.

Table 3.9 A two-stage cluster sample from the *Province'91* population. First stage: SRSWOR sample of four clusters (2, 3, 4 and 7). Second stage: four SRSWOR samples of two elements in each sampled cluster. (Sampled elements in sampled cluster are in bold).

Cluster identifiers			Estimated mean and total of UE91 for sampled clusters	
STR	CLU	Elements (municipalities included)	Mean $\overline{y}_i$	Total $\hat{t}_i$
1	1	Jyväskylä, Korpilahti, Muurame, Säynätsalo	···	···
1	2	Jämsä, Jämsänkoski, **Keuruu, Kuhmoinen**	473.5	1894
1	3	Saarijärvi, **Konginkangas, Äänekoski**, Sumiainen	454.5	1818
1	4	Kannonkoski, Karstula, **Kyyjärvi, Pylkönmäki**	96.0	384
1	5	Suolahti, Hankasalmi, Konnevesi, Laukaa	···	···
1	6	Joutsa, Leivonmäki, Luhanka, Toivakka	···	···
1	7	Jyväskylä mlk., Multia. **Petäjävesi, Uurainen**	241.0	962
1	8	Kinnula, Kivijärvi, Pihtipudas, Viitasaari	···	···

Sampling rates: First stage $4/8 = 0.50$. Second stage $2/4 = 0.50$.
···Nonsampled cluster.

In the analysis of the data from the two-stage CLU design, the following design identifiers are required: the stratum identifier STR, which is a constant 1 for all the sample elements, the cluster identifier CLU, which has the values 2, 3, 4 and 7 corresponding to the sampled clusters, and the weight variable WGHT, which is a constant (4) for all the sample elements. It should be noted that the weight would vary between the clusters if the cluster sizes varied and the selection rates in the clusters were not equal. Because SRSWOR was used at both stages, the first-stage sampling rate of 4/8 and the second-stage sampling rate of 2/4 are also supplied, giving the weights $w_{ik} = (M \times B)/(m \times b) = (8{\times}4)/(4{\times}2) = 4$ for all the sample elements. Estimation results on the total number of unemployed $\hat{t}$, the unemployment rate $\hat{r}$ and the median unemployment $\hat{m}$, as well as the values of the corresponding parameters *T*, *R* and *M* are displayed in Table 3.10.

The estimated design effects (deff) for the total, ratio and median estimators are close to one, indicating that the two-stage CLU sample does not differ greatly from SRSWOR in efficiency. But the efficiency differs considerably from that of the one-stage counterpart where design-effect estimates noticeably larger than one were obtained for all the estimators. In the one-stage design, the number of sample clusters was very small, thus resulting in serious instability in the variance estimates. In the two-stage design, on the other hand, one half of all the population clusters were drawn and, therefore, the design is not as sensitive to instability and, in addition, the population clusters were relatively heterogeneous. It should be noticed, however, that in this example the clustering was an illustration of the estimation under two-stage cluster sampling, not an example of cluster sampling in real surveys. These will be considered in later chapters.

Intra-cluster Correlation and Efficiency

Efficiency of cluster sampling depends strongly on the internal composition of the clusters. Cluster sampling would be as efficient as simple random sampling if the clusters were internally heterogeneous so that each of them closely mirrored the overall composition of the element population. Efficiency decreases if the clusters are internally homogeneous and if the between-cluster variation is large. In practice, many naturally formed population subgroups are of this latter type.

Table 3.10 Estimates from a two-stage CLU sample ($n = 8$); the *Province'91* population.

Statistic	Variables	Parameter	Estimate	s.e	c.v	deff
Total	UE91	15 098	10 116	2659	0.26	0.93
Ratio (%)	UE91, LAB91	12.65%	13.81%	0.5%	0.04	0.99
Median	UE91	229	192	49	0.25	0.84

The efficiency can be studied by *intra-cluster correlation*, which is a measure of the internal homogeneity of the clusters. This correlation can be included in the design variance equations of estimators from cluster sampling. Recall that in systematic sampling, a similar coefficient (intra-class correlation) also played a crucial role; SYS can indeed be taken as a special case of one-stage cluster sampling where only one cluster is drawn.

Let us assume equal-sized clusters $B_i = B$ in all the population clusters. We first study the ANOVA decomposition $SST = SSW + SSB$ of the total variation SST of the study variable y into the variation within the clusters (SSW) and between the clusters (SSB). The total variation SST can be written

$$\sum_{i=1}^{M}\sum_{k=1}^{B}(Y_{ik}-\overline{Y})^2 = \sum_{i=1}^{M}\sum_{k=1}^{B}(Y_{ik}-\overline{Y}_i)^2 + \sum_{i=1}^{M}B(\overline{Y}_i-\overline{Y})^2, \tag{3.16}$$

where Y_{ik} is the population value of the study variable for an element ik from cluster i, $\overline{Y}$ is the overall mean per element and $\overline{Y}_i$ is the cluster mean per element as previously given.

By using the formulae for intra-class correlation ρ_{int} derived in Section 2.4 under SYS for cluster sampling, we get

$$\rho_{int} = 1 - \frac{B}{B-1} \times \frac{SSW}{SST}. \tag{3.17}$$

The interpretation of intra-cluster correlation depends on the share of the total variation between the two variance components. First, if all the variation is within the clusters and there is no between-cluster variation, then the intra-cluster correlation coefficient is at minimum $\rho_{int} = -1/(B-1)$. If, on the other hand, all the variation is between the clusters, in which case the clusters are internally completely homogeneous, the coefficient has its maximum $\rho_{int} = 1$. And with the value $\rho_{int} = 0$ the elements are assigned to clusters at random.

Let us consider the efficiency of one-stage CLU sampling with respect to SRSWOR of the same size n. The design variance of an estimator $\hat{t}$ of the total T was in equation (3.12) under the CLU design given as

$$V_{clu-1}(\hat{t}) = (M \times B)^2 \left(1 - \frac{m}{M}\right)\frac{S_b^2}{m},$$

where S_b^2 is the between-cluster variance component. From equations (3.16) and (3.17) it follows that the between-cluster variance component can be written as

$$SSB = \frac{SST}{B}[1 + (B-1)\rho_{int}].$$

Inserting it into the variance formula above, we obtain

$$V_{clu-1}(\hat{t}) = (M \times B)^2 \left(1 - \frac{m}{M}\right) \frac{S^2}{m} \left[\frac{1}{B}(B-1)\rho_{int}\right] \times \frac{N-1}{N} \times \frac{M}{M-1}.$$

Assuming large N and M, the last two terms become close to one and can thus be dropped. We hence obtain, for the design variance of $\hat{t}$, an expression based on the total variance S^2 and the intra-cluster correlation ρ_{int}:

$$V_{clu-1}(\hat{t}) \doteq (M \times B)^2 \left(1 - \frac{m}{M}\right) \frac{S^2}{n}[1 + (B-1)\rho_{int}], \tag{3.18}$$

because $m \times B = n$. But the corresponding SRS design variance of $\hat{t}$ can be written as

$$V_{srs}(\hat{t}) = (M \times B)^2 \left(1 - \frac{n}{N}\right) \frac{S^2}{n},$$

which leads to the DEFF of $\hat{t}$ given by

$$\text{DEFF}_{clu-1}(\hat{t}) = \frac{V_{clu-1}(\hat{t})}{V_{srs}(\hat{t})} = 1 + (B-1)\rho_{int}, \tag{3.19}$$

because $m/M = n/N$ in the finite-population correction term of $V_{clu-1}(\hat{t})$.

The equation of DEFF indicates that when ρ_{int} is positive, which is usually the case in practice, then cluster sampling is less efficient than simple random sampling. And for a given ρ_{int}, the DEFF increases with increasing cluster size B. In the final example in this section, efficiency is further discussed as a function of cluster size and intra-cluster correlation.

Example 3.9

Intra-cluster correlation, cluster size and DEFF in the *Province'91* population. One-way of analysis of variance is calculated for the variable UE91 using the eight regional clusters as factor levels. Results are presented in Table 3.11.

Table 3.11 Population ANOVA table; one-stage cluster sampling with $M = 8$ and $B = 4$ from the *Province'91* population.

Source of variation	df	Sum of squares	Mean square
Between clusters	7	$\text{SSB} = 32.30 \times 10^5$	$\text{MSB} = 4.61 \times 10^5$
Within clusters	24	$\text{SSW} = 139.02 \times 10^5$	$\text{MSW} = 5.79 \times 10^5$
Total	31	$\text{SST} = 171.32 \times 10^5$	$S^2 = 5.53 \times 10^5 = 743^2$

Inserting the figures in equation (3.17), we get

$$\rho_{int} = 1 - \frac{4}{4-1} \times \frac{139.02 \times 10^5}{171.32 \times 10^5} = -0.082.$$

Design effect can be approximated from equation (3.19) (equal-sized clusters $B = 4$ are assumed) as $\text{DEFF}_{clu-1} = 1 + (B-1)\rho_{int} = 1 + (4-1)(-0.082) = 0.754$.

This figure is smaller than the exact DEFF = 0.84 computed in Example 3.4, because the formula (3.17) is an approximation more applicable for large populations with small sampling fraction.

The intra-class, or intra-cluster, correlation appeared to be an important design parameter in systematic sampling and in cluster sampling. The intra-class correlation measures the correlation between pairs of elements belonging to the same subgroup of population. In SYS, this subgroup was the elements in a sampling interval. In cluster sampling, intra-cluster correlation indicates dependency of elements belonging to the same cluster or a natural subgroup of population elements. We will consider a number of such grouping structures: pupils within a school, employees within a business firm and household members within a household. Several options are available when measuring the internal homogeneity of such clusters with an intra-cluster correlation coefficient. In systematic sampling and cluster sampling, the intra-cluster correlation coefficient was calculated in a design-based manner. In multivariate modelling, other options become more relevant. This includes the 'working' intra-cluster correlation coefficient to be introduced in Chapter 8 in the context of multivariate survey analysis estimation. In Section 9.4, intra-class correlation coefficients will be calculated in a model-based manner. This holds also for another way of forming clusters, namely, the workloads of interviewers (see Section 9.1).

Summary

Cluster sampling is commonly used in practice because many populations are readily clustered into natural subgroups. Typical clusters met in real surveys are regional administrative units, city blocks or block-like units, households, business firms or establishments and schools or school classes. Often, for practical and economical reasons, these kinds of clusters are used in sampling and in data-collection procedures. A practical motivation is that sampling frames for subsampling are needed only for the sampled clusters. And an economical motivation is that the cost efficiency of cluster sampling can be fairly high. Good examples of various cluster-sampling designs are to be found later in this book. A drawback in cluster sampling, however, is that due to the relative homogeneity of the clusters, as is often the case in practice, the statistical efficiency can be less than that of simple random sampling. However, high cost efficiency can successfully compensate for this inconvenience.

Our demonstration data, the *Province'91* population, appeared restrictive for a thorough demonstration of cluster sampling and was thus used for illustrating the basic principles of sampling and estimation in one-stage and two-stage designs. In large-scale surveys, there are usually a large number of clusters both in the population and in the sample. Moreover, the population of clusters can be stratified, and sampling can be achieved using several stages. In the analysis of such data, ratio-type estimators with approximative variance estimators are usually used in the estimation. These topics will be considered in detail in Chapter 5.

Cluster sampling is discussed in most textbooks on survey sampling. As further reading, Kish (1965), Lohr (1999), Levy and Lemeshow (1991) and Snijders and Bosker (2002) can be recommended, covering introductory, advanced and more theoretical topics on cluster sampling.

3.3 MODEL-ASSISTED ESTIMATION

Introduction

In the techniques discussed so far, auxiliary information of the population elements is used in the sampling phase to attain an efficient sampling design. We now turn to a different way of utilizing auxiliary information. Our aim is to introduce estimators that can be used for the selected sample to obtain better estimates of the parameters of interest, relative to the estimates calculated with estimators based on the sampling design used.

Let us assume that appropriate auxiliary data are available from the population as a set of auxiliary variables. Of these variables, some might be categorical and some continuous. Some auxiliary data are perhaps used for the sampling procedure. Others can be used for improving efficiency; a way to do this is, for example, to use an auxiliary variable z, which is related to our study variable y, for a reduction of the design variance of the original estimator of the population total of y. In Särndal *et al.* (1992), these techniques are discussed in the context of *model-assisted design-based estimation. Model-assisted* estimation refers to the property of the estimators that models such as linear regression are used in incorporating the auxiliary information in the estimation procedure for the finite-population parameters of interest, such as totals. Model-assisted estimation should be distinguished from the multivariate survey analysis methods to be discussed in Chapter 8. There, models are also used but for multivariate survey analysis purposes.

In the following text, a brief review is given on model-assisted estimation. More specifically, *poststratification, ratio estimation* and *regression estimation* are considered. The methods are special cases of so-called *generalized regression estimators*. All these methods are aimed at improving the estimation from a given sample by using available auxiliary information from the population. This can result in estimates closer to the true population value and a reduction in the design variance of an estimator calculated from the sampled data.

In model-assisted estimation, an auxiliary variable z, which is related to the study variable y, is required. If this variable is categorical, the target population U can be partitioned into subpopulations $U_1, \ldots, U_g, \ldots, U_G$ according to some classification principle. In poststratification, these subpopulations are called *poststrata*. If the poststrata are internally homogeneous, this partitioning can capture a great deal of the total variance of the study variable y, resulting in a decrease in the design-based variance of an estimator. Moreover, poststratification can be used to obtain more accurate point estimates and reduce the bias of sample estimates caused by nonresponse.

The auxiliary variable z is often continuous. If it correlates strongly with the study variable y, a linear regression model can be assumed with y as the dependent variable and z as the predictor. This regression can be estimated from the observed sample and used in the estimation of the original target parameter. For this, ratio estimation and regression estimation can be used. By these methods, substantial gains in efficiency and increased accuracy are often achieved.

To construct a model-assisted estimator, two kinds of weights are considered. The preliminary weights are the usual sampling weights w_k, which generally are the inverses of the inclusion probabilities π_k; these weights are extensively used in this book. The other type of weights are called *g weights* and their values g_k depend both on the selected sample and on the chosen estimator. The product $w_k^* = g_k w_k$ gives new weights known as *calibrated* weights, which are used in the model-assisted estimators. Thus, using calibrated weights, a model-assisted estimator can be written as $\hat{t}_{cal} = \sum_{k=1}^{n} w_k^* y_k$. A property of the calibrated weights is that for example for ratio estimation, the estimator $\hat{t}_{z,cal} = \sum_{k=1}^{n} w_k^* z_k$ of the total of the auxiliary z-variable reproduces exactly the known population total T_z. The g weights and calibrated weights will be explicitly given for poststratification, ratio estimation and regression estimation.

The basic principles of model-assisted estimation are most conveniently introduced for SRSWOR, although natural applications in practical situations are often under more complex designs. A further simplification is that only one auxiliary variable is assumed. Also, this assumption can be relaxed if multiple auxiliary variables are available as is assumed in discussing regression estimation. The concept of *estimation strategy* will be used referring to a combination of the sampling design and the appropriate estimator. The model-assisted strategies to be discussed are shown in Table 3.12. In the design-based reference strategies, no auxiliary information is used.

Poststratification

Poststratification can be used for improvement of efficiency of an estimator if a discrete auxiliary variable is available. This variable is used to stratify the sample data set after the sample has been selected. Recall from Section 3.1 that

Table 3.12 Estimation strategies for population total.

Strategy		Auxiliary information	Assisting model
Design-based strategies			
SRSWOR		Not used	None
SRSWR		Not used	None
Model-assisted strategies			
Poststratification	SRS*pos	Discrete	ANOVA
Ratio estimation	SRS*rat	Continuous	Regression (no intercept)
Regression estimation	SRS*reg	Continuous	Regression

stratification of the element population as part of the sampling design often gave a gain in efficiency. This was achieved by an appropriate choice of the stratification variables so that the variation in the study variable y within the strata would be small. Poststratification has a similar aim. To avoid confusion with the usual (pre)stratification, the population is partitioned into G groups that are called *poststrata*.

To carry out poststratification, the sample data are first combined with the appropriate auxiliary data obtained perhaps from administrative registers or official statistics. Combining the sampled data with poststratum information and the corresponding selection probabilities, we can proceed with the estimation in basically the same way as if it were being done by ordinary (pre)stratification. Certain differences exist, however. Because we are stratifying after the sample selection or, more usually, after the data collection, we cannot assume any specific allocation scheme. The sample size n is fixed but how it is allocated to the different strata is not known until the sample is drawn. This property causes no harm to the estimation of, for example, the total, but estimating of the variance of the total estimator requires more attention.

The *poststratified estimator* for the total T of y is given by

$$\hat{t}_{pos} = \sum_{g=1}^{G} \hat{t}_g = \sum_{g=1}^{G} \sum_{k=1}^{n_g} w^*_{gk} y_{gk}, \tag{3.20}$$

where $\hat{t}_g = N_g \overline{y}_g$ is an estimator of the poststratum total T_g and N_g is the size of the poststratum g. The poststratum weights are $w^*_{gk} = g_{gk} w_{gk}$, where the g weights are $g_{gk} = N_g / \hat{N}_g$ with the *estimated poststratum sizes* in the denominator, and w_{gk} are the original sampling weights. The calculation of w^*_{gk} will be illustrated in Example 3.9. The variance of $\hat{t}_{pos}$ can be determined in various ways, depending on how one uses the configuration of the observed sample. The configuration

refers to how the actual poststratum sample sizes n_g are distributed, and if this is taken as given, the *conditional variance* is simply the same as the usual variance for stratified samples:

$$V_{srs,con}(\hat{t}_{pos}|n_1,\ldots,n_g,\ldots,n_G) = \sum_{g=1}^{G} N_g^2 \left(1-\frac{n_g}{N_g}\right)\frac{S_g^2}{n_g}, \tag{3.21}$$

where the poststratum variances are given by $S_g^2 = \sum_{k=1}^{N_g}(Y_{gk}-\overline{Y}_g)^2/(N_g-1)$. By averaging (3.21) over all possible configurations of n, the *unconditional variance* is obtained. This gives an alternative variance formula,

$$V_{srs,unc}(\hat{t}_{pos}) = \sum_{g=1}^{G} N_g^2 \left(1-\frac{E(n_g)}{N_g}\right)\frac{S_g^2}{E(n_g)}, \tag{3.22}$$

where $E(n_g)$ is the expected poststratum sample size. This variance can be approximated in various ways. One of the approximations is

$$V_{srs,unc}(\hat{t}_{pos}) \doteq N^2\left(1-\frac{n}{N}\right)\left(\frac{1}{n}\right)\left[\sum_{g=1}^{G}\left(\frac{N_g}{N}\right)S_g^2 + \left(\frac{1}{n}\right)\sum_{g=1}^{G}\left(1-\frac{N_g}{N}\right)S_g^2\right]. \tag{3.23}$$

The difference between the conditional and unconditional variances could be considerable if the sample size is small. The corresponding variance estimators $\hat{v}_{srs,con}(\hat{t}_{pos})$ and $\hat{v}_{srs,unc}(\hat{t}_{pos})$ are obtained by inserting $\hat{s}_g^2$ for S_g^2, where $\hat{s}_g^2 = \sum_{k=1}^{n_g}(y_{gk}-\overline{y}_g)^2/(n_g-1)$. For illustrative purposes, both variances $V_{srs,con}$ and $V_{srs,unc}$ are estimated in the next example.

Example 3.10

Estimation with poststratification. The sample used is drawn with SRSWOR from the *Province'91* population in Section 2.3 (see Example 2.1). The sample is poststratified according to administrative division of the municipalities into urban and rural municipalities. The target population contains $N_1 = 7$ urban and $N_2 = 25$ rural municipalities. The two poststrata have the value 1 for urban and 2 for rural municipalities.

In Table 3.13, the sample information used for the estimation with poststratification is displayed.

Let us consider more closely the estimation of the total T. The poststratum totals of UE91 estimated from the table are $\hat{t}_1 = N_1\overline{y}_1 = 7\times 1868 = 13\,076$ and $\hat{t}_2 = N_2\overline{y}_2 = 25\times 201.2 = 5030$. Using these estimates, the poststratified estimate for T is $\hat{t}_{pos} = \hat{t}_1 + \hat{t}_2 = 18\,106$.

Alternatively, the total estimate $\hat{t}_{pos}$ can be calculated using the poststratum weights w_k^*. To calculate w_k^*, the original sampling weights w_k should be adjusted by the sample dependent g_k weights. For this, first the estimate of the poststratum size is determined. Denoting by w_{gk} the original element weight of a sample element that belongs to the poststratum g, an estimate for poststratum size $\hat{N}_g$ is given by summing up these original weights. Then, the corresponding g weight for an element k in poststratum g is simply $g_{gk} = N_g/\hat{N}_g$, where N_g is the exact size of the poststratum g. For example, in Table 3.13, the original sampling weight under SRS is $w_k = 4$, or a constant for each population element. In the first poststratum, the poststratum size is $N_1 = 7$ and its estimated size is $\hat{N}_1 = 4 + 4 + 4 = 12$, because there are three sampled elements in the first poststratum. Thus, the corresponding g weight is $g_{1k} = N_1/\hat{N}_1 = 7/12 = 0.5833$. Finally, the poststratum weights are given for the first poststratum by $w_{1k}^* = g_{1k} \times w_{1k} = 0.5833 \times 4 = 2.3333$. This value turns out to be the same for all the sampled elements for the first poststratum (urban municipalities). Using the poststratum weights, the estimate $\hat{t}_{pos}$ will be equal to that previously calculated.

Estimation results for the estimators of total and ratio are displayed in Table 3.14. The original setting of sample identifiers remains, say STR = 1 and CLU = ID, but the element weights are to be replaced by the poststratum weights, and the sampling rate is 0.43 for the first poststratum and 0.20 for the second poststratum. Original sampling weights are used and the sampling rate is 0.25 for both poststrata for estimation of unconditional variance. Note that this procedure roughly approximates the formula given in (3.23). For comparison, the design-based estimates $\hat{t}$ and $\hat{r}$ obtained under SRSWOR are included.

Table 3.13 A simple random sample drawn without replacement from the *Province'91* population with poststratum weights.

Sample design identifiers			Element	Study variables		Poststratification		
							g	Post.
STR	CLU	WGHT	LABEL	UE91	LAB91	POSTSTR	WGHT	WGHT
1	1	4	Jyväskylä	4123	33786	1	0.5833	2.3333
1	4	4	Keuruu	760	5919	1	0.5833	2.3333
1	5	4	Saarijärvi	721	4930	1	0.5833	2.3333
1	15	4	Konginkangas	142	675	2	1.2500	5.0000
1	18	4	Kuhmoinen	187	1448	2	1.2500	5.0000
1	26	4	Pihtipudas	331	2543	2	1.2500	5.0000
1	30	4	Toivakka	127	1084	2	1.2500	5.0000
1	31	4	Uurainen	219	1330	2	1.2500	5.0000

Sampling rate for calculation of *unconditional variance*: $8/32 = 0.25$
Sampling rates for calculation of *conditional variance*:
Stratum 1 (Urban) $= 3/7 = 0.43$
Stratum 2 (Rural) $= 5/25 = 0.20$

Table 3.14 Poststratified estimates from a simple random sample drawn without replacement from the *Province'91* population.

(1) Poststratified estimates (conditional)

Statistic	Variables	Estimate	s.e	c.v	deff
Total	UE91	18 106	6014	0.33	0.33
Ratio	UE91, LAB91	12.97%	0.45%	0.03	0.59

(2) Poststratified estimates (unconditional)

Statistic	Variables	Estimate	s.e	c.v	deff
Total	UE91	18 106	7364	0.41	0.50
Ratio	UE91, LAB91	12.97%	0.49%	0.03	0.70

(3) Design-based estimates

Statistic	Variables	Estimate	s.e	c.v	deff
Total	UE91	26 440	13 282	0.50	1.00
Ratio	UE91, LAB91	12.78%	0.41%	0.03	1.00

The comparison shows how poststratification affects point estimates. The big gain is obtained when estimating the population total. The estimate of the number of unemployed is $\hat{t}_{pos} = 18\,106$, which is closer to the true value $T = 15\,098$ than the design-based estimate $\hat{t} = 26\,440$. The ratio estimate changes only slightly.

The reason for a more accurate estimate for the total is obvious. Under SRSWOR, one should have drawn urban and rural municipalities approximately by their respective proportions: $(8/32) \times (7) \approx 2$ towns and $(8/32) \times (25) \approx 6$ rural municipalities. The urban municipalities have larger populations and unemployment figures. If by chance they are over-represented in the sample, then the design-based estimator will overestimate the population total. But poststratification can correct (at least partially) skewnesses. Therefore, we could also get a point estimate closer to its true value.

Poststratification can also improve efficiency. Again, this is true especially for the total. The estimated variance of $\hat{t}_{pos}$ under the conditional assumption is reduced to one-third when compared with the pure design-based estimate $\hat{t}$, which is indicated by $\text{deff} = 0.33$. If the unconditional variance is used as a basis, then $\text{deff} = 0.50$. The unconditional variance estimate is greater than the conditional variance estimate, because the poststratum sample sizes n_g are by definition random variables whose variance contribution increases the total variance.

Ratio Estimation of Population Total

The estimation of the population total T of a study variable y was considered previously under poststratification using the sample data and a discrete auxiliary variable. *Ratio estimation* can also be used to improve the efficiency of the estimation of T, if a continuous auxiliary variable z is available. The population total T_z and the n sample values z_k of z are required for this method. Such information can often be obtained from administrative registers or official statistics. This information can be used to improve the estimation of T by first calculating the sample estimator $\hat{r} = \hat{t}/\hat{t}_z$ of the ratio $R = T/T_z$ and multiplying $\hat{r}$ by the known total T_z. Ratio estimation of the total can be very efficient if the ratio Y_k/Z_k of the values of the study and auxiliary variables is nearly constant across the population.

Ratio estimators are usually effective but slightly biased. Because of bias, the mean squared error (MSE) could be used instead of the variance when examining the sampling error. It has been shown that the proportional bias of a ratio estimator is $1/n$ and so becomes small when the sample size increases. Thus, the variance serves as an approximation to the MSE in large samples. The properties of ratio estimators have been studied widely in classical sampling theory.

Let us consider ratio estimation of the total T of y under simple random sampling without replacement. We are interested in a *ratio-estimated total* given by

$$\hat{t}_{rat} = \hat{r} \times T_z = \sum_{k=1}^{n} w_k^* y_k, \tag{3.24}$$

where $\hat{r} = \hat{t}/\hat{t}_z = N\overline{y}/N\overline{z} = \sum_{k=1}^{n} yk/\sum_{k=1}^{n} z_k$ and T_z is the population total of the auxiliary variable z. The calibrated weights are $w_k^* = g_k w_k = (T_z/\hat{t}_z)w_k$.

In the estimator (3.24), $\hat{r}$ is a random variable and the total T_z is a constant. Thus, the variance of $\hat{t}_{rat}$ can be written simply as $V_{srs}(\hat{t}_{rat}) = T_z^2 \times V_{srs}(\hat{r})$. If the SRSWOR design variance of the estimator $\hat{r}$ of a ratio (equation (2.9)) is introduced here, an approximative variance of the ratio-estimated total is given by

$$V_{srs}(\hat{t}_{rat}) \doteq N^2 \left(1 - \frac{n}{N}\right)\left(\frac{1}{n}\right)\sum_{k=1}^{N} \frac{(Y_k - R \times Z_k)^2}{N-1}, \tag{3.25}$$

whose estimator is given by

$$\hat{v}_{srs}(\hat{t}_{rat}) = N^2 \left(1 - \frac{n}{N}\right)\left(\frac{1}{n}\right)\sum_{k=1}^{n} \frac{(y_k - \hat{r}z_k)^2}{n-1}. \tag{3.26}$$

By studying the sum of squares in the variance equation (3.25), it is possible to find the condition under which ratio estimation results in an improved estimate

of a total. The total sum of squares can be decomposed as follows:

$$\begin{aligned}\sum_{k=1}^{N}(Y_k - R \times Z_k)^2/(N-1) &= \sum_{k=1}^{N}[(Y_k - \overline{Y}) - R(Z_k - \overline{Z})]^2/(N-1) \\ &= \sum_{k=1}^{N}[(Y_k - \overline{Y})^2 - R^2(Z_k - \overline{Z})^2 \\ &\quad - 2R(Y_k - \overline{Y})(Z_k - \overline{Z})]/(N-1) \\ &= S_y^2 + R^2 S_z^2 - 2R\rho_{yz}S_y S_z,\end{aligned}$$

where ρ_{yz} is the finite-population correlation coefficient of the variables y and z. Consider the difference

$$V_{srs}(\hat{t}) - V_{srs}(\hat{t}_{rat}) = N^2\left(1 - \frac{n}{N}\right)\left(\frac{1}{n}\right)\{S_y^2 - [S_y^2 + R^2 S_z^2 - 2R\rho_{yz}S_y S_z]\}.$$

The ratio estimator improves efficiency if $V_{srs}(\hat{t}) > V_{srs}(\hat{t}_{rat})$, which occurs when

$$R^2 S_z^2 < 2R\rho_{yz}S_z S_y$$

is valid or

$$2\rho_{yz} > \frac{RS_z}{S_y}.$$

It should be noted that $R = \overline{Y}/\overline{Z}$, and that the former condition expressed in terms of coefficients of variation (C.V) of the variables z and y is given by

$$\rho_{yz} > \left(\frac{1}{2}\right)\frac{\text{C.V}_y}{\text{C.V}_z},$$

where $\text{C.V}_y = S_y/\overline{Y}$ and $\text{C.V}_z = S_z/\overline{Z}$ are the coefficients of variation of y and z respectively. Therefore, improvement in efficiency depends on the correlation between the study and auxiliary variables y and z and the C.V of each variable.

Example 3.11

Efficiency of a ratio-estimated total in the *Province'91* population. The variable UE91 is the study variable y and HOU85 is chosen as the auxiliary variable z. The correlation coefficient between UE91 and HOU85 is $\rho_{yz} = 0.9967$, and the corresponding coefficients of variation are $\text{C.V}_y = S_y/\overline{Y} = 743/472 = 1.57$ and $\text{C.V}_z = S_z/\overline{Z} = 4772/2867 = 1.66$. Thus, the condition given above is valid since

$$\rho_{yz} = 0.9967 > 0.4729 = \frac{1}{2} \times \frac{1.57}{1.66}.$$

It can be seen that the ratio estimation improves the efficiency. The improvement can also be measured directly as a design effect. In addition to the parameters given, the ratio $R = \overline{Y}/\overline{Z} = 472/2867 = 0.1646$ is required. The value of the design effect of the ratio-estimated total $\hat{t}_{rat}$ in the *Province'91* population is given by

$$\begin{aligned}\mathrm{DEFF}_{srs}(\hat{t}_{rat}) &= \frac{S_y^2 + R^2 S_z^2 - 2R\rho_{yz} S_y S_z}{S_y^2} \\ &= \frac{743^2 + 0.1646^2 \times 4772^2 - 2 \times 0.1646 \times 0.9967 \times 743 \times 4772}{743^2} \\ &= 0.0102\end{aligned}$$

which is close to 0. This substantial improvement in efficiency is due to the favourable relationship between UE91 and HOU85 such that the ratio Y_k/Z_k is nearly constant across the population.

The ratio-estimated total is in practice calculated using the available survey data under the actual sample design. If the design is, say, stratified SRS, the corresponding parameters would be estimated by using appropriate stratum weights. The present example was evaluated under simple random sampling without replacement, which will also be used in the following example. There, the use of g weights will also be illustrated.

Example 3.12

Calculating a ratio-estimated total from a simple random sample drawn without replacement from the *Province'91* population. Again we use UE91 as the study variable and HOU85 as the auxiliary variable. The estimated ratio is $\hat{r} = \overline{y}/\overline{z} = 0.1603$, which is calculated from the sample in Table 3.15. The sample identifiers are STR = 1, ID is the cluster identifier, and the weight is WGHT = 4.

Table 3.15 A simple random sample drawn without replacement from the *Province'91* population prepared for ratio estimation.

Sample design identifiers			Element	Study var.	Aux. var.	g	Adj.
STR	CLU	WGHT	LABEL	UE91	HOU85	WGHT	WGHT
1	1	4	Jyväskylä	4123	26 881	0.5562	2.2248
1	4	4	Keuruu	760	4896	0.5562	2.2248
1	5	4	Saarijärvi	721	3730	0.5562	2.2248
1	15	4	Konginkangas	142	556	0.5562	2.2248
1	18	4	Kuhmoinen	187	1463	0.5562	2.2248
1	26	4	Pihtipudas	331	1946	0.5562	2.2248
1	30	4	Toivakka	127	834	0.5562	2.2248
1	31	4	Uurainen	219	932	0.5562	2.2248

Sampling rate: $8/32 = 0.25$.

To carry out ratio estimation of the total, the calibrated weights w_k^* are first calculated. The sampling weight w_k is a constant $w_k = N/n = 32/8 = 4$ as before. The values of the g weight are $g_k = T_z/\hat{t}_z$. The population total of the auxiliary variable is $T_z = 91\,753$ and its estimate calculated from the sample is $\hat{t}_z = 164\,952$. Thus, the g weight is the constant $g_k = 91\,753/164\,952 = 0.5562$. Multiplying the weight w_k by the g weight gives the value for the calibrated weight $w_k^* = 4 \times 0.5562 = 2.2248$.

The ratio estimate for the total is calculated as

$$\hat{t}_{rat} = \sum_{k=1}^{n} w_k^* y_k = \hat{r} \times T_z = 0.1603 \times 91\,753 = 14\,707,$$

which is much closer to the population total $T = 15\,098$ than the SRSWOR estimate $\hat{t} = 26\,440$ for the total number of unemployed. The variance estimate for the total estimator is

$$\hat{v}_{srs}(\hat{t}_{rat}) = 32^2 \frac{(1 - 0.25)}{8} \times 91^2 = 892^2.$$

The corresponding deff estimate is

$$\text{deff}_{srs}(\hat{t}_{rat}) = \frac{\hat{v}_{srs}(\hat{t}_{rat})}{\hat{v}_{srs}(\hat{t})} = 892^2/13\,282^2 = 0.0045,$$

which also shows that ratio estimation improves the efficiency. The minimal auxiliary information of the population total T_z and the sample values of z yield good results.

It is also possible to calculate the DEFF when using the ratio-estimated total since the variance $V_{srs}(\hat{t}_{rat})$ is

$$\begin{aligned} V_{srs}(\hat{t}_{rat}) &\doteq N^2 \left(1 - \frac{n}{N}\right) \left(\frac{1}{n}\right) \sum_{k=1}^{N} \frac{(Y_k - R \times \mathrm{Z}_k)^2}{(N-1)} \\ &= 32^2 \frac{(1 - 0.25)}{8} \times 75^2 = 736^2. \end{aligned}$$

Division by the corresponding SRSWOR design variance of $\hat{t}$ gives

$$\text{DEFF}_{srs}(\hat{t}_{rat}) = \frac{V_{srs}(\hat{t}_{rat})}{V_{srs}(N\overline{y})} = 736^2/7283^2 = 0.0102,$$

which is the same figure presented previously in Example 3.11.

For these data, ratio estimation considerably improves efficiency and brings the point estimate for the total close to its population value. The value of the

ratio estimator is based on the fact that across the population, the ratio Y_k/Z_k remains nearly constant. It should be noted that even a high correlation between the variables does not guarantee this, because the ratio estimator assumes that the regression line of y and z goes near the origin. Thus, an intercept term is not included in the corresponding regression equation. The ratio estimator may therefore be unfavourable if the population regression line intercepts the y-axis far from the origin, even if the correlation is not close to zero. For these situations, the method presented next would be more appropriate.

Regression Estimation of Totals

Regression estimation of the population total T of a study variable y is based on the linear regression between y and a continuous auxiliary variable z. The linear regression can, for example, be given by $E_M(y_k) = \alpha + \beta \times z_k$ with a variance $V_M(y_k) = \sigma^2$, where y_k are independent random variables with the population values Y_k as their assumed realizations, α, β and σ^2 are unknown parameters, Z_k are known population values of z, and E_M and V_M refer respectively to the expectation and variance under the model. The finite-population analogues of α and β, denoted respectively by A and B, are estimated from the sample using weighted least squares estimation so that the sampling design is properly taken into account. It is immediately obvious that multiple auxiliary variables can also be incorporated in the model. Note that the model assumption introduces a new type of randomness; in the estimation considered previously, the sample selection was the only source of random variation.

We consider the basic principles of regression estimation for SRS without replacement using the above regression model with a single auxiliary variable. The finite-population quantities A and B are estimated by the ordinary least squares method giving $\hat{b} = \hat{s}_{yz}/\hat{s}_z^2$ as an estimator of the slope B and $\hat{a} = \overline{y} - \hat{b}\overline{z}$ as an estimator of the intercept A. Using the estimator $\hat{b}$, the *regression estimator* of the total T of y is given by

$$\hat{t}_{reg} = N(\overline{y} + \hat{b}(\overline{Z} - \overline{z})) = \hat{t} + \hat{b}(T_z - \hat{t}_z) \tag{3.27}$$

where $\hat{t} = N\overline{y}$ is the SRSWOR estimator of T, $\hat{t}_z = N\overline{z}$ is the SRSWOR estimator of T_z and $\overline{Z} = T_z/N$. Alternatively, if transformed values $z_k^* = \overline{Z} - z_k$ are used in the regression instead of z_k, an estimated intercept for this model is $\hat{a}^* = \hat{a} + \hat{b}\overline{Z}$ giving $\hat{t}_{reg} = N\hat{a}^*$, because (3.27) can be written also as $\hat{t}_{reg} = N\hat{a} + \hat{b}T_z$. Note that the regression estimation of the total T presupposes only knowledge of the population total T_z and the sample values z_k of the auxiliary variable z.

Regression estimators constitute a wide class of estimators. For example, the previous ratio estimator $\hat{t}_{rat} = \hat{r}T_z$ is a special case of (3.27) such that the intercept A is assumed 0 and the slope B is estimated by $\hat{b} = \hat{r} = \hat{t}/\hat{t}_z$.

Alternatively, we can calculate calibrated weights $w_k^* = w_k \times g_k$ where w_k is the sampling weight and the g weight is calculated from

$$g_k = \frac{N}{\hat{N}}\left[1 + \frac{\overline{Z} - \overline{z}}{\frac{n-1}{n}\hat{s}_z^2} \times (z_k - \overline{z})\right],$$

where $\overline{Z}$ is the population mean and $\overline{z}$ is the sample mean of the auxiliary variable z, the sum of the sampling weights is $\sum_{k=1}^{n} w_k = \hat{N}$ and

$$\hat{s}_z^2 = \frac{\sum_{k=1}^{n}(z_k - \overline{z})^2}{n-1}.$$

The weights g_k and calibrated weights w_k^* are presented under the model $E_M(y_k) = \alpha + \beta \times z_k$ in Table 3.16 for an SRSWOR sample from the *Province'91* Population. A regression estimate for the population total thus is the calibrated weight w_k^* multiplied by the observed value y_k and summed-up over all sample elements. The regression estimator given in (3.27) can thus also be expressed as $\hat{t}_{reg} = \sum_{k=1}^{n} w_k^* y_k$.

An approximate design variance of $\hat{t}_{reg}$ under SRSWOR is given by

$$V_{srs}(\hat{t}_{reg}) \doteq N^2\left(1 - \frac{n}{N}\right)\left(\frac{1}{n}\right)S_E^2, \tag{3.28}$$

where $S_E^2 = \sum_{k=1}^{N}(E_k - \overline{E})^2/(N-1)$, $E_k = Y_k - \hat{Y}_k$ and $\overline{E} = \sum_{k=1}^{N} E_k/N$ is the mean of population residuals. The fitted values $\hat{Y}_k = A + B \times Z_k$ are calculated from the population values. An approximate estimator of the design variance of $\hat{t}_{reg}$ under SRSWOR design is given by substituting S_E^2 by an estimate $\hat{s}_{\hat{e}}^2 = \sum_{k=1}^{n}(\hat{e}_k - \overline{\hat{e}})^2/(n-1)$, where $\hat{e}_k = y_k - \hat{y}_k$ and $\overline{\hat{e}} = \sum_{k=1}^{n}\hat{e}_k/n$. Fitted values $\hat{y}_k = \hat{a} + \hat{b} \times z_k$ are calculated from the sample values. An alternative, more conservative estimator, which uses g-weights is given by

$$\hat{v}_{srs}(\hat{t}_{reg}) = N^2\left(1 - \frac{n}{N}\right)\left(\frac{1}{n}\right)\left(\frac{n-1}{n-p}\right) \times \hat{s}_{\hat{e}^*}^2, \tag{3.29}$$

where $\hat{s}_{\hat{e}^*}^2 = \sum_{k=1}^{n}(\hat{e}_k^* - \overline{\hat{e}}^*)^2/(n-1)$, $\hat{e}_k^* = g_k \times \hat{e}_k$, $\overline{\hat{e}}^* = \sum_{k=1}^{n}\hat{e}_k^*/n$ and p is the number of estimated model parameters.

The improvement gained in regression estimation, as compared to the corresponding simple-random-sampling estimators, depends on the value of the finite-population correlation coefficient $\rho_{yz} = S_{yz}/(S_y S_z)$ between the variables y and z. This can be seen by writing the approximate variance (3.28) in the form

$$V_{srs}(\hat{t}_{reg}) \doteq N^2\left(1 - \frac{n}{N}\right)\left(\frac{1}{n}\right)S_y^2(1 - \rho_{yz}^2). \tag{3.30}$$

It will be noted that the value of the correlation coefficient has a decisive influence on the possible improvement of the regression estimation. If ρ_{yz} is zero, the variance of the regression estimator $\hat{t}_{reg}$ equals that of the SRSWOR counterpart $\hat{t}$. But with a nonzero correlation coefficient, the variance obviously decreases.

Under certain conditions, the regression estimator of a total is more efficient than the ratio estimator. This will be demonstrated below by considering the variances of the SRSWOR estimator, the ratio estimator and the regression estimator. Simple random sampling without replacement is assumed, and the constant (c) given in the formulae represents $c = N^2(1 - (n/N))(1/n)$ The variances are

Design-based estimator	$V_{srs}(\hat{t}) = cS_y^2$
Ratio estimator	$V_{srs}(\hat{t}_{rat}) = c(S_y^2 + R^2S_z^2 - 2R\rho_{yz}S_yS_z)$
Regression estimator	$V_{srs}(\hat{t}_{reg}) = cS_y^2(1 - \rho_{yz}^2)$

Studying the relationship between the regression coefficient B and the ratio $R = T/T_z$ will reveal the condition where the regression-estimated total is more efficient than the ratio-estimated total. To find this condition, the difference between the two variances is

$$\begin{aligned} V_{srs}(\hat{t}_{rat}) - V_{srs}(\hat{t}_{reg}) &= c[(S_y^2 + R^2S_z^2 - 2R\rho_{yz}S_yS_z) - S_y^2 + \rho_{yz}^2S_y^2] \\ &= c[(R^2S_z^2 - 2R\rho_{yz}S_yS_z) + \rho_{yz}^2S_y^2]. \end{aligned}$$

Regression estimation is more efficient if the difference is positive:

$$R^2S_z^2 - 2R\rho_{yz}S_yS_z + \rho_{yz}^2S_y^2 > 0.$$

The condition can be rewritten as

$$-\rho_{yz}^2S_{y-}^2 < R^2S_z^2 - 2R\rho_{yz}S_yS_z.$$

By dividing the inequality above by S_z^2 and inserting $\rho_{yz} = S_{yz}/S_yS_z$ and $B = S_{yz}/S_z^2$, gives

$$-B^2 < R^2 - 2RB.$$

Regression estimation, then, is more efficient than ratio estimation if

$$(B - R)^2 > 0.$$

Thus the squared difference between the finite-population regression coefficient and the ratio determines when the regression estimation is more efficient.

Regression estimation can also be applied using a multiple regression model as the assisting model. We postulate a linear regression model between the study variable y and p continuous auxiliary variables $z_1, z_2, \ldots, z_p$, given by

$y_k = \alpha + \beta_1 z_{1k} + \beta_2 z_{2k} + \cdots + \beta_p z_{pk} + \varepsilon_k$, where α refers to the intercept and $\beta_j, j = 1, \ldots, p$, are the slope parameters, and ε_k is the residual. For multiple regression estimation, we assume that the population totals $T_{z_1}, T_{z_2}, \ldots, T_{z_p}$ are known for each auxiliary variable. They can come from some source outside the survey, such as published official statistics. The regression estimator of the population total T of y is now given by

$$\hat{t}_{reg} = \hat{t} + \hat{b}_1(T_{z_1} - \hat{t}_{z_1}) + \hat{b}_2(T_{z_2} - \hat{t}_{z_2}) + \cdots + \hat{b}_p(T_{z_p} - \hat{t}_{z_p}), \tag{3.31}$$

where the estimated regression coefficients $\hat{b}_1, \hat{b}_2, \ldots, \hat{b}_p$ are obtained from the sample data set using weighted least squares estimation with $w_k = 1/\pi_k$ as the weights. The estimators $\hat{t}$ and $\hat{t}_{z_j}, j = 1, \ldots, p$, refer to Horvitz–Thompson estimators.

A different form, often referred to as the *generalized regression* (GREG) *estimator* (Särndal *et al.* 1992) is given by

$$\hat{t}_{reg} = \sum_{k=1}^{N} \hat{y}_k + \sum_{k=1}^{n} w_k(y_k - \hat{y}_k), \tag{3.32}$$

where $\hat{y}_k = \hat{a} + \hat{b}_1 z_{1k} + \hat{b}_2 z_{2k} + \cdots + \hat{b}_p z_{pk}$ are fitted values calculated using the estimated regression coefficients and the known values of z-variables. Note the difference between (3.31) and (3.32). In the former we only need to know the population totals of the auxiliary z-variables, but in the latter, the individual values of z-variables are assumed known for every population element (because the first summation is over all N population elements). Thus, (3.32) requires more detailed information on the population than (3.31). Micro-level auxiliary z-data may indeed be available, for example, in a statistical infrastructure where population census registers or similar statistical registers, compiled from various administrative registers, are used as sampling frames. In this case, the frame population often includes the necessary auxiliary z-data at a micro-level (see Chapter 6).

Let us consider the expression (3.32) for a multiple regression estimator in more detail. It is obvious that if the weights are equal for all sample elements, and ordinary least squares estimation had been used for a model that includes an intercept, then the latter part of (3.32) vanishes, and the regression estimate reduces to the sum of the fitted values over the population. This is the case for a self-weighting design such as simple random sampling. But if the weights vary between elements, then the sum of weighted residuals can differ from zero, as can happen for example in stratified SRS with non-proportional allocation. In such cases, the latter part of (3.32) serves as a bias adjustment factor protecting against model misspecification.

Under SRSWOR, an approximate design variance given in (3.28) can be applied by using the fitted values $\hat{Y}_k = A + B_1 Z_{1k} + \cdots + B_p Z_{pk}$. A variance estimator is

obtained by replacing $\hat{Y}_k$ by sample-based fitted values $\hat{y}_k = \hat{a} + \hat{b}_1 z_{1k} + \cdots + \hat{b}_p z_{pk}$. An alternative variance estimator is calculated as

$$\hat{v}_{srs}(\hat{t}_{reg}) = \hat{v}_{srs}(\hat{t})(1 - \hat{R}^2), \tag{3.33}$$

where the multiple correlation coefficient squared $\hat{R}^2$ is calculated for the sample data set. Because this term is always non-negative, the multiple regression estimator is always at least as efficient as simple random sampling without replacement. Efficiency improves when multiple auxiliary z-data that correlates with the study variable y are incorporated in the estimation procedure.

In the next example, we compute a regression-estimated total from a sample data set, first in a single auxiliary variable case and then in the context of multiple regression estimation.

Example 3.13

Single Auxiliary Variable

Regression estimation of the total in the *Province'91* population. The previously selected simple random sample is used. There, the study variable UE91 is regressed with the auxiliary variable HOU85. We conduct regression estimation in two ways, resulting in equal estimates. HOU85 is first used as the predictor and an estimate $\hat{t}_{reg}$ is computed using the estimated slope $\hat{b}$. In Table 3.16, the sample identifiers correspond to the SRSWOR case, and the sampling rate is, as previously, 0.25.

Using UE91 as the dependent variable and HOU85 as the predictor, the slope is estimated as $\hat{b} = 0.152$, giving

$$\hat{t}_{reg} = \hat{t} + \hat{b}(T_z - \hat{t}_z) = 26\,440 + 0.152(91\,753 - 164\,952) = 15\,312.$$

Table 3.16 A simple random sample drawn without replacement from the *Province'91* population prepared for regression estimation.

Sample design identifiers			Element	Study var.	Auxiliary information			
					Variable	Model	WGHT	
STR	CLU	WGHT	LABEL	UE91	HOU85	group	g-weight	w*-weight
1	1	4	Jyväskylä	4123	26 881	1	0.2844	1.1378
1	4	4	Keuruu	760	4896	1	1.0085	4.0341
1	5	4	Saarijärvi	721	3730	1	1.0469	4.1877
1	15	4	Konginkangas	142	556	1	1.1057	4.6058
1	18	4	Kuhmoinen	187	1463	1	1.1216	4.4863
1	26	4	Pihtipudas	331	1946	1	1.1391	4.4227
1	30	4	Toivakka	127	834	1	1.1423	4.5691
1	31	4	Uurainen	219	932	1	1.1515	4.5562

Sampling rate = 8/32 = 0.25.

The same point estimate is obtained using the calibrated weights by calculating $\hat{t}_{reg} = \sum_{k=1}^{8} w_k^* y_k = 15\,312$ (see Table 3.16). For variance estimation, the formula (3.29) or (3.33) can be used. The former gives a conservative estimate especially if the sample size is small as is the case here. Thus, by (3.29) we obtain

$$\begin{aligned}\hat{v}_{srs}(\hat{t}_{reg}) &= N^2 \left(1 - \frac{n}{N}\right) \left(\frac{1}{n}\right) \left(\frac{n-1}{n-p}\right) \times \hat{s}^2_{\hat{e}^*} \\ &= 32^2 \left(1 - \frac{8}{32}\right) \left(\frac{8-1}{8-2}\right) \left(\frac{1}{8}\right) \times 61.24^2 = 648^2.\end{aligned}$$

The corresponding design-based total estimate obtained under SRSWOR was $\hat{t} = 26\,440$, whose standard error was 13 282. Therefore, the deff estimate is deff $= 648^2/13\,282^2 = 0.002$, which is almost zero and is persuasive evidence of the superiority of regression estimation over design-based estimation for the present estimation problem. Improved efficiency is due to the strong linear relationship between UE91 and HOU85.

Multiple Regression Model

Multiple regression estimation of the total in the *Province'91* population. Here, the study variable UE91 is regressed with two auxiliary variables, HOU85 and a variable named URB85 with a value 1 for urban municipalities and zero otherwise (see Table 2.1). We use both the formula (3.31) and the GREG method with equation (3.32). First, the estimated regression coefficients $\hat{b}_1$ and $\hat{b}_2$ are calculated by fitting a two-predictor regression model for the sample data set of $n = 8$ municipalities, as given in Table 3.16. The estimates are $\hat{b}_1 = 0.14956$ and $\hat{b}_2 = 68.107$. The estimated totals of auxiliary variables are $\hat{t}_{z_1} = 164\,952$, as previously, and $\hat{t}_{z_2} = 12$. In addition, we use the known population totals $T_{z_1} = 91\,753$ and $T_{z_2} = 7$. Using (3.31), we obtain

$$\begin{aligned}\hat{t}_{reg} = \hat{t} + \hat{b}_1(T_{z_1} - \hat{t}_{z_1}) + \hat{b}_2(T_{z_2} - \hat{t}_{z_2}) &= 26\,440 + 0.14956\,(91\,753 - 164\,952) \\ &+ 68.107\,(7 - 12) = 15\,152.\end{aligned}$$

Using (3.32), we first calculate the fitted values for all population elements. The sum of the fitted values over the population provides the desired regression estimate. The GREG estimation procedure is summarized in Table 3.17. There also, the estimate 15 152 can be obtained. Note that in the SRSWOR case in which the sampling weights are equal, the sum of the residuals over the sample data set is equal to zero.

Calculating the multiple correlation coefficient squared $\hat{R}^2 = 0.998$ for the sample data set, we obtain the variance estimate of $\hat{t}_{reg}$ by (3.33), $\hat{v}(\hat{t}_{reg}) = 569^2$, which is smaller than in the previous case where HOU85 was used as the only auxiliary variable. There, an estimate $\hat{v}(\hat{t}_{reg}) = 648^2$ was obtained. Hence, multiple regression estimation appeared to be slightly more efficient in this case. The design effect estimate is now deff $= 569^2/13\,282^2 = 0.0018$.

Table 3.17 Population frame merged with sample data for multiple regression estimation. Simple random sample drawn without replacement from the *Province'91* population.

Population frame					Sample		Model fitting	
ID k	LABEL	URB85 z_{1k}	HOU85 z_{2k}	Sample indicator	WGHT w_k	UE91 y_k	Fitted value $\hat{y}_k$	Residual $\hat{e}_k$
1	**Jyväskylä**	**1**	**26 881**	**1**	**4**	**4123**	**4118.15**	**4.85**
2	Jämsä	1	4663	0	...	...	795.27	...
3	Jämsänkoski	1	3019	0	...	...	549.40	...
4	**Keuruu**	**1**	**4896**	**1**	**4**	**760**	**830.12**	**−70.12**
5	**Saarijärvi**	**1**	**3730**	**1**	**4**	**721**	**655.73**	**65.27**
6	Suolahti	1	2389	0	...	...	455.18	...
7	Äänekoski	1	4264	0	...	...	735.60	...
8	Hankasalmi	0	2179	0	...	...	355.66	...
9	Joutsa	0	1823	0	...	...	302.42	...
10	J:kylä mlk.	0	9230	0	...	...	1410.20	...
11	Kannonkoski	0	726	0	...	...	138.36	...
12	Karstula	0	1868	0	...	...	309.15	...
13	Kinnula	0	675	0	...	...	130.73	...
14	Kivijärvi	0	634	0	...	...	124.60	...
15	**Konginkangas**	**0**	**556**	**1**	**4**	**142**	**112.93**	**29.07**
16	Konnevesi	0	1215	0	...	...	211.49	...
17	Korpilahti	0	1793	0	...	...	297.93	...
18	**Kuhmoinen**	**0**	**1463**	**1**	**4**	**187**	**248.58**	**−61.58**
19	Kyyjärvi	0	672	0	...	...	130.28	...
20	Laukaa	0	4952	0	...	...	770.39	...
21	Leivonmäki	0	545	0	...	...	111.29	...
22	Luhanka	0	435	0	...	...	94.83	...
23	Multia	0	925	0	...	...	168.12	...
24	Muurame	0	1853	0	...	...	306.91	...
25	Petäjävesi	0	1352	0	...	...	231.98	...
26	**Pihtipudas**	**0**	**1946**	**1**	**4**	**331**	**320.82**	**10.18**
27	Pylkönmäki	0	473	0	...	...	100.52	...
28	Sumiainen	0	485	0	...	...	102.31	...
29	Säynätsalo	0	1226	0	...	...	213.13	...
30	**Toivakka**	**0**	**834**	**1**	**4**	**127**	**154.51**	**−27.51**
31	**Uurainen**	**0**	**932**	**1**	**4**	**219**	**169.16**	**49.84**
32	Viitasaari	0	3119	0	...	...	496.25	...
	Sum	**7**	**91 753**	**8**	**32**	**6610**	**15 151.98**	**0.00**

. . .Nonsampled element.

Regression estimation was illustrated in simple cases where one or two auxiliary variables were used and SRSWOR was assumed. The method can also be applied for more complex designs, and multiple auxiliary variables can be incorporated in the estimation. For this, weighted least squares regression can also be used. Although the use of multivariate regression models for regression estimation is technically straightforward, there are certain complexities when compared

to regression estimation under simple random sampling, such as the possible multicollinearity of the predictor variables. Another generalization is also obvious since discrete covariates can also be incorporated into a linear model. Using this kind of auxiliary variables for regression estimation leads to analysis-of-variance-type models. Further extensions are discussed in Chapter 6 in connection with the estimation for population subgroups.

Comparison of Estimation Strategies

For model-assisted estimation, we created three sets of new weights, denoted w^*. First, we check the calibration property of these weights. For ratio estimation, the calibration equation for the auxiliary variable z is

$$\sum_{k=1}^{n} w_k^* \times z_k = T_z$$

where $T_z = \sum_{k=1}^{N} Z_k = 91\,753$. This holds for the regression estimator as well.

We next compare the model-assisted estimation results obtained previously from a sample drawn with SRSWOR from the *Province'91* population. More specifically, poststratification, ratio estimation and regression estimation results for the population total T of UE91 are compared. The design-based estimate using the standard SRS formula is also included (see Table 3.18). The known population total $T = 15\,098$ of UE91 is the reference figure.

Two obvious conclusions can be drawn. Firstly, point estimates calculated using auxiliary information are closer to the population total than the design-based estimate. Secondly, the model-assisted estimators are much more efficient than SRSWOR.

The poststratified estimator uses, as discrete auxiliary information, the administrative division of municipalities into urban and rural municipalities. Improved

Table 3.18 Estimates for the population total of UE91 under different estimation strategies: an SRSWOR sample of eight elements drawn from the *Province'91* population.

Estimation strategy		Estimator	Estimate	s.e	deff
Desing-based					
SRSWOR		$\hat{t}_{srswor}$	26 440	13 282	1.0000
SRSWR		$\hat{t}_{srswr}$	26 440	15 095	1.2917
Design-based model-assisted					
Poststratified estimator		$\hat{t}_{pos}$	18 106	6021	0.3323
Ratio estimator		$\hat{t}_{rat}$	14 707	892	0.0045
Regression estimator	one z-variable	$\hat{t}_{reg,1}$	15 312	648	0.0020
	two z-variables	$\hat{t}_{reg,2}$	15 152	569	0.0018

estimates result, since this division is in relation to the variation of the study variable in such a way that the variation of unemployment figures is smaller in the poststrata than in the whole population. But the relation is not as strong as that between UE91 and the continuous auxiliary variable HOU85, the number of households. This can be seen from the ratio and regression estimation results. Because ratio estimation assumes that the regression line of UE91 and HOU85 goes through the origin, and this is not the case, regression estimation performs slightly better than ratio estimation.

Summary

Using auxiliary information from the population in the estimation of a finite-population parameter of interest is a powerful tool to get more precise estimates, if the variation of the study variable has some strong relationship with an auxiliary covariate. If so, efficient estimators can be obtained such that they produce estimates close to the true population value and have a small standard error. The auxiliary variable can be a discrete variable, in which case poststratification can be used. If the covariate is a continuous variable, ratio estimation or regression estimation is appropriate.

Model-assisted estimation is often used in descriptive surveys to improve the estimation of the population total of a study variable of interest, whereas in multi-purpose studies, where the number of study variables may be large, it may be difficult to find good auxiliary covariates for this purpose. In such surveys, however, poststratification is often used to adjust for nonresponse.

We have examined here the elementary principles of model-assisted estimation supplemented with computational illustrations. For more details, the reader is encouraged to consult Särndal *et al.* (1992); there, model-assisted survey sampling covering poststratification, ratio estimation and regression estimation is extensively discussed. These methods are considered as special cases of *generalized regression estimation* which is used in many statistical agencies in the production of official statistics (for example Estevao *et al.* 1995). A clear overview of poststratification can be found in Holt and Smith (1979). Further, as a generalization of poststratification, Deville and Särndal (1992) and Deville *et al.* (1993) consider a class of weights calibrated to known marginal totals. Silva and Skinner (1997) address the problem of variable selection in regression estimation.

3.4 EFFICIENCY COMPARISON USING DESIGN EFFECTS

The design effect provides a convenient tool for the comparison of efficiency of the estimation of the population parameter of interest under various sampling designs. In this section, we summarize the findings on efficiency evaluations from the preceding sections.

Efficiency is derived by comparing the variance of an estimator with that obtained under SRSWOR, and is measured as the population design effect (DEFF), or as an estimated design effect (deff) calculated from the selected sample. We previously evaluated the efficiency in three ways: (1) analytically, by deriving the corresponding design variance formulae, (2) population-based, by calculating from the small fixed population, the *Province'91* population, the true value of the design variances, and (3) sample-based, by estimating the design variances from one realization of a sampling design applied to the *Province'91* population. Evaluation by these methods covered all the basic sampling techniques considered. In the sample-based evaluation of the design effect using an estimated deff, we considered the estimators of a total, a ratio and a median.

Let us consider first the evaluation of efficiency for the estimation of the total T of a study variable y. The design effect is defined as a ratio of two design variances: the actual variance $V_{p(s)}(\hat{t}^*)$ of an estimator $\hat{t}^*$ of the total, reflecting properly the sampling design, and the variance $V_{srs}(N\overline{y})$ derived assuming SRSWOR, where $\hat{t}^*$ is the design-based estimator of the total under the design $p(s)$ and $N\overline{y} = \hat{t}$ is the corresponding SRSWOR estimator. Note that the two estimators of the total may be different, and the same sample size is assumed as for the actual sampling design. The DEFF is thus

$$\text{DEFF}_{p(s)}(\hat{t}^*) = V_{p(s)}(\hat{t}^*)/V_{srs}(N\overline{y}) \tag{3.34}$$

as defined in Section 2.1. The equation indicates that if DEFF > 1, then the actual design is less efficient than SRSWOR; if DEFF is approximately 1, then the designs are equally efficient; and if DEFF < 1, the efficiency of the actual design is superior to SRSWOR.

Analytical Evaluation of Design Effect

The analytical evaluation of DEFF is possible if the population parameters in the variance equations, such as the population variance S^2, cancel out in the formula of the design effect. For example, the design effect under simple random sampling with replacement (SRSWR) can be calculated for a given sample size n and population size N. Hence, we have DEFF $= (N-1)/(N-n)$ with the result that the design effect for SRSWR is greater than or equal to 1. It is also sometimes possible to identify conditions when the DEFF will be less than 1 and the actual design will be more efficient than SRSWOR.

Analytical evaluation of the design effect for an estimator of a total is illustrated for stratified simple random sampling, sampling with probabilities proportional to a size measure, and cluster sampling. Systematic sampling is excluded because it can be considered a special case of cluster sampling.

1. *Stratified sampling with proportional allocation (STR)* Factors affecting efficiency under STR are the possible heterogeneity of separate strata and internal

homogeneity within each stratum. The design effect for an estimator $\hat{t}^* = \hat{t}$ of the total T of the study variable y is

$$\text{DEFF}_{str}(\hat{t}) \doteq \frac{\sum_{h=1}^{H} W_h S_h^2}{S^2}, \tag{3.35}$$

where S_h^2 are stratum variances and S^2 is the population variance of y (see Section 3.1). In stratified sampling, the DEFF is usually less than one, which happens when the strata are internally homogeneous with respect to the variation of the study variable, i.e. if the stratum variances are small.

2. *Sampling with probability proportional to a measure of size (PPS)* The value of an auxiliary variable z measuring the size of a population element is required from all the units in the population. Assuming that the population regression line of y and z intercepts the y-axis near to the origin, an approximate equation of the design effect of an estimator $\hat{t} = \hat{t}_{HT}$ (the *Horvitz–Thompson* estimator) is given by

$$\text{DEFF}_{pps}(\hat{t}_{HT}) \doteq (1 - \rho_{yz}^2), \tag{3.36}$$

where ρ_{yz} is the finite-population correlation coefficient between the study variable y and the size measure z (see Section 2.5). Given the above condition, if z is a good size measure correlating strongly with y, a DEFF smaller than one is obtained.

3. *Cluster sampling (CLU)* The design effect under CLU depends on the value of the intra-cluster coefficient ρ_{int} of the study variable y measuring internal homogeneity of the population clusters. Assuming equal-sized clusters, an approximative equation of the design effect of an estimator $\hat{t}$ is given by

$$\text{DEFF}_{clu}(\hat{t}) \doteq 1 + (B - 1)\rho_{int} \tag{3.37}$$

where B is the cluster size (see Section 3.2). Because in cluster sampling, the clusters are usually internally homogeneous, resulting in a positive ρ_{int}, the design effects tend to be greater than one.

To fully utilize the above formulae in planning a sampling design, it is necessary to know the variation of the study variable in the population. In choosing a sampling design, the planner would also need knowledge about the variation at stratum and cluster levels, and information on the correlation of the study variable and the size measure. In practice, however, this kind of information is rarely available, but in some cases approximations can be taken from auxiliary sources, or by carrying out a smaller pilot study.

Population Design Effects

We next perform a numerical evaluation of the population design effects for the total by calculating the design variances by the corresponding formulae for the

Table 3.19 Population DEFFs for a total estimator under various sampling designs for the *Province'91* population (fixed sample size $n = 8$).

Sampling design		S.E	DEFF
Sampling proportional to size (wr)	PPS	720	0.01
Stratified sampling (power alloc.)	STR	4852	0.44
Systematic sampling (random start)	SYS	5420	0.55
Cluster sampling (two-stage)	CLU2	6532	0.80
Cluster sampling (one-stage)	CLU1	6663	0.84
Simple random sampling	SRSWOR	7283	1.00

six sampling designs considered for the *Province'91* population. The fixed sample size is eight municipalities ($n = 8$) drawn from the population of 32 municipalities ($N = 32$). The values of the population design effects are displayed in Table 3.19.

PPS sampling with probability proportional to a measure of size appears to be the most efficient sampling design for the estimation of a total. The population DEFF is 0.01, which is very small. Improved efficiency is due to the relationship between UE91 and HOU85 (which was used as the size measure) such that the ratio of these variables is nearly constant across the population. It should be noted that the shape of the population distribution of the study variable UE91 also affects efficiency. The distribution of UE91 in the *Province'91* population is very skewed. However, under PPS, large selection probabilities are given for large clusters, such that the possible samples drawn from the population will vary to a rather small extent in their composition. Sample totals are thus not expected to vary much from sample to sample and this leads to efficient estimation. For improved efficiency, it is also beneficial if the study variable and the size measure are strongly correlated. In the case considered, the correlation was close to one.

Stratified sampling also appears to be quite efficient for the estimation of a total because the DEFF is 0.44, but the difference in favour of PPS is still noticeable. The stratification divided the municipalities into urban and rural ones, and it appeared that in urban municipalities there are more unemployed on average than in rural municipalities. The strata were thus internally homogeneous, a property that increases efficiency. The efficiency of systematic sampling is close to the STR design. Since there is a monotonic trend in the sampling frame, intra-class correlation becomes close to zero, leading to improved efficiency. Efficiency of two-stage cluster sampling is somewhat less than that of SYS, and one-stage cluster sampling is slightly less efficient than two-stage cluster sampling.

Sample Design Effects

The previous efficiency comparisons were theoretical in the sense that we considered the design variances at the population level. We next evaluate the efficiency

from a selected sample of size $n = 8$ units drawn from the *Province'91* population. We thus obtain an estimated design effect, calculated by the corresponding variance estimates $\hat{v}_{p(s)}(\hat{\theta}^*)$ and $\hat{v}_{srs}(\hat{\theta})$, which for an estimator of a population parameter θ is given by

$$\text{deff}_{p(s)}(\hat{\theta}^*) = \frac{\hat{v}_{p(s)}(\hat{\theta}^*)}{\hat{v}_{srs}(\hat{\theta})}, \tag{3.38}$$

where $\hat{\theta}^*$ is a design-based estimator of θ and $\hat{\theta}$ is the SRSWOR counterpart.

Using the sample deff (see Table 3.20), the efficiency of estimation under the given sample obtained with the various sampling designs $p(s)$ is compared for the estimators $\hat{t}^*$ (total), $\hat{r}^*$ (ratio) and $\hat{m}^*$ (median). There is a natural interpretation for these estimators in the *Province'91* population. The total measures the total number of unemployed (UE91) in the province, the ratio measures the unemployment rate, and the median gives an average number of unemployed per municipality.

The deff estimates vary not only between the sampling designs but also between the estimators for a given design. PPS and STR are the most efficient designs for the total because the deff estimates are close to zero. For the ratio, PPS and STR are superior to the others but have larger design effects than those calculated for the total. For the median, the deff estimates are close to zero under SYS with implicit stratification and under STR.

Summary

The design effect provides a practical tool for the evaluation of efficiency of an estimator under a given sampling design. Using design effects, it is also possible to compare the efficiency of different sampling designs. The design effect clearly shows the effect of complex sampling relative to simple random sampling. Even for a scalar-type estimator, the sampling design can affect the design effect in various ways depending on the type of the estimator being considered. An estimator of

Table 3.20 The sample design effect estimates of the estimators of the total, the ratio and the median under the six different sampling designs; the *Province'91* population.

Sampling design		deff($\hat{t}^*$)	deff($\hat{r}^*$)	deff($\hat{m}^*$)
Sampling proportional to size	PPS	0.0035	0.19	0.92
Stratified sampling (power alloc.)	STR	0.21	0.38	0.19
Systematic sampling (implicit str.)	SYS	0.76	1.29	0.21
Cluster sampling (two-stage)	CLU2	0.93	0.99	0.84
Cluster sampling (one-stage)	CLU1	1.92	1.44	1.29
Simple random sampling	SRSWOR	1.00	1.00	1.00

a total is a linear-type estimator, a ratio estimator is a nonlinear estimator and a median is a robust estimator of a mean. These represent the types of estimator commonly used in statistical analysis. It is important to note that if an optimal design were desired for a given estimator, say for the total, so as to minimize its standard error, i.e. to produce a deff estimate close to zero, the optimality criterion would not necessarily be fulfilled for another estimator. In our examples, an estimator $\hat{m}$ for median seemed to be almost untouched by the design effect.

Design effects can be successfully utilized in the analysis of complex survey data. In the preceding sections, we used design effects mainly for descriptive purposes to solve estimation problems concerning a small fixed population. In the following chapters, we present several analytical situations and give further practical examples of the use of design effects. There, estimation and testing problems are considered for complex survey data from large populations. It will be shown, for example, that using design effects (or their generalizations) it is possible to estimate standard errors and calculate observed values of test statistics so that the complexities of a sampling design are properly accounted for. For both descriptive and analytical purposes, design effects can be obtained by using commercial software for survey analysis. Moreover, design effects are good indicators of the effects of complex sampling inherent in the computations. The papers by Kish and Frankel (1974) and Kish (1995) are recommended as further reading on this topic.

4

Handling Nonsampling Errors

In the survey estimation methodology discussed so far, the only source of variation has been the sampling error, which has been measured by the standard error of an estimator. In addition to the sampling error, there are also other sources of variation in surveys causing so-called *nonsampling errors*. In particular, these errors can be present in large-scale surveys. Survey organizations make efforts to minimize nonsampling errors occurring in the data-collection and data-processing phases. A good coverage of the frame population, carefully planned and tested measurement instruments, well-trained and motivated interviewers and well-implemented fieldwork and data-processing operations can guarantee a high response rate and minor measurement and processing errors and, thus, good total survey quality.

The important types of nonsampling errors are *nonresponse, coverage errors, measurement errors* and *processing errors*. Nonresponse implies that the intended measurements have not been obtained from all sample units. Coverage errors include the possible imperfections in the frame population. Measurement errors describe the difference between the observed value and the true value of a study variable. Processing errors cover such components as data entry and coding and editing errors, which can occur when the collected survey data are transformed to machine-readable form.

Nonsampling errors can cause biased estimation. Various techniques are available for adjusting for this undesirable effect of nonsampling errors. In the following two sections, we discuss in greater detail methods for adjusting for a particular source of nonsampling error, namely that caused by nonresponse. We will also demonstrate the adjusting for nonresponse by using the methods that have been described in previous sections. This chapter will be closed by a summary section covering a brief discussion on *total survey quality*. References for further reading will also be given.

Practical Methods for Design and Analysis of Complex Surveys Risto Lehtonen and Erkki Pahkinen
 ISBN: 0-470-84769-7

Nonresponse

Failure to obtain all the intended measurements or responses from all the selected sample members is called *nonresponse*. Nonresponse causes *missing data*, i.e. results in a data set whose size for the study variable y is smaller than planned. Two types of missing data are distinguished for a sample element. First, all intended measurements for a selected sample element could be missing, e.g. owing to a refusal to participate in a personal interview. The *unit nonresponse* has thus arisen, because all values of study variables are missing for that sample element. On the other hand, if an interviewed person does not respond to all of the questions, an *item nonresponse* has arisen, because a measurement for at least one study variable is missing for that element. Missing data of either type can give biased estimates and erroneous standard error estimates.

To illustrate typical response rates in large-scale surveys organized by governmental bodies, we summarize in Table 4.1 the response rates of six real large-scale surveys used in this book (see Chapter 1). Response rate refers here to the share of completed data-collection operations out of the total number of planned operations (usually, the share of completed interviews, or completed questionnaires, of the total sample size), for a given type of sampling unit. In the multinational PISA 2000 survey, the median of country-level response rates is presented owing to heavy country-wise variation.

The figures indicate a clear variation in response rates between surveys. The highest response rate is for the Mini-Finland Health Survey (96%) and the lowest is for the Passenger Transport Survey (65%). There can be different reasons for this variation: the attractiveness of the subject matter area of the survey, effectiveness of the fieldwork or the data-collection mode chosen, just to mention a few possibilities.

For example, in the Passenger Transport Survey (6), computer-assisted telephone interviewing (CATI) was used. A problem in this case was the identification of a phone number for every sampled unit, reducing possibilities for contact-making and thus excluding some sampled units out of the interview. In the two establishment surveys, (2) and (5), and in the PISA 2000 Survey (3), a

Table 4.1 Response rate in different surveys.

Name of the survey and section where the survey is described	Sampling unit	Sample size	Response rate (%)
(1) Mini-Finland Health Survey, Section 5.1	Person	8000	96
(2) Occupational Health Care Survey, Section 5.1	Establishment	1542	88
(3) PISA 2000 Survey, Section 9.4	School	6638	85
(4) Health Security Survey, Section 9.3	Household	6998	84
(5) Wages Survey, Section 9.2	Business firm	1572	80
(6) Passenger Transport Survey, Section 9.1	Person	18 250	65

self-administered questionnaire was used. Paper and pencil interview provides a traditional data-collection mode where trained interviewers contact the sample units and was used successfully in (1) and (4).

Readers more interested in nonresponse issues are advised to consult the brief technical descriptions included in each section in which the survey in question is discussed. The methods are demonstrated further in the web extension of the book.

The type of missing data, unit nonresponse or item nonresponse guides the selection of an appropriate method for adjusting for the nonresponse in an estimation procedure. Various *reweighting methods* are available for appropriate adjustment for the unit nonresponse. And for the item nonresponse, the missing values can be *imputed* by various imputation methods. Reweighting and imputation are discussed separately in the following two sections. Next, an example of a possible unfavourable impact of unit nonresponse is shown.

Impact of Unit Nonresponse

Unit nonresponse results in a sample data set whose size $n_{(r)}$ is smaller than the intended sample size n, thus increasing the standard errors of the estimates. This can be seen by considering the variance of an estimator $\hat{t}_{HT}$ of a population total T. Under simple random sampling without replacement (SRSWOR), this variance is $V_{srs}(\hat{t}_{HT}) = N^2(1 - n/N)S^2/n$, where the denominator is the original sample size n. If the number of respondents decreases because of unit nonresponse, the denominator decreases, and thus the variance increases.

A more serious consequence of unit nonresponse is that the estimation can become biased because of missing observations. This is particularly true if the probability θ_k of the kth population unit to respond depends on the value Y_k of the study variable y. Little and Rubin (1987) call this *nonignorable* nonresponse. This means that there is an association between the study variable and the probability to respond. For example, if the probability of responding to income-related questions decreases with increasing income level, then nonignorable nonresponse takes place. On the other hand, nonresponse is *ignorable* if Y_k is independent of θ_k. We point out two trivial situations when this is true. This happens when the value Y_k of the study variable is a constant ($Y_k = \overline{Y}$) for each population unit or the probability θ_k to respond is a constant θ for all k.

The following example concerns the effect of nonignorable nonresponse. As an extreme case, let us suppose that in an interview survey, a certain subgroup of the sample totally refuses to participate. In this case, the total population can be divided into two subpopulations, one for the response group and one for the nonresponse group, whose sizes are N_1 and N_2. After the fieldwork, all the sample data available for the estimation come only from the first group, thus covering only the response cases. Let the estimator for the total T be $\hat{t}_{HT(r)} = N \times \overline{y}_{(r)}$, where the mean of the respondent data is $\overline{y}_{(r)}$. Because all the respondents are from group 1, the expectation of the respondent mean $\overline{y}_{(r)}$ equals, say $\overline{Y}_1$, the population

mean of that group. If the population group means are unequal or $\overline{Y}_1 \neq \overline{Y}_2$, then the estimator $\hat{t}_{HT(r)}$ is a biased estimator for the population total T, since

$$\text{BIAS}(\hat{t}_{HT(r)}) = E(\hat{t}_{HT(r)}) - T = N\overline{Y}_1 - (N_1\overline{Y}_1 + N_2\overline{Y}_2) = N_2(\overline{Y}_1 - \overline{Y}_2) \qquad (4.1)$$

In practice, it is difficult to evaluate this bias. Although the subpopulation size N_2 could be roughly estimated, the subpopulation mean $\overline{Y}_2$ remains totally unknown. Moreover, the mean squared error (MSE) should be examined instead of the variance, where the MSE for an estimator $\hat{t}_{HT(r)}$ of the total can be written as

$$\text{MSE}(\hat{t}_{HT(r)}) = V_{p(s)}(\hat{t}_{HT(r)}) + \text{BIAS}^2(\hat{t}_{HT(r)}). \qquad (4.2)$$

A further inconvenience is that the variance of the estimated total will be underestimated. The bias due to unit nonresponse is illustrated in the following example.

Example 4.1

Unit nonresponse bias in the *Province'91* population. Let us assume that the southern municipalities were not able to complete the records for the unemployed in time. These municipalities are Kuhmoinen, Joutsa, Luhanka, Leivonmäki and Toivakka. The population of municipalities can thus be divided into two subpopulations, the group of the respondents ($N_1 = 27$) and the group of the nonrespondents ($N_2 = 5$), whose group totals, sizes and means are as follows:

$T_1 = 14\,475$	$N_1 = 27$ (group of respondents)	$\overline{Y}_1 = 536.11$
$T_2 = 623$	$N_2 = 5$ (group of nonrespondents)	$\overline{Y}_2 = 124.60$
$T = 15\,098$	$N = 32$ (whole province)	$\overline{Y} = 471.81$

When drawing the sample by SRSWOR, the selected sample would include both the response and the nonresponse municipalities. Thus, the expected value of the total estimator, based on the response group sample total $\hat{t}_{HT(r)}$, will be $E(\hat{t}_{HT(r)}) = N \times \overline{Y}_1 = 32 \times 536.11 = 17\,156$. If this estimator is taken as the estimator of the population total, a biased estimate results, where the bias due to the unit nonresponse is

$$\text{BIAS}(\hat{t}_{HT(r)}) = E(\hat{t}_{HT(r)}) - T = N_2(\overline{Y}_1 - \overline{Y}_2) = 5 \times (536.11 - 124.60) = 2058,$$

and is noticeably large.

Framework for Handling Nonresponse

In the first part of this book, we examined randomness generated by a sampling design $p(s)$. In the case of nonresponse, we meet another source of

randomness, which is generated by an unknown *response mechanism*, which creates the unknown conditional probability that response set $s_{(r)}$ is realized, given the sample s of size n under the sampling design $p(s)$. This motivates us to consider that in the presence of nonresponse, the point estimators may differ from that of the full design-based estimators and the corresponding variances of estimators include two components: the first component is due to the sampling design used and the second is due to the unknown response mechanism. Taking this dualism as a framework for handling nonresponse presupposes that we guess or model the unknown response probability. This view is clearly presented in the technical report by Lundström and Särndal (2002).

The two main methods for adjustment for nonresponse are reweighting and imputation. The adjustment for unit response can be done by *reweighting*. The sampling weights $w_k = 1/\pi_k$ are adjusted by the inverses $1/\hat{\theta}_k$ of estimated response probabilities $\hat{\theta}_k$, providing new analysis weights or reweight $w_k^* = 1/(\pi_k \hat{\theta}_k)$. Reweighting methods for unit nonresponse are commonly used, for example, by national statistical agencies. In Section 4.1, reweighting techniques will be discussed.

Imputation for item nonresponse means that a missing value of a measurement y_k is filled in by a predicted value $\hat{y}_k$. The goal of imputation is to achieve a complete data matrix for further analysis. Imputation can be performed under single or multiple imputation methods. Little and Rubin (1987) consider the main lines of multiple imputation techniques in theory and practice. Section 4.2 focuses on different imputation methods.

4.1 REWEIGHTING

Unit nonresponse refers to the situation in which data are not available within the survey data set for a number of sampling units. Reweighting can then be used and applied to the observations from the respondents, with the auxiliary information available for both the respondents and the nonrespondents. As a simple example, consider the estimation of a population total. The values obtained from the respondents can be multiplied by an expansion or raising factor to produce a data set, which better agrees with the initial or intended sample size. A simple expansion factor is the inverse of the response rate. For example, if the overall response rate in a survey is 71%, a suitable raising factor would be $1/0.71 = 1.41$. In this nonresponse model, it is assumed that each population element has the same probability θ of responding if selected in the sample, i.e. $\theta_k = \theta$ for all the population elements $k = 1, \ldots, N$, and θ is estimated by $\hat{\theta} = n_{(r)}/n$. Under this rather naive assumption of a nonresponse mechanism, a *reweighted Horvitz–Thompson (HT)*

estimator based on the constant response probability assumption for the population total would be

$$\hat{t}^*_{HT} = \sum_{k=1}^{n_{(r)}} w^*_{HT,k} \times y_k = \frac{1}{\hat{\theta}} \sum_{k=1}^{n_{(r)}} w_k \times y_k = \frac{n}{n_{(r)}} \sum_{k=1}^{n_{(r)}} w_k \times y_k, \tag{4.3}$$

where y_k is observed value for the respondent k of the study variable y, $w^*_{HT,k} = (1/\hat{\theta}) \times w_k$ is the analysis weight, and the subscript '(r)' refers to the respondents; so, $n_{(r)}$ denotes the number of respondents in the sample.

Although these kinds of expansion factors are sometimes used in practice, better estimation can be attained by modelling the response probability. A commonly used model is to divide the population into *response homogeneity groups*, denoted as RHG. These groups are denoted by $1, \ldots, c, \ldots, C$. The group sample sizes and the numbers of respondents in each group are denoted correspondingly by $n_1, \ldots, n_c, \ldots, n_C$ and $n_{1(r)}, \ldots, n_{c(r)}, \ldots, n_{C(r)}$. The homogeneity of RHGs means that all the elements in a group c are assumed to have the same response probability θ_c, which is estimated by the group response rate $\hat{\theta}_c = n_{c(r)}/n_c$. Between the RHGs, however, the response probabilities can vary. And in the reweighting, the inverses of the estimated response probabilities, i.e. estimated group response rates $\hat{\theta}_c$, can be used, giving an analysis weight $w^*_{rhg,k} = (1/\hat{\theta}_c) \times w_k$. Hence, the *reweighted HT estimator based on the RHG method* is

$$\hat{t}^*_{rhg} = \sum_{k=1}^{n_{(r)}} w^*_{rhg,k} \times y_k = \sum_{c=1}^{C} \left(\frac{1}{\hat{\theta}_c}\right) \sum_{k=1}^{n_{c(r)}} w_{ck} \times y_{ck} = \sum_{c=1}^{C} \frac{n_c}{n_{c(r)}} \sum_{k=1}^{n_{c(r)}} w_{ck} \times y_{ck}, \tag{4.4}$$

where w_{ck} and y_{ck} are the sampling weight and the value of y for responding unit k in group c, respectively.

This adjustment for the unit nonresponse can be more powerful than the previous one, because the response probabilities are modelled by using, more efficiently, the information about the structure of the nonresponse. If the value z_k of an auxiliary variable z is known for every sample unit and z correlates with the study variable y, one can try to apply a *reweighted ratio estimator*, whose weights are $w^*_{rat,k} = [(1/\hat{\theta}) \times (\overline{z}/\overline{z}_{(r)})] \times w_k$, where $\overline{z}$ is the mean of an auxiliary variable z calculated from all sampled units and $\overline{z}_{(r)}$ is that calculated from responding units, and $\hat{\theta} = n_{(r)}/n$. Correspondingly, the *reweighted HT estimator based on the ratio model* is

$$\hat{t}^*_{rat} = \sum_{k=1}^{n_{(r)}} w^*_{rat,k} \times y_k = \frac{\overline{z}}{\hat{\theta} \times \overline{z}_{(r)}} \sum_{k=1}^{n_{(r)}} w_k \times y_k = \frac{n \times \overline{z}}{n_{(r)} \times \overline{z}_{(r)}} \sum_{k=1}^{n_{(r)}} w_k \times y_k. \tag{4.5}$$

Next we turn to variance estimation of a reweighted HT estimator of a total. In the context of design-based inference, the sample weights are known constants

$w_k = \pi_k^{-1}$. In reweighting, these constants are multiplied by a sample-dependent weighting factor specific for each reweighting method. This causes an additional variance component to be measured and included into a design-based variance of the estimator of a total. We denote this component as V_{rew}, where 'rew' refers to reweighting. The variance component V_{rew} can be estimated under the above-defined framework for handling unit nonresponse. For this, we conceptually decompose the sample selection procedure into two phases: the selection of the sample s according to a sampling design $p(s)$ and the realization of the set $s_{(r)}$ of the respondents from the selected sample s. This sampling scheme gives an opportunity to estimate separately the variance components of the first and second phases. The first component, denoted as V_{sam}, represents the variance due to the sampling design and the second component, denoted as V_{rew}, represents that due to the unknown *response mechanism*. If assuming, as in Särndal (1996), that these two components are independent, the variance of a reweighted HT estimator $\hat{t}^*_{HT}$ for a total T can be decomposed as

$$V(\hat{t}^*_{HT}) = V_{sam}(\hat{t}^*_{HT}) + V_{rew}(\hat{t}^*_{HT}), \tag{4.6}$$

where $V_{sam}(\hat{t}^*_{HT})$ is the design variance of the basic HT estimator $\hat{t}_{HT}$, defined for respondent data, and $V_{rew}(\hat{t}^*_{HT})$ is the variance component due to the reweighting method used. In Example 4.2, three reweighted estimators $\hat{t}^*_{HT}$, $\hat{t}^*_{rhg}$ and $\hat{t}^*_{rat}$ and their variance components will be calculated.

Example 4.2

Adjustment for unit nonresponse by reweighting for an SRSWOR sample drawn from the *Province'91* population. The data set is presented in Table 4.2. Let us assume two unit nonresponse cases, namely, Kuhmoinen and Toivakka. Note that the value of the auxiliary variable HOU85 is available for the nonresponse cases also. The initial sample size is eight municipalities. Thus, the estimated response rate $\hat{\theta} = n_{(r)}/n = 6/8 = 0.75$. In addition, three of the sampled municipalities are towns (response homogeneity group $c = 1$) and the other five are rural municipalities (response homogeneity group $c = 2$). Because all the towns responded, estimated response probabilities are correspondingly $\hat{\theta}_1 = 3/3 = 1.00$ for the first group and $\hat{\theta}_2 = 3/5 = .60$ for the second group. The mean of the auxiliary variable HOU85 calculated for the total sample ($n = 8$) is $\overline{z} = 5154.75$. The mean of HOU85 calculated for the respondent data set ($n = 6$) is $\overline{z}_{(r)} = 6490.17$. Furnished with this background information, we are ready to calculate the three previously introduced reweighted HT estimators $\hat{t}^*_{HT}$, $\hat{t}^*_{rhg}$ and $\hat{t}^*_{rat}$ for the total T of the variable UE91.

For calculating the reweights, we should first define the appropriate response homogeneity groups. In this case, a natural group for estimators $\hat{t}^*_{HT}$ and $\hat{t}^*_{rat}$ is the total sample, and for the estimator $\hat{t}^*_{rhg}$, two response homogeneity groups are created according to urbanicity. For the estimator $\hat{t}^*_{HT}$, we adopt a naive

Table 4.2 A simple random sample from the *Province'91* population including two nonresponse cases, constructed response homogeneity groups and weights for adjustment for unit nonresponse.

Sample design identifiers			Element	Response data		Response homogeneity	Reweight by nonresponse model		
							REW_HT	RHG	RATIO
STR	CLU	WGHT	LABEL	UE91	HOU85	group (RHG)	$w^*_{HT,k}$	$w^*_{rhg,k}$	$w^*_{rat,k}$
1	18	4	Kuhmoinen	••	1463	2	••	••	••
1	30	4	Toivakka	••	834	2	••	••	••
1	26	4	Pihtipudas	331	1946	2	5.3333	6.6667	4.2359
1	31	4	Uurainen	219	932	2	5.3333	6.6667	4.2359
1	15	4	Konginkangas	142	556	2	5.3333	6.6667	4.2359
1	1	4	Jyväskylä	4123	26 881	1	5.3333	4.0000	4.2359
1	4	4	Keuruu	760	4896	1	5.3333	4.0000	4.2359
1	5	4	Saarijärvi	721	3730	1	5.3333	4.0000	4.2359

A missing value is denoted as '••'.

reweighting method; the reweight is $w^*_{HT,k} = (1/\hat{\theta}) \times w_k = (1/0.75) \times 4 = 5.3333$ for the respondents. For the estimator $\hat{t}^*_{rhg}$, the reweight in the first response homogeneity group (towns) is $w^*_{rhg,1} = (1/\hat{\theta}_1) \times w_k = (1/1) \times 4 = 4$. It is equal to the sampling weight because all towns responded. In the second response homogeneity group (rural municipalities), $w^*_{rhg,2} = (1/\hat{\theta}_2) \times w_k = (1/0.60) \times 4 = 6.6667$. In the case of the ratio estimator, the total sample is taken again as the response homogeneity group. We use the same formula as given in the case of calculation of adjusted weights (see ratio estimation in Section 3.3), but this time the population mean (or total) of the auxiliary variable is replaced by that calculated from the sample. Reweights for the respondents are $w^*_{rat,k} = (1/\hat{\theta}) \times (\overline{z}/\overline{z}_{(r)}) \times w_k = [(n \times \overline{z})/(n_{(r)} \times \overline{z}_{(r)})] \times w_k$, and for SRSWOR they have the same value for each respondent. Empirical values are calculated from the selected sample: $w^*_{rat} = [(8 \times 5154.75)/(6 \times 6490.17)] \times 4 = 4.2359$ for responding units.

Using the calculated reweights, the point estimates and their variance estimates can be calculated. Point estimates for the total T of UE91 are simply reweighted HT estimators calculated from the respondent data set. Estimates are presented in Table 4.3. We focus on the variance estimation because it now includes two components: the variance estimator $\hat{v}_{sam}$ due to the sampling design and the variance estimator $\hat{v}_{rew}$ caused by the response mechanism. We assume that nonresponse is ignorable within each response homogeneity group.

Because the sample design is SRSWOR, we use the appropriate design variance of the total

$$V_{sam}(\hat{t}^*_{HT}) = N^2 \left(1 - \frac{n}{N}\right) \times S^2_{(r)}/n_{(r)} \tag{4.7}$$

where $S^2_{(r)} = \sum_{k=1}^{N_{(r)}} (Y_k - \overline{Y}_{(r)})^2/(N_{(r)} - 1)$ is calculated from the respondent part $U_{(r)}$ of the population U.

The estimated value of this component in this case is

$$\hat{v}_{sam}(\hat{t}^*_{HT}) = N^2 \left(1 - \frac{n}{N}\right) \times \hat{s}^2_{(r)}/n_{(r)} = 32^2 \left(1 - \frac{8}{32}\right) \times 1527.59^2/6 = 14\,967^2,$$

where $\hat{s}^2_{(r)} = \sum_{k=1}^{n_{(r)}} (y_k - \overline{y}_{(r)})^2/(n_{(r)} - 1)$ and is estimated from the respondent data set.

This variance component is the same for each reweighted estimator. The reweighting component of the total variance depends on the reweighting method used. For the reweighting methods, the estimation of V_{rew} is next carried out. Note that the HT estimator $\hat{t}_{HT(r)}$, when calculated from the respondent data set, does not include a variance component because of reweighting.

1. Reweighted estimator $\hat{t}^*_{HT}$. In the case of the first reweighted HT estimator, the variance component $V_{rew}(t^*_{HT})$ is

$$V_{rew}(\hat{t}^*_{HT}) = N^2 \left(1 - \frac{n_{(r)}}{n}\right) \times S^2_{(r)}/n_{(r)}, \tag{4.8}$$

   where $S^2_{(r)} = \sum_{k=1}^{N_{(r)}} (Y_k - \overline{Y}_{(r)})^2/(N_{(r)} - 1)$ is calculated from the respondent part $U_{(r)}$ of the population U. The estimated value of this component is

$$\begin{aligned}\hat{v}_{rew}(\hat{t}^*_{HT}) &= N^2 \left(1 - \frac{n_{(r)}}{n}\right) \times \hat{s}^2_{(r)}/n_{(r)} \\ &= 32^2 \left(1 - \frac{6}{8}\right) \times 1527.59^2/6 = 9978.18^2,\end{aligned}$$

   where $\hat{s}^2_{(r)}$ is estimated from the respondent data.
2. Response homogeneity group estimator $\hat{t}^*_{rhg}$. We have two RHGs whose sample sizes are $n_1 = 3$ and $n_2 = 5$. From these figures, one can estimate the sizes of the corresponding subpopulations, which are as follows:

   $\hat{N}_1 = (n_1/n) \times N = (3/8) \times 32 = 12$ for the first subpopulation and

   $\hat{N}_2 = (n_2/n) \times N = (5/8) \times 32 = 20$ for the second.

   The reweight component of variance for the response homogeneity group estimator $\hat{t}^*_{rhg}$ is

$$V_{rew}(\hat{t}^*_{rhg}) = \sum_{c=1}^{C} \hat{N}_c^2 \left(1 - \frac{n_{c(r)}}{n_c}\right) \times S^2_{c(r)}/n_{c(r)}, \tag{4.9}$$

where $S^2_{c(r)} = \sum_{k=1}^{N_{c(r)}} (Y_{ck} - \overline{Y}_{c(r)})^2/(N_{c(r)} - 1)$ is calculated separately for each response homogeneity group. The number of responding units in the subpopulation U_c, where $c = 1, 2$, is denoted as $N_{c(r)}$. The corresponding estimate $\hat{v}_{rew}(\hat{t}_{rhg})$ is calculated from (4.9) by substituting each variance $S^2_{c(r)}$ by its estimated value $\hat{s}^2_{c(r)}$ calculated from the respondent data set. Thus, we get

$$\begin{aligned}\hat{v}_{rew}(\hat{t}^*_{rhg}) &= 12^2\left(1 - \frac{3}{3}\right) \times 1952.99^2/3 + 20^2\left(1 - \frac{3}{5}\right) \times 95.04^2/3 \\ &= 0 + 694.07^2 \\ &= 694.07^2.\end{aligned}$$

3. Reweighted ratio estimator $\hat{t}^*_{rat}$. First, we derive a variable of residuals $E_{k(r)} = Y_{k(r)} - (\overline{Y}_{(r)}/\overline{Z}_{(r)}) \times Z_{k(r)}$. Note that the residuals are calculated from the responding part of the population. The reweight component of the variance of the estimator $\hat{t}^*_{rat}$ is given by

$$V_{rew}(\hat{t}^*_{rat}) = N^2\left(1 - \frac{n_{(r)}}{n}\right) \times S^2_{E_{(r)}}/n_{(r)}, \tag{4.10}$$

where $S^2_{E_{(r)}} = \sum_{k=1}^{N_{(r)}} (E_{k(r)} - \overline{E})^2/(N_{(r)} - 1)$ and $\overline{E} = \sum_{k=1}^{N_{(r)}} E_{k(r)}/N_{(r)}$.

The residuals $E_{k(r)}$ are estimated from the respondent data set as $\hat{e}_{k(r)} = y_{k(r)} - (\overline{y}_{(r)}/\overline{z}_{(r)}) \times z_{k(r)}$.

In this particular case, the reweight component of variance $\hat{v}_{rew}(\hat{t}^*_{rat})$ is

$$\hat{v}_{rew}(\hat{t}^*_{rat}) = N^2\left(1 - \frac{n_{(r)}}{n}\right)\hat{s}^2_{\hat{e}_{(r)}}/n_{(r)} = 32^2\left(1 - \frac{6}{8}\right) \times 120.29^2/6 = 785.73^2,$$

where $\hat{s}^2_{\hat{e}_{(r)}} = \sum_{k=1}^{n_{(r)}} (\hat{e}_{k(r)} - \overline{\hat{e}}_{(r)})^2/(n_{(r)} - 1)$ is calculated from the respondent data set.

The sampling rates are defined as the number of respondents in the sample divided by the estimated or actual size of the response homogeneity group in the target population. Estimators $\hat{t}^*_{HT}$ and $\hat{t}^*_{rat}$ have the whole sample as the response homogeneity group. Thus, the sampling rate is $n_{(r)}/N = 6/32 = 0.1875$ for both. For the estimator $\hat{t}^*_{rhg}$, the sampling rate in the first response homogeneity group is $n_{1(r)}/\hat{N}_1 = 3/12 = 0.25$ and that in the second response homogeneity group is $n_{2(r)}/\hat{N}_2 = 3/20 = 0.15$.

Table 4.3 Estimates of the total and components of its variance estimate under various reweighting methods; a simple random sample from *Province'91* population presented in Table 4.2.

Method and estimator	Estimate for the total	$\hat{v}(\hat{t})$	$\hat{v}_{sam}$	$\hat{v}_{rew}$
Respondent data ($n_{(r)} = 6$) $\hat{t}_{HT(r)}$	33 579	$17\,988^2$	$17\,988^2$	0
Reweighted estimator $\hat{t}^*_{HT}$	33 579	$17\,988^2$	$14\,967^2$	9978^2
Response homogeneity group $\hat{t}^*_{rhg}$	27 029	$14\,983^2$	$14\,967^2$	694^2
Ratio estimator $\hat{t}^*_{rat}$	26 669	$14\,988^2$	$14\,967^2$	786^2
'Full response' ($n = 8$) $\hat{t}_{HT}$	26 440	$13\,282^2$	$13\,282^2$	0

Results are summarized in Table 4.3. In addition to the reweighted estimators, two reference estimators are included. The estimator $\hat{t}_{HT(r)} = N \times \overline{y}_{(r)}$ is calculated directly from the respondent data. In this case, the sampling rate is $n_{(r)}/N = 6/32 = 0.1875$. For a fair comparison, the basic design-based estimator $\hat{t}_{HT}$ for a total is calculated from the figures presented in the last row headed as 'Full response'. The sampling rate is, in this case, $n/N = 8/32 = 0.25$. The variance component $\hat{v}_{sam}$ is estimated separately for the respondent data ($n = n_{(r)} = 6$) and the 'Full response' ($n = 8$), respectively, by (4.7). The last column headed as $\hat{v}_{rew}$ shows that for the respondent data (first row) and 'Full response' (bottom row) there is no variance component due to reweighting.

A desired property of a reweighted estimator is that it reproduces, as closely as possible, the value of the full response estimator. In this sense, both the response data estimator $\hat{t}_{HT(r)}$ and the reweighted HT estimator $\hat{t}^*_{HT}$ give poor results. The point estimate $\hat{t}_{HT(r)} = \hat{t}^*_{HT} = 33\,579$ is very far from that of the 'Full response' estimator $\hat{t}_{HT} = 26\,440$. The same holds for variance estimates $\hat{v}(\hat{t}_{HT(r)}) = \hat{v}(\hat{t}^*_{HT}) = 17\,988^2 > 13\,282^2$. The reason for the reweighted HT estimator to produce poor results is that a simple response mechanism was assumed involving a constant response probability $\hat{\theta}$ for all population elements. The response homogeneity group estimator $\hat{t}^*_{rhg}$ and the ratio estimator $\hat{t}^*_{rat}$ use auxiliary information gathered from the sample data set. The use of these estimators is based on more appropriate model assumptions, and if the assumptions hold closely, as seems to be the case, these two estimators reproduce closely the 'Full response' estimate.

4.2 IMPUTATION

Item nonresponse means that in the data set to be analysed some values are present for a sample element, but at least for one item a value is missing for that element.

When using this kind of data matrix with some computer programs for survey estimation, each observation with a missing value for any of the variables included in the analysis is excluded. Moreover, in some programs, a complete data matrix is required. This leads to loss of information for the other variables for which data are not missing. Therefore, efforts are often made to get a more complete data set. To attain this goal, different imputation techniques have been devised.

Imputation implies simply that a missing value of the study variable y for a sample element k in the data matrix is substituted by an imputed value $\hat{y}_k$. For example, in some computer packages, a technique called *mean imputation* is available, in which an overall respondent mean $\overline{y}_{(r)}$, calculated from the respondent values of the study variable, is inserted in place of the missing values for that variable. Then the imputed value for element k is $\hat{y}_k = \overline{y}_{(r)}$. However, there are certain disadvantages in this method, as will be demonstrated in Example 4.3. In more advanced methods, auxiliary information available from the frame population or from the original sample is utilized to model the missing values more realistically.

Mean imputation does not use any auxiliary information possibly available in the sample data set. Here, as before, an auxiliary variable z, which is correlated with the study variable y and whose values are known for all sampled units, could be used in an imputation method. For example, we could use the sample values of the auxiliary variable z to create distances $|z_l - z_k|$ between two sampling units where $l \neq k$. The sample element for which the distance reaches the minimum is called a *nearest neighbour*. If the element k belongs to the group of nonrespondents and the element l to the group of respondents, we substitute the value y_l for the missing value of element k, providing an imputed value $\hat{y}_k = y_l$. Thus, the sample element l is a *donor* for the element k. Note that this estimate is a real measurement actually observed. And *ratio estimation* can be applied here as in the context of reweighting. Now, we predict an individual value for each missing value through the equation $\hat{y}_k = z_k \times (\overline{y}_{(r)}/\overline{z}_{(r)})$, where $\overline{y}_{(r)}$ and $\overline{z}_{(r)}$ are respectively the respondent means of the study and auxiliary variables. For example, the incomplete data set can be imputed using *hot-deck imputation.* In hot-deck imputation, a measurement value is selected randomly from the response data and is applied for the missing value. All these methods belong to *single imputation* methods.

A single missing value may be replaced by two or more imputed values, as in the method of *multiple imputation.* This is done independently for each missing value. When we repeat this procedure m times for each missing item, we get m complete data sets, which are ready for statistical analysis. The original weighting system derived from the sample design $p(s)$ can be used. In all imputation methods, we are faced with a similar problem as formerly in reweighting. We should evaluate and add the imputation variance component to the variance formula of an estimator. In the case of single imputation, where one predicted value $\hat{y}_k$ is substituted for a missing value, the formula (4.6) can be used by replacing the component V_{rew} by

a new component V_{imp} due to imputation, giving

$$V(\hat{t}^*) = V_{sam} + V_{imp}. \tag{4.11}$$

In multiple imputation, we predict m values $\hat{y}_1, \ldots, \hat{y}_j, \ldots, \hat{y}_m$ for each missing item. We thus create m 'completed' data sets. In order to combine the results, we define first the multiple imputation estimate of our parameter of interest. For example, for a total, an estimate is

$$\hat{t}^*_{mi} = \frac{1}{m} \times \sum_{j=1}^{m} \hat{t}^*_j, \tag{4.12}$$

where $\hat{t}^*_j$ is an estimate for the total and $\hat{v}_{p(s)}(\hat{t}^*_j)$ is the variance estimate from the jth 'completed' data set, $j = 1, \ldots, m$. The variance estimate of $\hat{t}^*_{mi}$ includes two components, the within-imputation variance component and the between-imputation variance component. The within-imputation variance is calculated as the mean of the m variance estimates $\hat{v}_{p(s)}(\hat{t}^*_j)$, representing the variance V_{sam}. The between-imputation variance component is associated with the variability of $\hat{t}^*_j$. This component is interpreted here as the variance V_{imp} due to imputation. Under multiple imputation, the variance estimate of the total is thus

$$\begin{aligned} \hat{v}(\hat{t}^*_{mi}) &= \hat{v}_{sam} + \hat{v}_{imp} \\ &= \left[\frac{1}{m} \times \sum_{j=1}^{m} \hat{v}_{p(s)}(\hat{t}^*_j) \right] + \left[\left(1 + \frac{1}{m} \right) \times \sum_{j=1}^{m} \frac{(\hat{t}^*_j - \hat{t}^*_{mi})^2}{m-1} \right] \end{aligned} \tag{4.13}$$

where $\hat{v}_{p(s)}(t^*_j)$ is the variance estimate calculated under the sample design $p(s)$ from the jth completed data set and $(1 + (1/m))$ is an adjustment for a finite m.

In practice, m is usually taken to be a small number. $m = 2$ is a minimum but 3 to 5 is preferred. Example 4.3 illustrates the estimation of the variance components for different imputation methods.

Example 4.3

We impute two missing values for the sample selected from the *Province'91* population with SRSWOR. The same sample is used as in Example 4.2, and it includes $n = 8$ municipalities. A missing value is created for the study variable UE91 in two municipalities (Kuhmoinen and Toivakka). Missing values are marked as '••' in the data set displayed in Table 4.4. Variable HOU85 serves as an auxiliary variable having no missing values.

Four imputation techniques will be applied for completing the sample data set. The first is the *respondent mean imputation* method. The mean of the respondents ($n_{(r)} = 6$) is $\overline{y}_{(r)} = 1049.33$. The two missing values are replaced by this overall mean. The second and the third nonresponse models use the variable HOU85 as an auxiliary variable z. The second method is called *nearest neighbour imputation*.

Table 4.4 Completed data sets obtained by single imputation methods (The *Province'91* population).

ID	Element	Response data		Imputed data sets by model			
k	LABEL	UE91	HOU85	Respondent mean	Nearest neighbour	Ratio estimation	Full response
18	Kuhmoinen	••	1463	1049.33*	331*	236.54*	187
30	Toivakka	••	834	1049.33*	219*	134.84*	127
1	Jyväskylä	4123	26 881	4123	4123	4123	4123
4	Keuruu	760	4896	760	760	760	760
5	Saarijärvi	721	3730	721	721	721	721
15	Konginkangas	142	556	142	142	142	142
26	Pihtipudas	331	1946	331	331	331	331
31	Uurainen	219	932	219	219	219	219

Imputed values are flagged with '*' and missing values with '••'.

For nonresponding unit k, we select the value of responding unit l, for which the distance $|z_k - z_l|$ attains the minimum over all potential donors (a potential donor is a sample unit that belongs to the group of respondents for variable y). The minimum is reached when Pihtipudas ($y_{26} = 331$) is the donor for Kuhmoinen ($|1949 - 1463| = 486$) and Uurainen ($y_{31} = 219$) is the donor for Toivakka ($|932 - 834| = 98$). In the third model, we use *ratio estimation*. We calculate the ratio $\hat{B} = \overline{y}_{(r)}/\overline{z}_{(r)} = 1049.33/6490.17 = 0.1617$ from the response data set and then evaluate the predicted values $\hat{y}_k = \hat{B} \times z_k$, which are $\hat{y}_{18} = 0.1617 \times 1463 = 236.57$ for Kuhmoinen and $\hat{y}_{30} = 0.1617 \times 834 = 134.86$ for Toivakka. The sample data set amended with the imputed values is displayed in Table 4.4. Note that in mean imputation and ratio estimation, a predicted value is used for a missing observation. For this reason, these values are with two decimal digits. On the other hand, a nearest neighbour as a donor gives an integer value for imputation. This holds also for multiple imputation, because we have used hot-deck imputation, where every responding unit is a potential donor.

Three complete data sets, one for each single imputation method, are now created for estimation. We use sampling weights, which are here a constant $w_k = 4$, because our sampling design is an SRSWOR design. However, a new aspect is revealed in the variance estimator, which now includes two components (see formula 4.11). In the estimation of a total, an estimator of the sampling variance is

$$\hat{v}_{sam}(\hat{t}^*_{HT}) = N^2 \left(1 - \frac{n}{N}\right) \times \hat{s}^2_{(r)}/n_{(r)} = 32^2 \left(1 - \frac{8}{32}\right) \times 1527.59^2/6 = 14\,967^2$$

where $\hat{s}^2_{n_{(r)}} = \sum_{k=1}^{n_{(r)}} (y_k - \overline{y}_{(r)})^2/(n_{(r)} - 1)$ is computed from the respondent data set. This variance component is the same for each imputation method. The

imputation variance component V_{imp} for all single imputation methods is estimated by

$$\hat{v}_{imp}(\hat{t}^*_{HT}) = N^2 \left(1 - \frac{n_{(r)}}{n}\right) \times \frac{\sum_{k=1}^{n_{(r)}} (\hat{e}_k - \overline{\hat{e}})^2}{n_{(r)} - 1} / n_{(r)} \tag{4.14}$$

where $\overline{\hat{e}} = \sum_{k=1}^{n_{(r)}} \hat{e}_k / n_{(r)}$ is the mean of residuals $\hat{e}_k = y_k - \hat{y}_k$. For mean imputation, the residuals are $\hat{e}_k = y_k - \overline{y}_{(r)}$. Using a nearest neighbour as the donor results in residuals $\hat{e}_k = y_k - y_{k(l)}$, where $y_{k(l)}$ is the y-value of the donor l. Ratio estimation results in $\hat{e}_k = y_k - (\overline{y}_{(r)}/\overline{z}_{(r)}) \times z_k$. Incorporating these variables in 4.14, we get the estimated imputation variance components as follows:

$$\hat{v}_{imp}(\hat{t}^*_{rm}) = 32^2 \left(1 - \frac{6}{8}\right) \times 1527.59^2/6 = 9978.18^2$$

$$\hat{v}_{imp}(\hat{t}^*_{nn}) = 32^2 \left(1 - \frac{6}{8}\right) \times 1365.62^2/6 = 8920.20^2$$

$$\hat{v}_{imp}(\hat{t}^*_{ra}) = 32^2 \left(1 - \frac{6}{8}\right) \times 120.29^2/6 = 785.73^2.$$

Note that the smallest variance due to imputation is for the ratio model.

Next, we turn to the *multiple imputation* method. For this simple exercise, we use five independent repetitions of *hot-deck imputation*. Note that hot-deck imputation is used here just to illustrate the basic principles of multiple imputation for this quite restricted small-scale data set. For practical purposes, much more sophisticated multiple imputation techniques have been developed and computerized. There is much literature on the alternative techniques; the reader is advised to consult the book by Schafer (2000) and the paper by Rubin (1996) for further details.

For each run, the missing responses are replaced by values selected randomly from the respondent data set. This procedure results here in five complete data sets, which are presented in Table 4.5. A point estimate $\hat{t}^*_{mi}$ of the total of unemployed persons is here the mean value of the five individual datawise estimates $\hat{t}^*_j$ of the same total. Thus, we get from (4.12)

$$\hat{t}^*_{mi} = (1/5)(28\,792 + 31\,108 + 28\,944 + 44\,716 + 29\,100) = 32\,532.$$

By (4.13), the variance of the estimator $\hat{t}^*_{mi}$ is decomposed to within-imputation variability and between-imputation variability. The elements of within-imputation variation are the five datawise estimates of the design variance estimates of the estimator $\hat{t}^*_j$. Thus, the first term of (4.13) is

$$\begin{aligned}\hat{v}_{sam} = \frac{1}{m} \times \sum_{j=1}^{m} \hat{v}_{p(s)}(\hat{t}^*_j) &= \frac{1}{5} \times \left(1 - \frac{8}{32}\right) \times 32^2 \\ &\times (1330.715^2 + 1298.982^2 + 1325.416^2 \\ &+ 1699.989^2 + 1324.716^2)/8 = 13758.87^2\end{aligned}$$

Table 4.5 Imputed data sets obtained by multiple imputation ($m = 5$). Hot-deck imputation is used for each completed data set (the *Province'91* population).

ID	Element	Response data UE91	Repeated samples including imputed values and flagged as '*' 1	2	3	4	5	Full response
18	Kuhmoinen	••	760*	760*	721*	4123*	760*	187
30	Toivakka	••	142*	721*	219*	760*	219*	127
1	Jyväskylä	4123	4123	4123	4123	4123	4123	4123
4	Keuruu	760	760	760	760	760	760	760
5	Saarijärvi	721	721	721	721	721	721	721
15	Konginkangas	142	142	142	142	142	142	142
26	Pihtipudas	331	331	331	331	331	331	331
31	Uurainen	219	219	219	219	219	219	219
	Mean	1049.33	899.75	972.12	904.50	1397.38	909.37	826.25
	STD (y)	1527.59	1330.71	1298.98	1325.42	1699.99	1324.72	1355.15

Imputed values are flagged with '*' and missing values with '••'.

where $\hat{v}_{p(s)}(\hat{t}_j^*) = \hat{v}_{srswor}(\hat{t}_j^*)$ or a variance estimator for a total when the sampling design is SRSWOR.

The corresponding between-variability or imputation variance is estimated by

$$\hat{v}_{imp} = \left(1 + \frac{1}{m}\right) \times \sum_{j=1}^{m} \frac{(\hat{t}_j^* - \hat{t}_{mi}^*)^2}{m-1}$$
$$= 1.2 \times 6876.444^2 = 7532.39^2$$

Summing up these two components gives a variance estimate for the estimator $\hat{t}_{mi}^*$ as

$$\hat{v}(\hat{t}_{mi}^*) = \hat{v}_{sam} + \hat{v}_{imp} = 13758.87^2 + 7532.39^2 = 15686.86^2.$$

Results from all imputation methods are summarized in Table 4.6. Again, for a comparison, the bottom row represents the estimate and its variance estimate in the case of 'full response'. This row serves as the reference. If an imputation method works well, it should produce a value close to the 'full response' estimate. This is expected to happen for the point estimate (but not for the variance estimator because it includes an additional imputation variance term).

Respondent mean imputation gives the same total estimate as the 'no adjustment' method (33 579) but leads to underestimation of the variance unless the imputation variance ($\hat{v}_{imp} = 9978^2$) is added. The more advanced nonresponse models, 'nearest neighbour' and 'ratio estimation', result in estimates that are closer to the reference value calculated from the data set of 'full response'. For the

Table 4.6 Estimates of a total and its standard error under various imputation methods (the *Province'91* population).

Model type	Estimator	Estimate for a total	$\hat{v}(\hat{t}_*)$	$\hat{v}_{sam}$	$\hat{v}_{imp}$
No adjustment ($n_{(r)} = 6$)	$\hat{t}_{HT(r)}$	33 579	$17\,988^2$	$17\,988^2$	0
Respondent mean	$\hat{t}^*_{ma}$	33 579	$17\,988^2$	$14\,967^2$	9978^2
Multiple imputation ($m = 5$)	$\hat{t}^*_{mi}$	32 532	$15\,686^2$	$13\,759^2$	7532^2
Nearest neighbour	$\hat{t}^*_{nn}$	27 384	$17\,424^2$	$14\,967^2$	8920^2
Ratio estimation	$\hat{t}^*_{ra}$	26 669	$14\,988^2$	$14\,967^2$	786^2
Full response ($n = 8$)	$\hat{t}_{HT}$	26 440	$13\,282^2$	$13\,282^2$	0

'nearest neighbour method', we obtain $\hat{t}^*_{nn} = 27\,384$. Using auxiliary information by the ratio model, the calculated estimate is $\hat{t}^*_{ra} = 26\,669$, which is close to the reference value $\hat{t}_{HT} = 26\,640$. Though, a penalty due to imputation causes only moderate variance increase: the imputation variance is $\hat{v}_{imp} = 786^2$.

Multiple imputation behaves differently. The point estimate $\hat{t}^*_{mi} = 32\,532$ is clearly greater than that of the 'full response'. On the other hand, the total sample variance calculated according to the formula (4.13) is $\hat{v}(\hat{t}^*_{mi}) = 15\,686^2$and is thus smaller than that of nearest neighbour imputation and respondent mean imputation.

Imputation has two different impacts. Firstly, a substitute value can be imputed for a missing value and, secondly, imputation has an effect on the standard error of the estimator that we are interested in. An obvious gain from imputation is that the analyst has a complete data matrix for analysis, but if the imputation model gives biased values, the results of analysis may be misleading. All depends on how successfully the imputation model catches the nonresponse. If nonresponse is ignorable within a response homogeneity group, then the respondent mean is an unbiased estimate for response homogeneity group all the elements belonging to this group including the missing values. But, because imputed values are also estimates, they have their own variance component that is to be added to the variance of the basic estimator.

4.3 CHAPTER SUMMARY AND FURTHER READING

The aim of this section is to give a broader perspective on survey production as could be achieved by considering only design-based estimators and their sampling errors generated by the randomness due to probability sampling. We discuss on nonsampling errors by considering briefly different sources of survey errors that are components of the *total survey error.* The concept of total survey error is difficult to define and even more difficult to measure in practice. One reason is that the different survey errors are not independent of each other. However, for practical purposes it is reasonable to consider different types of survey errors

separately and to search for strategies to reduce them one by one. Then, the total survey error can be expected to decrease.

Nonresponse is present in large-scale sample surveys, causing an incomplete data set. Because most computer packages for data analysis presuppose complete data, as a first step after the data-collection phase, the data are cleaned and adjusted for those that are missing. Nonresponse involves missing data in the form of unit nonresponse or item nonresponse, which can cause biased estimation and erroneous standard error estimates. Effective operations are important during the data-collection phase to reduce the nonresponse.

Nonresponse can be adjusted for by various techniques. We introduced a practical way to perform an adjustment by modelling the nonresponse by using auxiliary information available in the sampled data set. Alternatively, auxiliary data can be extracted from a census or business register. The difference between these methods depends upon the extent to which auxiliary information is utilized. Nonresponse in social surveys is discussed, for example, in Groves *et al.* (2001) and that in business surveys in Dillman (1999).

Coverage errors, processing errors and measurement errors and often met in the context of large-scale surveys and are discussed, for example, in a policy paper published by the U.S. Federal Committee on Statistical Methodology (2001). In the following text, definitions given in that paper are used.

Coverage error is an error associated with the failure to include some target population elements in the frame used for sample selection (undercoverage) and the error associated with the failure to exclude units, which do not belong to the target population (overcoverage). The source of coverage error is the sampling frame itself. It is important, therefore, that information about the quality of the sampling frame, and its completeness for the target population, is assessed. Measurement methods for coverage error rely on methods external to the survey operations: for example, comparing survey estimates to independent sources or implementing a case-by-case matching of two registers. Coverage errors do not leave any apparent indication of their existence; they can be measured only by a reference to an outside source. Often-used methods are aggregate comparison to another source and case-by-case matching. It is possible to compare the distribution of age, sex and other population characteristics in a study population with that of a census register. A second approach for measuring coverage error is based on case-by-case matching. This method presupposes that an alternative list of population units exists or can be constructed using the census/survey/record system. The population not on either list is, of course, not observable. However, it can be estimated when two lists are, approximately, independent. An often-used measure in CATI is the number of identified phone numbers of sample units divided by the nominal sample size. An example is given in Section 9.1.

Processing error can occur after the survey data are collected, usually during the process of converting collected data to consistent machine-readable form for statistical analysis. Processing errors include data entry, coding and editing errors, thus inducing damaged data records. Error rates are determined through quality

control samples; however, in recent years authors have advocated continuous management practices. For example, editing is the procedure for detecting and adjusting individual errors in data records resulting from data collection. Edit rules or, simply, edits are used for identifying missing or erroneous, or suspicious values. Generally, this procedure is performed by computer-assisted methods. For example, Cox *et al.* (1995) and Couper *et al.* (1998) devote many chapters to the methods for detecting and handling processing errors in business and social surveys. Producers of official statistics such as national statistical agencies have developed automated procedures to monitor and adjust for processing errors.

Measurement error is characterized as the difference between the observed value of a variable and the true but unobserved value of that variable. Measurement error comes from four primary sources in survey data collection: the questionnaire (as a formal presentation or request for information), the effect the interviewer has on the response to a question (interviewer effect), the data-collection mode and the respondent (as the recipient of the request for information). These sources comprise the entity of data collection, and each source can introduce error into the measurement process. For example, measurement error may occur in respondents' answers to survey questions, including misunderstanding the meaning of the question, failing to recall the information accurately and failing to construct the response correctly (e.g. by summing the components of an amount correctly). Measurement errors are difficult to quantify, usually requiring special, expensive studies. Re-interview programs, record check studies, behaviour coding, cognitive testing and randomized experiments are a few of the approaches used to quantify measurement error. An example of measurement error is the interviewer effect, which has been generally measured by the intra-class correlation coefficient ρ_{int} introduced in Section 2.3. For example, the book by Biemer *et al.* (1991) addresses, with strong empirical background, the measurement error both in business and in social surveys.

Total survey quality is an interesting concept in this context. It refers to a multidimensional characteristic covering sampling and different nonsampling components of survey error. Groves (1989) discusses this concept from a different point of view and presents an interesting dualism that survey errors can be classified into observational errors and errors of nonobservation. In addition, he analysed the effect of different types of errors separately and how they influence the bias and variance of estimators. Several examples of survey practice conducted led to the conclusion that the survey quality indeed is a multidimensional property and its different components could be inter-correlated. Then the quality profile, including a set of well-defined indicators, of a social or business survey can be constructed and communicated rather than a single figure of total survey error. This idea is strongly supported, for example, in a paper of Platek and Särndal (2001).

5

Linearization and Sample Reuse in Variance Estimation

In this chapter and in Chapters 7 and 8, we discuss estimation, testing and modelling methods for complex analytical surveys common, for example, in social, health and educational sciences. In analytical surveys, variance estimation is needed to obtain standard error estimates of sample means and proportions for the total population and, more importantly, for various subpopulations. In modelling procedures, variance estimates of estimated model coefficients, such as regression coefficients, are needed for proper test statistics. Subpopulation means and proportions are defined as ratio estimators in Section 5.2. Approximation techniques are required for the estimation of the variances of these nonlinear estimators. These techniques supplement those examined for descriptive surveys in Chapters 2 and 3. The linearization method, considered in Section 5.3, is used as the basic approximation method. Alternative methods (balanced half-samples, jackknife and bootstrap) based on sample reuse techniques are examined in Section 5.4, and all the methods are compared numerically in Section 5.5. The variance approximation methods are demonstrated for the Mini-Finland Health Survey, providing a complex analytical survey in which stratified cluster sampling is used with regional stratification and two regional sample clusters per stratum. A more complex setting is introduced in Section 5.6, in which the Occupational Health Survey (OHC) data is introduced. The sampling design of the OHC Survey is a combination of stratified one-stage and two-stage sampling with industrial establishments as clusters. These data will be used in extending variance estimation to the estimation of the covariance matrix of several ratio estimators, which are each calculated for a specific population subgroup. Covariance-matrix estimates of such ratio estimators as subpopulation proportions and means are needed, for example, to conduct logit modelling and other types of modelling

Practical Methods for Design and Analysis of Complex Surveys Risto Lehtonen and Erkki Pahkinen
 ISBN: 0-470-84769-7

procedures. The other extension is to consider non-epsem complex designs. This is done by incorporating appropriate element weights in the estimators.

All the approximation methods for variance estimation of a ratio estimator under a complex sampling design, introduced in previous sections, would also be available for the covariance-matrix estimation. We choose the linearization method because of its practical importance. Covariance-matrix estimation using linearization is considered in Section 5.7. There, the concept of the design effect of a ratio estimator is extended to a *design-effects matrix* of a vector of several ratio estimators. The design-effects matrix is also used when assessing the contribution from clustering on a covariance-matrix estimate. The chapter summary is given in Section 5.8.

5.1 THE MINI-FINLAND HEALTH SURVEY

The Mini-Finland Health Survey was designed to obtain a comprehensive picture of health and of the need for care in Finnish adults, and to develop methods for monitoring health in the population. The sampling design of the survey belongs to the class of two-stage stratified cluster sampling. A variety of data collection methods were used; one aim of the survey was to compare the reliability of these various methods (Heliövaara *et al.* 1993). A large part of the data was collected in health examinations using a Mobile Clinic Unit, and by personal interviews. Cluster sampling with regional clusters was thus motivated by cost efficiency.

The target population of the survey was the Finnish population aged 30 years or over. A two-stage stratified cluster-sampling design was used in such a way that one cluster was sampled from each of the 40 geographical strata. The one-cluster-per-stratum design was used to attain a deep stratification of the population of the clusters. The sample of 8000 persons was allocated to achieve an *epsem sample* (equal probability of selection method; see Section 3.2). Recall that an epsem sample refers to a design involving a constant overall element-sampling fraction.

Original Sampling Design

The 320 population clusters in the original sampling design consisted of one municipality or, in some cases, two regionally neighbouring municipalities. The clusters were stratified by whether they were urban or rural and the shares of the population in manufacturing industry and agriculture. From the largest towns, 8 self-representing strata were formed. The other 32 strata consisted of several nearly equal-sized clusters and consisted of 40 000–60 000 eligible inhabitants. One cluster was sampled from these noncertainty strata using PPS sampling with a cumulative method in which the inclusion probabilities were proportional to the size of the target population in a stratum (see Section 2.5). Second-stage sample

sizes were obtained by proportional allocation, resulting in an epsem design. Sample sizes from the sampled clusters varied between 50 and 500 people, the mean being 150. The person-level samples were drawn by systematic sampling in each stratum using a register database as the sampling frame, which covered the relevant population of the sampled clusters.

Modified MFH Survey Sampling Design

The estimation of between-cluster variance was not possible in the noncertainty strata because only one cluster was drawn from each stratum. The original design was thus modified for variance estimation using the so-called *collapsed stratum technique*. A total of 16 pseudo-strata were formed from the 32 noncertainty strata so that there were two clusters in each of the new strata. A pair of strata was formed by combining two of the original strata that were approximately equal-sized and had similar values for the stratification variables. To obtain a manageable design for analysis, which is also useful for our pedagogical purposes, two pseudo-clusters were formed in the eight self-representing strata by randomly dividing the sample into two approximately equal-sized parts in each stratum. Note that, alternatively, one could assume an element-sampling design in the eight certainty strata such that each element constitutes a cluster of its own and, then, the modified overall design would consist of 8 one-stage strata and 16 two-stage strata with 2 sample clusters in each of them. In the modified design, called the *MFH Survey sampling design*, there are 24 strata and 48 sample clusters. The MFH design is described in more detail in Lehtonen and Kuusela (1986).

The relatively small number of sample clusters in the MFH Survey sampling design can cause a problem in the estimation of variances and covariances. The number of clusters determines the degrees of freedom available for variance and covariance estimation. These degrees of freedom are defined as the number of sample clusters less the number of strata, i.e. $48 - 24 = 24$ in the MFH design. This small number can cause instability in variance and covariance estimates, possibly resulting in difficulties in testing and modelling procedures. The situation is different, for example, in the Occupational Health Care Survey and in the Finnish Health Security Survey, where the number of sample clusters is much larger (these surveys will be described in Sections 5.6 and 9.3, respectively).

Data Collection and Nonresponse

The main phases of the field survey were a health interview, a health examination, which consisted of two phases, and an in-depth examination. The field survey was carried out in 1978–1981. The main methods were interviews, questionnaires, tests of performance, physical and biochemical measurements, observer

assessments and a clinical examination by a doctor. The interview was carried out by local public health or hospital nurses, and the health examination was carried out by a Mobile Clinic Unit.

Of the 8000 people in the sample, 7703 (96%) completed the health interview, and 7217 (90%) took part in the screening phase of the health examination. Over 6000 persons of those examined during the screening phase had at least one symptom, or finding, or gave a disease history that led to their being asked to attend the clinical phase of health examination; 94% attended. Almost 5300 of those examined during the screening phase were asked to attend the doctor's clinical examination; 4840 participated. The data for non-attendance were amended after the field study. Thus, clinical data based on a doctor's examination, or data similar to these data, are available for all 5292 persons invited to the doctor's examinations. The response rates are thus very high for each phase of the survey.

Design Effects

The regional clusters in the MFH Survey sampling design had quite large and heterogeneous populations. Because of the type of clusters, only slight intra-cluster correlations can be expected in most study variables. But there are also variables for which clustering effects are noticeable. Design-effect estimates of sample means or proportions of selected study variables are displayed in Table 5.1, which covers data from the screening phase of the health examination. The design-effect estimates vary between 3.2 and 0.9, the largest estimate being for the mean of a continuous variable, systolic blood pressure. The design-effect estimates in many study variables were close to one, and in some cases less than one, indicating a weak clustering effect.

Table 5.1 Design-effect estimates of sample means or proportions of selected study variables in the MFH Survey data set.

Study variable	deff
Systolic blood pressure	3.2
Chronic morbidity	2.0
Number of physician visits	1.4
Body mass index	1.4
Serum cholesterol	1.2
Number of dental visits	1.0
Number of sick days	0.9

Demonstration Data Set

In examining variance approximation techniques for subpopulation means and proportions, we used a subgroup of the MFH Survey data consisting of 30–64-year-old males who took part in the screening phase of the health examination and who also belonged to an active labour force or had a past labour history. These data consist of 2699 eligible males. The data set includes sampling identifiers STRATUM, CLUSTER and WEIGHT; and two binary response variables, CHRON (presence of chronic illness) and PHYS (suffering or having suffered from physical health hazards at work); and a continuous response variable SYSBP (systolic blood pressure). Information on these data is displayed in Table 5.2. Note that the selected subgroup is of a cross-classes type, properly reflecting all essential properties of the MFH Survey sampling design such as the number of strata (24) and the number of sample clusters (48) covered.

Our aim is to estimate the variances of the subpopulation proportion estimator of CHRON and the subpopulation mean estimator of SYSBP by using approximation methods based on linearization and sample reuse. Both response variables indicated relatively strong intra-cluster correlation from the total MFH Survey data. The response variable PHYS is used in a test for two-way tables in Chapter 7. Before turning to these tasks, we briefly discuss the issue of weighting in the relevant MFH Survey subgroup.

Poststratification

The MFH Survey data set can be regarded as self-weighting because the design is epsem and adjustment for nonresponse is not necessary. However, for further

Table 5.2 Age distribution, proportions (%) of chronically ill persons (CHRON) and persons exposed to physical health hazards at work (PHYS), and average of systolic blood pressure (SYSBP) in the MFH survey subgroup of 30–64-year-old males.

	Sample		CHRON	PHYS	SYSBP
Age	*n*	%	%	%	Mean
30–34	508	18.8	13.8	12.8	134.0
35–39	384	14.2	21.4	17.4	136.2
40–44	437	16.2	28.4	18.8	138.5
45–49	395	14.6	44.8	18.5	141.9
50–54	379	14.0	52.2	17.4	144.7
55–59	336	12.4	68.5	21.4	151.2
60–64	260	9.6	73.8	21.2	154.3
Total sample	2699	100.0	39.8	17.8	141.8

demonstration of poststratification as considered in Sections 3.3 and 4.1, we develop the poststratification weights, and we compare the unweighted and weighted estimation results. For this, let us suppose for a moment that we are working with a simple random sample (although this is not actually true for the MFH Survey data set).

We construct the poststratification weights using the regional age distributions for both sexes, which are available on the population level. We first divide the target population into 30 regional age–sex poststrata with five regions and three age groups. Let us consider the selected MFH Survey subgroup of 30–64-year-old males; the corresponding population and sample frequency distributions and proportions are displayed in Table 5.3. Using these distributions, two different weights are derived for the sample elements in poststratum g: a weight $w_g^* = N_g/n_g$, and a rescaled weight $w_g^{**} = w_g^* \times n/N$, where N_g and n_g denote the population size and sample size in poststratum g, respectively, and N and n are the corresponding sizes of the population and the sample data set.

The weights w_g^* indicate the amount of population elements 'represented' by a single sample element. Over an n-element sample data set, these weights sum up to the relevant population size N. The rescaled weights w_g^{**} sum up to n. In Table 5.3, these weights vary only slightly around their mean value of one, indicating the self-weighting property of the MFH data set. In a strictly self-weighting data set,

Table 5.3 Poststratification weight generation for the MFH Survey subgroup of

rare in practice, the weights w_g^* would be constant and the rescaled weights w_g^{**} would be equal to one for all sample elements.

When using the weights, it is obvious that the weight w_g^* is suitable for proper estimation of population totals and the rescaled weight w_g^{**} is convenient in testing and modelling procedures when population totals are not of interest.

Developing a weight variable for poststratification is more complicated for a non-epsem data set from a complex sampling design, because there may already exist an element weight to compensate for unequal inclusion probabilities. For the simplest case, an adjusted weight to account for nonresponse can be derived by multiplying the sampling weight by the response rate in a poststratum, and then the product can be used as a weight variable in an analysis program. Strictly speaking, however, the variance estimators of poststratified estimates are different from the estimators obtained using the adjusted weights. However, in practice, the differences in variance estimates are usually small.

Let us compare the estimation results from an unweighted and a weighted MFH Survey data set, and using poststratified estimators. For simplicity, we ignore the original stratification and clustering; the MFH sample data set is thus taken as a simple random sample (drawn with replacement) for the unweighted analysis (SRSWR), a stratified simple random sample with non-proportional allocation in the weighted analysis (STRWR) and a poststratified simple random sample in the third case. Weighted estimates are obtained using the weights w_g^* or w_g^{**} in the weight variable, and the poststratification is carried out by supplying the population sizes N_g in each poststratum in the estimation procedure. The corresponding sample means and standard-error estimates of CHRON, PHYS and SYSBP are displayed below:

Study		SRSWR		STRWR		Poststratified	
variable	n	Mean	s.e	Mean	s.e	Mean	s.e
CHRON	2699	0.398	0.0094	0.386	0.0084	0.386	0.0085
PHYS	2699	0.178	0.0074	0.176	0.0073	0.176	0.0073
SYSBP	2699	141.8	0.3677	141.4	0.3353	141.4	0.3375

The unweighted and poststratified means differ for CHRON and somewhat for SYSBP because of their dependence on the demographic decomposition of the poststrata, especially on age, which is stronger than for PHYS. It should be noted that poststratification can increase efficiency. Poststratification executed as a usual stratified analysis decreases standard error estimates for CHRON and SYSBP. The extra variance owing to the poststratification can be seen from the last column (especially for SYSBP) where the standard errors are estimated using the most appropriate variance estimators. However, when compared to the stratified analysis, the differences are still quite small.

5.2 RATIO ESTIMATORS

In the estimation of variances we concentrate on *ratio estimators*, which are the simplest examples of *nonlinear* estimators. The means and proportions estimated in population subgroups, for example, the mean of systolic blood pressure and the proportion of chronically ill persons in the MFH Survey subgroup, are typical nonlinear ratio estimators. Variance estimation is examined under a stratified cluster-sampling design, which is epsem like the MFH Survey sampling design. This kind of sampling design is simple for variance estimation and is popular in practice.

Nonlinear Estimators

A linear estimator constitutes a linear function of the sample observations. Totals such as $\hat{t} = N\sum_{k=1}^{n} y_k/n$ are linear estimators when calculated from a simple random sample whose size n is fixed in advance. Under cluster sampling, situations are often encountered in which a fixed-size sample cannot be assumed. This occurs, for example, in one-stage cluster sampling if the cluster sizes B_i vary. Then, in the total estimator $\hat{t}_{rat} = N\sum_{i=1}^{m} y_i/\sum_{i=1}^{m} B_i$ (considered in Section 3.2) where y_i is the sample sum of the response variable in cluster i, the denominator should also be taken as a random variate whose value depends on which clusters are drawn. Because of this, $\hat{t}_{rat}$ turns out to be a nonlinear estimator.

The estimator $\hat{t}_{rat}$ is a special case of the ratio estimation considered in Section 3.3, where ratio estimation refers to the estimation of the population total T of a response variable using auxiliary information. There, the estimator $\hat{t}_{rat} = \hat{r} \times T_z$ was derived, where $\hat{r} = \hat{t}/\hat{t}_z$ is a ratio of the total estimators $\hat{t}$ and $\hat{t}_z$ of the response variable of interest and an auxiliary variable z, respectively, and T_z is the known population total of z. For the estimation of the population ratio $R = T/T_z$, the estimator $\hat{r}$ is directly available, and it can be written as $\hat{r} = \sum_{i=1}^{m} y_i/\sum_{i=1}^{m} z_i$. The estimator $\hat{r}$ is called an *estimator of a ratio*, or a *ratio estimator*. In this estimator, the denominator is the sample size, which is not assumed to be fixed. In practice, subpopulation means and proportions estimated from a subgroup of a sample such that the subgroup sample size is not fixed, as in the MFH Survey subgroup of 30–64-year-old males, provide the most common examples of ratio estimators. We shall consider such ratio estimators here.

Combined Ratio and Separate Ratio Estimators

Let the population clusters be divided into H strata so that there are M_h clusters in stratum h. A first-stage sample of m_h (≥ 2) clusters is drawn from each stratum h, and a second-stage sample of a total of $n = \sum_{h=1}^{H} n_h$ elements is drawn from the $m = \sum_{h=1}^{H} m_h$ sample clusters. As we often work with subgroups of the sample

whose sizes are not fixed in advance, we will use x_h in place of n_h. Note that we do not use z_h to avoid confusion with notation used for an auxiliary variable. We assume that the sample is self-weighting, i.e. the inclusion probability of each of the N population elements is constant over the strata and adjustment for nonresponse is not necessary. Element weights are thus constant for all sample elements. Further, let $y_{hi} = \sum_{k=1}^{x_{hi}} y_{hik}$ denote the subgroup sample sum of the response variable in sample cluster i of stratum h, and let x_{hi} denote the corresponding sample size. Two types of ratio estimators are derived by using the sample sums y_{hi} and x_{hi}. A *combined ratio* (across-stratum ratio) *estimator* is given by

$$\hat{r} = \frac{\sum_{h=1}^{H} y_h}{\sum_{h=1}^{H} x_h} = \frac{\sum_{h=1}^{H}\sum_{i=1}^{m_h} y_{hi}}{\sum_{h=1}^{H}\sum_{i=1}^{m_h} x_{hi}}, \tag{5.1}$$

which is a ratio estimator of a mean $\overline{Y} = T/N$ or of a proportion $P = N_1/N$, where T is the population total of a continuous response variable and N_1 is the count of persons having the value one on a binary response variable in the population subgroup considered. It is essential to note that in the ratio estimator $\hat{r}$ not only the numerator quantities y_{hi} vary between clusters but the denominator quantities x_{hi} may also do so.

For (5.1) y_{hi} and x_{hi} were first summed over the strata and clusters. A *separate ratio* (stratum-by-stratum ratio) *estimator* is a weighted sum of stratum ratios y_h/x_h. It is given by

$$\hat{r}_s = \sum_{h=1}^{H} W_h \hat{r}_h, \tag{5.2}$$

where $W_h = N_h/N$ are known stratum weights, and

$$\hat{r}_h = \frac{y_h}{x_h} = \frac{\sum_{i=1}^{m_h} y_{hi}}{\sum_{i=1}^{m_h} x_{hi}}, \quad h = 1, \ldots, H.$$

The separate ratio estimator is often used in descriptive surveys, whereas the combined ratio estimator is more common in complex analytical surveys. We will exclusively use combined ratio estimators in this chapter and in subsequent chapters and call them ratio estimators. In the case of a continuous response variable, we put $\hat{r} = \overline{y}$ (a sample mean) and in the case of a binary response, $\hat{r} = \hat{p}$ (a sample proportion). We will often denote the ratio estimator in (5.1) simply as $\hat{r} = y/x$, where $y = \sum_{h=1}^{H} y_h$ and $x = \sum_{h=1}^{H} x_h$. The quantities y and x thus

refer to the sample sum of the response variable and the sample size, respectively, in a subgroup of the sample. Note that the above discussion applies equally to an estimator $\hat{r}$ calculated from the whole sample if its size is not fixed by the sampling design.

The ratio estimator $\hat{r}$ is not unbiased but is consistent. The bias of $\hat{r}$ depends on the variability of the cluster sample sizes in the subgroup. The coefficient of variation of the cluster sample sizes x_{hi} can be used as a measure of this variability. If the coefficient of variation is small, the ratio estimator $\hat{r}$ is nearly linear and hence nearly unbiased. The bias is not disturbing if the coefficient of variation is less than, say, 0.2.

Various kinds of subgroups can be formed in which the bias properties of ratio estimators can vary. In *cross-classes*, which cut smoothly across the strata and sample clusters, the decrease in the subgroup sample sizes x_{hi} within clusters is proportional to the decrease in the subgroup sample size relative to the total sample size. The coefficient of variation of the subgroup sample sizes hence has the same magnitude as for the total sample. For this kind of a subgroup, basic features of the sampling design are well reflected; for example, the number of strata and sample clusters covered by a cross-class are usually the same as for the entire sample. Alternatively, in *segregated classes* covering only a part of the sample clusters, the coefficient of variation of the subgroup sample sizes can increase substantially. These are, for example, regional subgroups. It should be noted that, in contrast to a cross-classes-type domain, a segregated class does not properly reflect the properties of the sampling design, possibly leading to instability problems in variance estimation (see Section 5.7). Between these extremes are *mixed classes*, which are perhaps the most common subgroup types in practice. Demographic subgroups often constitute cross-classes while socioeconomic subgroups tend to be mixed classes. Moreover, a property of design-effect estimates of subpopulation ratio estimators for cross-classes is that they tend to approach unity with decreasing subgroup sample size. This property is not shared by the other types of subgroups.

Variance Estimation of a Ratio Estimator

For the ratio estimator (5.1), not only the cluster-wise variation in the numerator $\sum_{h=1}^{H} y_h$ but also the variation in the denominator $\sum_{h=1}^{H} x_h$ contributes to the total variance. Therefore, variance estimation of a ratio estimator is more complicated than that of a linear estimator. Analytical variance estimators for linear estimators, such as for population totals considered in Chapter 2, were derived according to the special features of each basic sampling technique. For nonlinear estimators, analytical variance estimators can be cumbersome or may not be available. Other types of variance estimators are thus needed. To be successful, these estimators, and the corresponding computational techniques, should have multi-purpose

properties that cover the most common types of complex sampling designs and nonlinear estimators.

Approximative variance estimators can be used for variance estimation of a nonlinear estimator. These variance estimators are not sampling-design-specific, unlike those for linear estimators. Approximative variance estimators are flexible so that they can be applied for different kinds of nonlinear estimators, including the ratio estimator, under a variety of multi-stage designs covering all the different real sampling designs selected for this book. We use the *linearization method* as the basic approximation method. Alternative methods are based on *sample reuse techniques* such as *balanced half-samples, jackknife* and *bootstrap*. Approximative techniques for variance estimation are available in statistical software products for variance estimation in complex surveys.

Certain simplifying assumptions are often made when using approximative variance estimators. In variance estimation under a multi-stage design, each sampling stage contributes to the total variance. For example, under a two-stage design, an analytical variance estimator of a population total is composed of a sum of the between-cluster and within-cluster variance components as shown in Section 3.2. In the simplest use of the approximation methods, a possible multi-stage design is reduced to a one-stage design, and the clusters are assumed to be drawn with replacement. Variances are then estimated using the between-cluster variation only. In more advanced uses of the approximation techniques, the variation of all the sampling stages can be properly accounted for.

5.3 LINEARIZATION METHOD

Linearization Method for a Nonlinear Estimator

In estimating the variance of a general nonlinear estimator, denoted by $\hat{\theta}$, we adopt a method based on the so-called *Taylor series expansion*. The method is usually called the *linearization* method because we first reduce the original nonlinear quantity to an approximate linear quantity by using the linear terms of the corresponding Taylor series expansion, and then construct the variance formula and an estimator of the variance of this linearized quantity.

Let an s-dimensional parameter vector be denoted by $\mathbf{Y} = (Y_1, \ldots, Y_s)'$ where Y_j are population totals or means. The corresponding estimator vector is denoted by $\hat{\mathbf{Y}} = (\hat{Y}_1, \ldots, \hat{Y}_s)'$ where $\hat{Y}_j$ are estimators of Y_j. We consider a nonlinear parameter $\theta = f(\mathbf{Y})$ with a consistent estimator denoted by $\hat{\theta} = f(\hat{\mathbf{Y}})$. A simple example is a subpopulation mean parameter $\theta = \overline{Y} = Y_1/Y_2$ with a ratio estimator $\hat{\theta} = \overline{y} = \hat{Y}_1/\hat{Y}_2 = y/x$, where $y = \sum_{h=1}^{H} \sum_{i=1}^{m_h} y_{hi}$ is the subgroup sample sum of the response variable and $x = \sum_{h=1}^{H} \sum_{i=1}^{m_h} x_{hi}$ is the subgroup sample size, both regarded as random quantities.

Suppose that for the function $f(\mathbf{y})$, continuous second-order derivatives exist in an open sphere containing $\mathbf{Y}$ and $\hat{\mathbf{Y}}$. Using the linear terms of the Taylor series expansion, we have an approximative linearized expression,

$$\hat{\theta} - \theta \doteq \sum_{j=1}^{s} \frac{\partial f(\mathbf{Y})}{\partial y_j}(\hat{Y}_j - Y_j), \tag{5.3}$$

where $\partial f(\mathbf{Y})/\partial y_j$ refers to partial derivation. Using the linearized equation (5.3), the variance approximation of $\hat{\theta}$ can be expressed by

$$V(\hat{\theta}) \doteq V\left(\sum_{j=1}^{s} \frac{\partial f(\mathbf{Y})}{\partial y_j}(\hat{Y}_j - Y_j)\right) = \sum_{j=1}^{s}\sum_{l=1}^{s} \frac{\partial f(\mathbf{Y})}{\partial y_j}\frac{\partial f(\mathbf{Y})}{\partial y_l} V(\hat{Y}_j, \hat{Y}_l), \tag{5.4}$$

where $V(\hat{Y}_j, \hat{Y}_l)$ denote variances and covariances of the estimators $\hat{Y}_j$ and $\hat{Y}_l$. We have hence reduced the variance of a nonlinear estimator $\hat{\theta}$ to a function of variances and covariances of s linear estimators $\hat{Y}_j$. A variance estimator $\hat{v}(\hat{\theta})$ is obtained from (5.4) by substituting the variance and covariance estimators $\hat{v}(\hat{Y}_j, \hat{Y}_l)$ for the corresponding parameters $V(\hat{Y}_j, \hat{Y}_l)$. The resulting variance estimator is a first-order Taylor series approximation where justification for ignoring the remaining higher-order terms is essentially based on practical experience derived from various complex surveys in which the sample sizes have been sufficiently large.

As an example of the linearization method, let us consider further a ratio estimator. The parameter vector is $\mathbf{Y} = (Y_1, Y_2)'$ with the corresponding estimator vector $\hat{\mathbf{Y}} = (\hat{Y}_1, \hat{Y}_2)'$. The nonlinear parameter to be estimated is $\theta = f(\mathbf{Y}) = Y_1/Y_2$, and the corresponding ratio estimator is $\hat{\theta} = f(\hat{\mathbf{Y}}) = \hat{Y}_1/\hat{Y}_2$. The partial derivatives are

$$\partial f(\mathbf{Y})/\partial y_1 = 1/Y_2 \quad \text{and} \quad \partial f(\mathbf{Y})/\partial y_2 = -Y_1/Y_2^2.$$

Hence we have

$$\begin{aligned} V(\hat{\theta}) &\doteq \sum_{j=1}^{2}\sum_{l=1}^{2} \frac{\partial f(\mathbf{Y})}{\partial y_j}\frac{\partial f(\mathbf{Y})}{\partial y_l} V(\hat{Y}_j, \hat{Y}_l) \\ &= \frac{1}{Y_2}\frac{1}{Y_2}V(\hat{Y}_1) + \frac{1}{Y_2}\left(-\frac{Y_1}{Y_2^2}\right)V(\hat{Y}_1, \hat{Y}_2) \\ &\quad + \left(-\frac{Y_1}{Y_2^2}\right)\frac{1}{Y_2}V(\hat{Y}_2, \hat{Y}_1) + \left(-\frac{Y_1}{Y_2^2}\right)\left(-\frac{Y_1}{Y_2^2}\right)V(\hat{Y}_2) \\ &= (1/Y_2^2)(V(\hat{Y}_1) + \theta^2 V(\hat{Y}_2) - 2\theta V(\hat{Y}_1, \hat{Y}_2)) \\ &= \theta^2(Y_1^{-2}V(\hat{Y}_1) + Y_2^{-2}V(\hat{Y}_2) - 2(Y_1Y_2)^{-1}V(\hat{Y}_1, \hat{Y}_2)). \end{aligned} \tag{5.5}$$

Basic principles of the linearization method for variance estimation of a nonlinear estimator under complex sampling are due to Keyfitz (1957) and Tepping (1968). Woodruff (1971) suggested simplified computational algorithms for the approximation by transforming an s-dimensional situation to a one-dimensional case. A good reference for the method is Wolter (1985). The linearization method can also be used for more complex nonlinear estimators such as correlation and regression coefficients. The linearization method is used in most survey analysis software products for variance estimation of ratio estimators and for more complicated nonlinear estimators. We next consider the estimation of the approximative variance of a ratio estimator using the linearization method.

Linearization Method for a Combined Ratio Estimator

A variance estimator of the ratio estimator $\hat{r} = y/x = \sum_{h=1}^{H} \sum_{i=1}^{m_h} y_{hi} / \sum_{h=1}^{H} \sum_{i=1}^{m_h} x_{hi}$ given by (5.1) should, according to equation (5.5), include the following terms: first, a term accounting for cluster-wise variation of the subgroup sample sums y_{hi}, second, a term accounting for cluster-wise variation of the subgroup sample sizes x_{hi}, and finally, a term accounting for joint cluster-wise variation of the sample sums y_{hi} and x_{hi}, i.e. their covariance. A variance estimator of $\hat{r}$ can thus be obtained from equation (5.5) by substituting the estimators $\hat{v}(y)$, $\hat{v}(x)$ and $\hat{v}(y, x)$ for the corresponding variance and covariance terms $V(y)$, $V(x)$ and $V(y, x)$. Hence we have

$$\hat{v}_{des}(\hat{r}) = \hat{r}^2(y^{-2}\hat{v}(y) + x^{-2}\hat{v}(x) - 2(yx)^{-1}\hat{v}(y, x)), \tag{5.6}$$

as the *design-based* variance estimator of $\hat{r}$ based on the linearization method, where $\hat{v}(y)$ is the variance estimator of the subgroup sample sum y, $\hat{v}(x)$ is the variance estimator of the subgroup sample size x, and $\hat{v}(y, x)$ is the covariance estimator of y and x.

The variance estimator (5.6) is consistent if the estimators $\hat{v}(y)$, $\hat{v}(x)$ and $\hat{v}(y, x)$ are consistent. The cluster sample sizes x_{hi} should not vary too much for the reliable performance of the approximation based on the Taylor series expansion. The method can be safely used if the coefficient of variation of x_{hi} is less than 0.2. If the cluster sample sizes are equal, the variance and covariance terms $\hat{v}(x)$ and $\hat{v}(y, x)$ are zero and the variance approximation reduces to $\hat{v}_{des}(\hat{r}) = \hat{v}(y)/x^2$. And for a binary response from simple random sampling with replacement, this variance estimator reduces to the binomial variance estimator $\hat{v}_{des}(\hat{p}) = \hat{v}_{bin}(\hat{p}) = \hat{p}(1 - \hat{p})/x$, where $x = n$, the size of the available sample data set.

The variance estimator (5.6) is a large-sample approximation in that a good variance estimate can be expected if not only a large element-level sample is available but a large number of sample clusters is also present. In the case of a small number of sample clusters, the variance estimator can be unstable; this will be examined in Section 5.7.

Strictly speaking, the variance and covariance estimators in (5.6) depend on the actual sampling design. But assuming that at least two sample clusters are drawn from each stratum and by using the with-replacement assumption, i.e. assuming that clusters are drawn independently of each other, we obtain relatively simple variance and covariance estimators, which can be generally applied for multi-stage stratified epsem samples:

$$\hat{v}(y) = \sum_{h=1}^{H} m_h \hat{s}_{yh}^2, \quad \hat{v}(x) = \sum_{h=1}^{H} m_h \hat{s}_{xh}^2$$

and

$$\hat{v}(y, x) = \sum_{h=1}^{H} m_h \hat{s}_{yxh},$$

where

$$\hat{s}_{yh}^2 = \sum_{i=1}^{m_h} (y_{hi} - y_h/m_h)^2/(m_h - 1),$$

$$\hat{s}_{xh}^2 = \sum_{i=1}^{m_h} (x_{hi} - x_h/m_h)^2/(m_h - 1),$$

and

$$\hat{s}_{yxh} = \sum_{i=1}^{m_h} (y_{hi} - y_h/m_h)(x_{hi} - x_h/m_h)/(m_h - 1). \tag{5.7}$$

Note that by using the with-replacement approximation, only the between-cluster variation is accounted for. Therefore, the corresponding variance estimators underestimate the true variance. This bias is negligible if the stratum-wise first-stage sampling fractions are small, which is the case when there are a large number of population clusters in each stratum (see Section 3.2).

For the estimation of the between-cluster variance, at least two sample clusters are needed. If the sampling design is such that exactly two clusters are drawn from each stratum, the estimators (5.7) can be further simplified:

$$\hat{v}(y) = \sum_{h=1}^{H} (y_{h1} - y_{h2})^2, \quad \hat{v}(x) = \sum_{h=1}^{H} (x_{h1} - x_{h2})^2$$

and

$$\hat{v}(y, x) = \sum_{h=1}^{H} (y_{h1} - y_{h2})(x_{h1} - x_{h2}). \tag{5.8}$$

This kind of design is popular in practice because of the simplicity of the variance and covariance estimators. The modified MFH Survey sampling design is of this type. The linearization method is demonstrated in the MFH Survey's two-stage design in Example 5.1.

Example 5.1

Linearization method in the MFH Survey. We consider the estimation of the variance of a subpopulation proportion estimator $\hat{r} = \hat{p}$ for the binary response variable CHRON (chronic morbidity) and a subpopulation mean estimator $\hat{r} = \overline{y}$ for the continuous response variable SYSBP (systolic blood pressure) by the linearization method. The MFH Survey subgroup covers 30–64-year-old males. The subgroup sample size is $x = 2699$ and the data set is self-weighting. In the modified MFH Survey sampling design, described in Section 5.1, there are $H = 24$ regional strata and $m = 48$ regional sample clusters. Two sample clusters are thus drawn from each stratum. Recall that the subgroup maintains these properties of the sampling design because it constitutes a cross-classes-type domain. The data set is displayed in Table 5.4.

For the binary response variable CHRON, we obtain:

$$y = \sum_{h=1}^{24} \sum_{i=1}^{2} y_{hi} = \sum_{h=1}^{24} (y_{h1} + y_{h2}) = 1073$$

chronically ill males in the sample, and a sample sum of

$$x = \sum_{h=1}^{24} \sum_{i=1}^{2} x_{hi} = \sum_{h=1}^{24} (x_{h1} + x_{h2}) = 2699$$

males in the subgroup. The subpopulation proportion estimate of CHRON is

$$\hat{p} = y/x = 1073/2699 = 0.3976.$$

For the variance estimate $\hat{v}_{des}(\hat{p})$ of $\hat{p}$, we calculate the variance and covariance estimates $\hat{v}(y)$, $\hat{v}(x)$ and $\hat{v}(y, x)$. By using equation (5.8), these are:

$$\hat{v}(y) = \sum_{h=1}^{24} (y_{h1} - y_{h2})^2 = 1545, \quad \hat{v}(x) = \sum_{h=1}^{24} (x_{h1} - x_{h2})^2 = 2527$$

and

$$\hat{v}(y, x) = \sum_{h=1}^{24} (y_{h1} - y_{h2})(x_{h1} - x_{h2}) = 1435.$$

Table 5.4 Cluster sample sums y_{hi} of the response variables CHRON and SYSBP and the corresponding cluster sample sizes x_{hi} for the subgroup of 30–64-year-old males in the MFH Survey.

Stratum h	Cluster i	CHRON y_{hi}	SYSBP y_{hi}	x_{hi}	Cluster i	CHRON y_{hi}	SYSBP y_{hi}	x_{hi}
1	1	70	29 056	204	2	74	29 417	210
2	1	12	3692	26	2	14	4564	30
3	1	15	7741	59	2	16	8585	63
4	1	9	6277	45	2	14	5668	43
5	1	10	2322	17	2	16	3960	30
6	1	10	3080	21	2	6	3252	22
7	1	10	3966	27	2	4	3261	24
8	1	12	4156	28	2	6	2852	20
9	1	15	6617	46	2	23	6616	48
10	1	37	10 552	73	2	25	11 032	77
11	1	11	8759	60	2	25	9876	72
12	1	33	9901	69	2	24	6828	47
13	1	31	8624	61	2	27	9390	66
14	1	22	6960	48	2	20	7130	49
15	1	18	6646	49	2	22	7094	49
16	1	24	9841	69	2	37	11 786	83
17	1	19	6910	48	2	23	6446	45
18	1	25	10 742	73	2	29	9026	61
19	1	36	9350	65	2	34	8912	62
20	1	9	3810	26	2	22	7098	51
21	1	18	6998	53	2	34	9970	69
22	1	29	11 146	79	2	41	13 215	94
23	1	22	6596	48	2	18	6002	41
24	1	15	3808	27	2	7	3148	22
Over both clusters in all strata						1073	382 678	2699

Using these estimates, we obtain a variance estimate (5.6):

$$\begin{aligned}\hat{v}_{des}(\hat{p}) &= \hat{p}^2(y^{-2}\hat{v}(y) + x^{-2}\hat{v}(x) - 2(y \times x)^{-1}\hat{v}(y, x)) \\ &= 0.3976^2 \times (1073^{-2} \times 1545 + 2699^{-2} \times 2527 \\ &\quad - 2 \times (1073 \times 2699)^{-1} \times 1435) = 0.1103 \times 10^{-3}.\end{aligned}$$

For the continuous response variable SYSBP, we obtain the sample sum

$$y = \sum_{h=1}^{24} \sum_{i=1}^{2} y_{hi} = \sum_{h=1}^{24} (y_{h1} + y_{h2}) = 382\,678.$$

Hence the subpopulation mean estimate of SYSBP is

$$\overline{y} = y/x = 382\,678/2699 = 141.785.$$

For the variance estimate $\hat{v}_{des}(\overline{y})$ of $\overline{y}$, we obtain:

$$\hat{v}(y) = \sum_{h=1}^{24}(y_{h1} - y_{h2})^2 = 50\,469\,516$$

and

$$\hat{v}(y, x) = \sum_{h=1}^{24}(y_{h1} - y_{h2})(x_{h1} - x_{h2}) = 349\,962.$$

Using these estimates, we obtain a variance estimate (5.6):

$$\begin{aligned}\hat{v}_{des}(\overline{y}) &= \overline{y}^2(y^{-2}\hat{v}(y) + x^{-2}\hat{v}(x) - 2(y \times x)^{-1}\hat{v}(y, x)) \\ &= 141.785^2 \times (382\,678^{-2} \times 50\,469\,516 + 2699^{-2} \times 2527 \\ &\quad - 2 \times (382\,678 \times 2699)^{-1} \times 349\,962) = 0.2788.\end{aligned}$$

All these variances could be estimated from the cluster-level data set given in Table 5.4. For CHRON we next calculate a binomial variance estimate of $\hat{p}$ corresponding to simple random sampling with replacement, and the corresponding design-effect estimate. The variance estimate is

$$\hat{v}_{bin}(\hat{p}) = \hat{p}(1 - \hat{p})/x = 0.3976 \times (1 - 0.3976)/2699 = 0.0887 \times 10^{-3},$$

where $\hat{v}_{bin}$ is the standard binomial variance estimator. The design-effect estimate is $\hat{d}(\hat{p}) = \hat{v}_{des}(\hat{p})/\hat{v}_{bin}(\hat{p}) = 1.24$. Note that the design-effect estimate is noticeably smaller than that for the total survey data because the subgroup is a cross-class. The design-effect estimate also indicates that intra-cluster correlation in CHRON in the subgroup is only slight. For SYSBP, on the other hand, access to the individual-level data set is required for the calculation of the variance estimate of $\overline{y}$ with an assumption of simple random sampling with replacement. This turns out to be

$$\hat{v}_{srswr}(\overline{y}) = \sum_{k=1}^{2699}(y_k - \overline{y})^2/(2699(2699 - 1)) = 0.1352,$$

and hence the design-effect estimate is $\hat{d}(\overline{y}) = 2.06$. The estimate indicates a substantial intra-cluster correlation in the response SYSBP in the subgroup, even the estimate is considerably smaller than that for the total survey data. The coefficient of variation of the subgroup sample size is $\text{c.v}(x) = \text{s.e}(x)/x = 0.019$,

which is small enough to justify the use of the Taylor series linearization. We finally collect the estimation results below.

Study variable	Estimate $\hat{r}$	Standard-error estimate $\text{s.e}_{des}(\hat{r})$	Standard-error estimate $\text{s.e}_{srs}(\hat{r})$	deff
CHRON	0.3976	0.0105	0.0094	1.24
SYSBP	141.785	0.5280	0.3677	2.06

In practice, the estimation of the variance of a ratio-type proportion or mean estimator can be carried out by suitable software for survey analysis. Instead of the cluster-level data set, an individual-level data set is usually used as input in applying survey analysis software. For further training, the user is encouraged to visit the web extension of the book.

5.4 SAMPLE REUSE METHODS

Sample reuse methods can be used as an alternative to the linearization method in variance approximation of a nonlinear estimator $\hat{\theta}$ under complex multi-stage designs. The term *reuse* refers to a procedure in which variance estimation is based on repeated utilization of the sampled data set that itself is obtained as a single sample from the population. Therefore, these methods are sometimes called *pseudoreplication* techniques. Pseudoreplication should be distinguished from techniques such as the *random groups methodology*, which rely on true replication where several independent samples are actually drawn from the same population. These methods are excluded here because of their limited practical applicability in complex analytical surveys.

In this section, we consider three particular sample reuse techniques: *balanced half-samples*, *jackknife* and *bootstrap*. They all share the following basic variance estimation procedure (which actually originates from random groups methodology):

1. From the sample data set, we draw K *pseudosamples* by a particular technique with a value of K that is specific to each reuse method.
2. An estimate $\hat{\theta}_k$ mimicking the parent estimator $\hat{\theta}$ is obtained from each of the K pseudosamples.
3. The variance $V(\hat{\theta})$ of the estimator $\hat{\theta}$ is estimated by using the observed variation of the pseudosample estimates $\hat{\theta}_k$, essentially based on squared differences of the form $(\hat{\theta}_k - \hat{\theta})^2$. Typically, sample reuse estimators are of the form $\hat{v}(\hat{\theta}) = c\sum_{k=1}^{K}(\hat{\theta}_k - \hat{\theta})^2$, where c is a constant, specific for each sample reuse method. An average $\overline{\hat{\theta}} = \sum_{k=1}^{K}\hat{\theta}_k/K$ of the K pseudosample estimates $\hat{\theta}_k$ can be used in place of $\hat{\theta}$ to form the squared differences.

The estimator $\hat{\theta}$ is usually a nonlinear estimator, a ratio estimator or an estimator of a regression coefficient. In the linearization method, analytical expressions for partial derivatives of such nonlinear functions were needed in the construction of a variance estimator. This is not so in sample reuse techniques. In fact, the basic variance estimation procedure described above is independent of the type of estimator and, therefore, the methods are applicable for any kind of nonlinear estimator. Pseudoreplication techniques, especially the bootstrap, however, involve much more computation than the linearization method; thus they are flexible but computer-intensive.

The technique of balanced half-samples was introduced by McCarthy (1966, 1969) for variance approximation of a nonlinear estimator under an epsem design, where a large number of strata are formed and exactly two clusters are drawn with replacement from each stratum. For variance estimation in a similar design, McCarthy (1966) also introduced the jackknife method, which was originally developed by Quenouille (1956) for bias reduction of an estimator. A key property of the jackknife method is compactly stated as *jack-of-all-trades and master of none*. Both methods have been generalized for more complex designs involving more than two clusters per stratum and without-replacement sampling of clusters. Good introductions to balanced half-samples and jackknife techniques for complex surveys may be found in Wolter (1985) and Rao *et al.* (1992).

Bootstrapping was introduced by Efron (1982) for a general nonparametric methodology for various statistical problems: '*Our goal is to understand a collection of ideas concerning the nonparametric estimation of bias, variance and more general measures of error*' (Efron 1982, p. 1). Since then, the technique has been extensively applied, using computer-intensive simulation, for a variety of non-standard variance and confidence-interval approximation problems when working with independent observations. Originating, like the jackknife, outside the survey-sampling framework, the bootstrap technique has been only recently applied for variance estimation of nonlinear estimators in complex surveys. One of the first developments for finite-population without-replacement sampling was given in McCarthy and Snowden (1985). Extensions of the bootstrap technique are given in Rao and Wu (1988), Rao *et al.* (1992), covering non-smooth functions such as quantiles, and Sitter (1992, 1997). A brief summary of the bootstrap technique for complex surveys is given in Särndal *et al.* (1992).

Here we only introduce the basic principles of the sample reuse techniques and concentrate on their practical application within the MFH Survey setting. As an example of a nonlinear estimator we again consider the (combined) ratio estimator $\hat{r} = y/x$, given by (5.1), where $y = \sum_{h=1}^{H} \sum_{i=1}^{m_h} y_{hi}$ is the sum of the cluster-level subgroup sample sums of a response variable and $x = \sum_{h=1}^{H} \sum_{i=1}^{m_h} x_{hi}$ is the corresponding sum of the cluster-level subgroup sample sizes. A two-stage epsem sampling design is assumed such that the clusters are drawn with replacement. The with-replacement assumption involves bias to the approximative variance estimates but the bias is negligible if the first-stage sampling fraction is small. Note

that the cluster-level data set used for variance approximation in all the reuse methods is similar to that used in the linearization method.

Balanced half-samples and jackknife techniques for variance approximation of a ratio estimator $\hat{r}$ are examined in a design in which exactly two clusters are drawn from each stratum. Note that the MFH Survey sampling design is of this type. The bootstrap technique is applied to a more general design in which at least two clusters are drawn from each stratum but the number of sample clusters is constant over the strata. Under these designs, the techniques are here called *balanced repeated replications* (*BRR*), *jackknife repeated replications* (*JRR*) and *bootstrap repeated replications* (*BOOT*). Because there are several alternative versions of BRR and JRR suggested in the literature, our aim is also to compare estimation results with each other, and also with the results attained by the linearization method. An overall comparison is given in Section 5.5.

Sample reuse methods differ in their asymptotic and other properties, computational requirements and practicality. Comparative results for the properties of the sample reuse methods for nonlinear estimators from complex sampling are reported by Kish and Frankel (1970, 1974), Bean (1975), Krewski and Rao (1981), Rao and Wu (1985, 1988), Rao *et al.* (1992) and Shao and Tu (1995). We discuss briefly the relative merits of the methods in Section 5.5.

The BRR Technique

In its basic form, the technique of *balanced repeated replications* can be applied to variance approximation in epsem designs where exactly two clusters are drawn with replacement from each stratum, and the number of strata is large. We consider using this design, the BRR method for a ratio estimator $\hat{r} = y/x$, which is a subpopulation mean or proportion estimator, where $y = \sum_{h=1}^{H}(y_{h1} + y_{h2})$ and $x = \sum_{h=1}^{H}(x_{h1} + x_{h2})$ and y_{hi}, x_{hi}, $i = 1, 2$, are the cluster-level sample sums previously given.

The way of forming pseudosamples in the BRR technique starts from the fact that, with H strata and $m_h = 2$ sample clusters per stratum, the total sample can be split into 2^H overlapping half-samples each with H sample clusters. For each half sample, one of the pairs (y_{11}, x_{11}) and (y_{12}, x_{12}) from the first stratum, one of the pairs (y_{21}, x_{21}) and (y_{22}, x_{22}) from the second stratum, and so forth, is selected. A ratio estimator

$$\hat{r}_k = \frac{\sum_{h=1}^{H}\sum_{i=1}^{2}\delta_{hik}y_{hi}}{\sum_{h=1}^{H}\sum_{i=1}^{2}\delta_{hik}x_{hi}}, \quad k = 1, \ldots, 2^H \tag{5.9}$$

is derived for each half-sample k, where the weights $\delta_{hik} = 1$ if the cluster hi is selected in the kth half-sample, and $\delta_{hik} = 0$ otherwise.

Variance estimator of the mean of $\hat{r}_k$ over all half-samples, namely,

$$\overline{\hat{r}} = \sum_{k=1}^{2^H} \hat{r}_k / 2^H, \tag{5.10}$$

and that of the parent estimator, $\hat{r}$, can be constructed using $\hat{r}_k$ obtained from the half-samples. Hence we have:

$$\hat{v}(\overline{\hat{r}}) = \sum_{k=1}^{2^H} (\hat{r}_k - \overline{\hat{r}})^2 / 2^H$$

and

$$\hat{v}(\hat{r}) = \sum_{k=1}^{2^H} (\hat{r}_k - \hat{r})^2 / 2^H. \tag{5.11}$$

If $\hat{r}$ is a linear estimator, an identity $\hat{r} = \overline{\hat{r}}$ holds, and the two variance estimators in (5.11) are equal. Although for a ratio estimator the identity does not hold, in practice the parent estimate and the mean of the half-sample estimates are usually close and either of the variance estimators (5.11) could be used as a variance estimator for the parent estimator $\hat{r}$. But it is obvious that these variance estimators are not useful in practice because they often presuppose forming a very large number of half-samples, e.g. in the MFH Survey setting about 17 million. To avoid the onerous task of constructing all possible pseudosamples, a subset of them may be selected. But if this subset is chosen at random, a nonzero cross-stratum covariance term will appear in the corresponding variance estimator. In the BRR technique, a subset of K half-samples is selected by a *balanced* method. Balancing involves the selection of the half-samples in such a way that the cross-stratum covariance term is zero. This considerably reduces the number of half-samples needed. In practice, the number K should be selected such that it is at least equal to the number of strata H.

The balanced selection of half-samples is achieved by applying a method developed by Plackett and Burman (1946) for the construction of $K \times K$ orthogonal matrices where K is an integer multiple of 4. An example of such an orthogonal *Hadamard* matrix $\mathbf{B}$ with $K = 12$ such that $\mathbf{B'B} = 12 \times \mathbf{I}$, where $\mathbf{I}$ denotes an identity matrix, is given below. The rows in the matrix refer to the half-samples and the columns to the strata. A $+1$ in a cell (k, h) of the matrix denotes that the first cluster $h1$ in a stratum h is included in the kth half sample, whilst -1 denotes that the cluster $h2$ is included. Note that complement half-samples can be obtained simply by reversing the signs in the matrix. The number of half-samples,

$K = 12$, is thus noticeably smaller than the total amount of possible half-samples, which in this case is $2^{12} = 4096$.

Half-sample k	Stratum h											
	1	2	3	4	5	6	7	8	9	10	11	12
1	+1	−1	+1	−1	−1	−1	+1	+1	+1	−1	+1	−1
2	+1	+1	−1	+1	−1	−1	−1	+1	+1	+1	−1	−1
3	−1	+1	+1	−1	+1	−1	−1	−1	+1	+1	+1	−1
4	+1	−1	+1	+1	−1	+1	−1	−1	−1	+1	+1	−1
5	+1	+1	−1	+1	+1	−1	+1	−1	−1	−1	+1	−1
6	+1	+1	+1	−1	+1	+1	−1	+1	−1	−1	−1	−1
7	−1	+1	+1	+1	−1	+1	+1	−1	+1	−1	−1	−1
8	−1	−1	+1	+1	+1	−1	+1	+1	−1	+1	−1	−1
9	−1	−1	−1	+1	+1	+1	−1	+1	+1	−1	+1	−1
10	+1	+1	−1	−1	+1	+1	+1	−1	+1	+1	−1	−1
11	−1	−1	−1	−1	+1	+1	+1	−1	+1	+1	+1	−1
12	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1

If the actual number of strata is 12, we use the full matrix in the balanced construction of the half-samples. If H is smaller than K, e.g. 10, we can choose any 10 rows of the matrix. In the MFH Survey design, we will use $K = 24$, which equals the number of strata. When working with linear estimators, *full orthogonal balance* is reached, which involves equality of a full-sample mean estimate with the estimate obtained as an average of the half-sample estimates, by choosing K as an integer multiple of 4, which is greater than H. Hadamard matrices of orders 2 to 100 are given in Wolter (1985); such matrices can also be easily reproduced by a suitable computer algorithm.

Several BRR variance estimators are suggested in the literature for the variance $V(\hat{r})$ of the parent estimator $\hat{r}$. The variance estimator based on estimators $\hat{r}_k$ from the K half-samples and the full-sample estimator $\hat{r}$ is

$$\hat{v}_{1.brr}(\hat{r}) = \sum_{k=1}^{K} (\hat{r}_k - \hat{r})^2 / K, \tag{5.12}$$

which is equal to (5.11) based on all 2^H half-samples. As a counterpart to the variance estimator $\hat{v}_{1.brr}(\hat{r})$, an estimator based on estimates $\hat{r}_k^c$ obtained from the K complement half-samples is given by

$$\hat{v}_{2.brr}(\hat{r}) = \sum_{k=1}^{K} (\hat{r}_k^c - \hat{r})^2 / K. \tag{5.13}$$

Using the variance estimators (5.12) and (5.13), a combined variance estimator

$$\hat{v}_{3.brr}(\hat{r}) = (\hat{v}_{1.brr}(\hat{r}) + \hat{v}_{2.brr}(\hat{r}))/2 \tag{5.14}$$

is derived. Counterparts to the variance estimators (5.12)–(5.14) can be derived on the basis of the averages of $\hat{r}_k$ and $\hat{r}_k^c$. An estimator corresponding to $\hat{v}_{1.brr}$ is hence

$$\hat{v}_{4.brr}(\hat{r}) = \sum_{k=1}^{K}(\hat{r}_k - \overline{\hat{r}})^2/K, \quad \text{where} \quad \overline{\hat{r}} = \sum_{k=1}^{K}\hat{r}_k/K, \tag{5.15}$$

and that formed by using the complement half-samples is

$$\hat{v}_{5.brr}(\hat{r}) = \sum_{k=1}^{K}(\hat{r}_k^c - \overline{\hat{r}}^c)^2/K, \quad \text{where} \quad \overline{\hat{r}}^c = \sum_{k=1}^{K}\hat{r}_k^c/K. \tag{5.16}$$

Using $\hat{v}_{4.brr}$ and $\hat{v}_{5.brr}$ we obtain a counterpart to $\hat{v}_{3.brr}$:

$$\hat{v}_{6.brr}(\hat{r}) = (\hat{v}_{4.brr}(\hat{r}) + \hat{v}_{5.brr}(\hat{r}))/2. \tag{5.17}$$

Using the estimators $\hat{r}_k$ and $\hat{r}_k^c$ from all the half-samples, we finally obtain

$$\hat{v}_{7.brr}(\hat{r}) = \sum_{k=1}^{K}(\hat{r}_k - \hat{r}_k^c)^2/4K. \tag{5.18}$$

For a linear estimator, all these variance estimators coincide. However, this is not so for a ratio estimator. For example, there is a relationship between $\hat{v}_{3.brr}$ and $\hat{v}_{7.brr}$:

$$\hat{v}_{3.brr}(\hat{r}) = \hat{v}_{7.brr}(\hat{r}) + \sum_{k=1}^{K}(\overline{\hat{r}} - \hat{r})^2/K,$$

and hence $\hat{v}_{3.brr}(\hat{r}) \geq \hat{v}_{7.brr}(\hat{r})$. According to Wolter (1985), $\hat{v}_{7.brr}$ could be regarded as the most natural BRR variance estimator for the parent estimator $\hat{\theta}$. In practice, however, all the estimators should yield nearly equal variance estimates, as appears to be true in the MFH Survey.

Example 5.2

The BRR technique in the MFH Survey. We continue working with variance approximation of ratio-type subpopulation mean and proportion estimators from the MFH Survey data, as considered in the previous section for the linearization method. The binary response variable CHRON (chronic morbidity) and the

continuous response variable SYSBP (systolic blood pressure) are used. The subgroup consists of 30–64-year-old males; the subgroup size is 2699. A proportion estimator for CHRON is denoted by $\hat{r} = \hat{p}$ and a mean estimator for SYSBP is denoted by $\hat{r} = \overline{y}$. We calculate all the seven BRR variance estimators for $\hat{p}$ and $\overline{y}$.

Recall that there are $H = 24$ strata and $m = 48$ sample clusters in the modified MFH Survey design, with exactly two clusters drawn from each stratum. Variance estimation by BRR starts with forming the K half-samples and the corresponding complement half-samples. We choose $K = 24$, i.e. the number of strata, and use the whole matrix in forming the half-samples and their complements. Note that for a full orthogonal balance we would choose $K = 28$. We work out a weight matrix from the 24×24 Hadamard matrix to perform the computations, which are based on the cluster-level data set given in Example 5.1.

The parent ratio and mean estimates $\hat{p}$ and $\overline{y}$, and the corresponding means of the half-sample estimates $\hat{p}_k$ and $\overline{y}_k$ with their complement half-sample estimates $\hat{p}_k^c$ and $\overline{y}_k^c$, are first calculated. These are:

$$\hat{p} = 0.3976, \quad \overline{\hat{p}} = \sum_{k=1}^{24} \hat{p}_k/24 = 0.3953 \quad \text{and} \quad \overline{\hat{p}}^c = \sum_{k=1}^{24} \hat{p}_k^c/24 = 0.3997,$$

$$\overline{y} = 141.785, \quad \overline{\hat{y}} = \sum_{k=1}^{24} \overline{y}_k/24 = 141.804 \quad \text{and} \quad \overline{\hat{y}}^c = \sum_{k=1}^{24} \overline{y}_k^c/24 = 141.768.$$

All three CHRON proportion estimates and SYSBP mean estimates are close. We next calculate the BRR variance estimates (5.12)–(5.18). For CHRON, using $\hat{p}$ we obtain from the half-samples and their complements:

$$\hat{v}_{1.brr}(\hat{p}) = \sum_{k=1}^{24} (\hat{p}_k - 0.3976)^2/24 = 0.1104 \times 10^{-3},$$

$$\hat{v}_{2.brr}(\hat{p}) = \sum_{k=1}^{24} (\hat{p}_k^c - 0.3976)^2/24 = 0.1103 \times 10^{-3},$$

and

$$\hat{v}_{3.brr}(\hat{p}) = (\hat{v}_{1.brr}(\hat{p}) + \hat{v}_{2.brr}(\hat{p}))/2 = 0.1103 \times 10^{-3}.$$

Using the mean estimates $\overline{\hat{p}}$ and $\overline{\hat{p}}^c$, we obtain the counterparts:

$$\hat{v}_{4.brr}(\hat{p}) = \sum_{k=1}^{24} (\hat{p}_k - 0.3953)^2/24 = 0.1052 \times 10^{-3},$$

$$\hat{v}_{5.brr}(\hat{p}) = \sum_{k=1}^{24} (\hat{p}_k^c - 0.3997)^2/24 = 0.1056 \times 10^{-3},$$

and

$$\hat{v}_{6.brr}(\hat{p}) = (\hat{v}_{4.brr}(\hat{p}) + \hat{v}_{5.brr}(\hat{p}))/2 = 0.1054 \times 10^{-3}.$$

From all the half-samples we finally obtain:

$$\hat{v}_{7.brr}(\hat{p}) = \sum_{k=1}^{24} (\hat{p}_k - \hat{p}_k^c)^2/(4 \times 24) = 0.1103 \times 10^{-3}.$$

For CHRON the first three BRR variance estimates, and the last one, happen to be equal to those obtained by the linearization method. Those based on the mean of the half-sample estimates are somewhat, but not very much, smaller.

For SYSBP, we obtain the following BRR variance estimates:

$$\hat{v}_{1.brr}(\overline{y}) = \sum_{k=1}^{24} (\overline{y}_k - 141.785)^2/24 = 0.2791,$$

$$\hat{v}_{2.brr}(\overline{y}) = \sum_{k=1}^{24} (\overline{y}_k^c - 141.785)^2/24 = 0.2790,$$

$$\hat{v}_{3.brr}(\overline{y}) = (\hat{v}_{1.brr}(\overline{y}) + \hat{v}_{2.brr}(\overline{y}))/2 = 0.2791,$$

$$\hat{v}_{4.brr}(\overline{y}) = \sum_{k=1}^{24} (\overline{y}_k - 141.804)^2/24 = 0.2787,$$

$$\hat{v}_{5.brr}(\overline{y}) = \sum_{k=1}^{24} (\overline{y}_k^c - 141.768)^2/24 = 0.2788,$$

$$\hat{v}_{6.brr}(\overline{y}) = (\hat{v}_{4.brr}(\overline{y}) + \hat{v}_{5.brr}(\overline{y}))/2 = 0.2787,$$

$$\hat{v}_{7.brr}(\overline{y}) = \sum_{k=1}^{24} (\overline{y}_k - \overline{y}_k^c)^2/(4 \times 24) = 0.2790.$$

For SYSBP, all the BRR variance estimates (and that obtained by the linearization method) are equal to 0.279 when rounded to three digits.

All the BRR variance estimators provided similar results for a ratio estimator, a subpopulation proportion or a mean, for the response variables considered. These results equal those drawn from other comparable empirical studies. Also on theoretical grounds, no definite preference for the BRR variance estimators of a nonlinear estimator can be given. In addition to the BRR variance estimators introduced, other versions have been also developed, such as a BRR variant, called the Fay's method (Judkins 1990), resembling jackknife-type estimation.

The JRR Technique

The particular jackknife method based on *jackknife repeated replications* has many features of the BRR technique, since only the method of forming the pseudosamples is different. Application of the JRR technique to a design where more than two sample clusters are drawn from a stratum is more straightforward than for BRR. We, however, consider the JRR technique in the simplest case where the number of sample clusters per stratum is exactly two, and the clusters are assumed to be drawn with replacement, i.e. with a design similar to that required for BRR. JRR variance estimators are derived for a ratio estimator $\hat{r}$, which is a subpopulation proportion or mean estimator.

We construct the pseudosamples following the method suggested by Frankel (1971). For the first pseudosample, we exclude the first cluster $h1$ from the first stratum and weight the second cluster $h2$ by the value 2, leaving the remaining $H-1$ strata unchanged. By repeating this procedure for all strata, we get a total of H pseudosamples. For a similar set of H complement pseudosamples, we change the order of the clusters that are excluded. The JRR variance estimators are derived using these two sets of pseudosamples.

Like the BRR technique, several alternative JRR variance estimators can be constructed for the parent ratio estimator $\hat{r}$. For these, we first derive the pseudosample estimators for each stratum. Let $\hat{r}_h$ denote a pseudosample estimator based on excluding cluster $h1$ and duplicating cluster $h2$ in stratum h:

$$\hat{r}_h = \frac{2y_{h2} + \sum_{h' \neq h}^{H} \sum_{i=1}^{2} y_{h'i}}{2x_{h2} + \sum_{h' \neq h}^{H} \sum_{i=1}^{2} x_{h'i}}, \quad h = 1, \ldots, H. \tag{5.19}$$

These estimators are constructed for each pseudosample. From the complement pseudosamples, we obtain corresponding estimators $\hat{r}_h^c$ by excluding cluster $h2$ and duplicating cluster $h1$. Using the pseudosample estimators and the complement pseudosample estimators, we can derive the first set of JRR variance estimators for the parent estimator $\hat{r}$. Hence we have

$$\hat{v}_{1.jrr}(\hat{r}) = \sum_{h=1}^{H} (\hat{r}_h - \hat{r})^2, \tag{5.20}$$

and from the complement pseudosamples

$$\hat{v}_{2.jrr}(\hat{r}) = \sum_{h=1}^{H} (\hat{r}_h^c - \hat{r})^2. \tag{5.21}$$

A combined variance estimator is

$$\hat{v}_{3.jrr}(\hat{r}) = (\hat{v}_{1.jrr}(\hat{r}) + \hat{v}_{2.jrr}(\hat{r}))/2. \tag{5.22}$$

Another set of variance estimators can be obtained using the so-called *pseudovalues* introduced by Quenouille (1956) to reduce the bias of an estimator. In the case considered above, pseudovalues are of the form

$$\hat{r}_h^p = 2\hat{r} - \hat{r}_h, \quad h = 1, \ldots, H, \tag{5.23}$$

and for the complement pseudosamples they are denoted by $\hat{r}_h^{pc}$. By using the first set of H pseudovalues $\hat{r}_h^p$, we obtain a bias-corrected estimator given by

$$\bar{\hat{r}}^p = \sum_{h=1}^{H} \hat{r}_h^p / H, \tag{5.24}$$

and using the pseudovalues $\hat{r}_h^{pc}$ from the complement pseudosamples we obtain

$$\bar{\hat{r}}^{pc} = \sum_{h=1}^{H} \hat{r}_h^{pc} / H. \tag{5.25}$$

Counterparts to the variance estimators (5.20)–(5.22) can be derived from the pseudovalues and the bias-corrected estimators, giving

$$\hat{v}_{4.jrr}(\hat{r}) = \sum_{h=1}^{H} (\hat{r}_h^p - \bar{\hat{r}}^p)^2, \tag{5.26}$$

and from the complement pseudosamples

$$\hat{v}_{5.jrr}(\hat{r}) = \sum_{h=1}^{H} (\hat{r}_h^{pc} - \bar{\hat{r}}^{pc})^2. \tag{5.27}$$

A combined variance estimator can also be derived:

$$\hat{v}_{6.jrr}(\hat{r}) = (\hat{v}_{4.jrr}(\hat{r}) + \hat{v}_{5.jrr}(\hat{r}))/2. \tag{5.28}$$

Finally, from all the $2H$ pseudosamples we obtain:

$$\hat{v}_{7.jrr}(\hat{r}) = \sum_{h=1}^{H} (\hat{r}_h - \hat{r}_h^c)^2 / 4. \tag{5.29}$$

A similar way of constructing the JRR variance estimators was used to that given for the BRR technique. For a linear estimator, the bias-corrected JRR estimators reproduce the parent estimator, and all the JRR variance estimators coincide. This is not the case for nonlinear estimators, but in practice all JRR variance estimators should give closely related results. Like BRR, the variance estimator $\hat{v}_{7.jrr}$ could be taken as the most natural estimator of the variance of the parent estimator $\hat{\theta}$.

The JRR technique can be extended to a more general case in which more than two clusters are drawn from each stratum, for without-replacement sampling of clusters. Pseudosamples and their complements are constructed by consecutively excluding a cluster and weighting the remaining clusters appropriately in a stratum (see Section 4.6 in Wolter 1985).

Like BRR, we use the JRR technique for variance estimation of a ratio estimator $\hat{r}$ for the MFH Survey design.

Example 5.3

The JRR technique in the MFH Survey. We continue to consider the estimation of variance of a ratio-type subpopulation proportion estimator $\hat{p}$ of CHRON (chronic morbidity) and a subpopulation mean estimator $\overline{y}$ of SYSBP (systolic blood pressure) for 30–64-year-old males. Using the cluster-level data set available, we calculate all the seven JRR variance estimates for $\hat{p}$ and $\overline{y}$.

Because $H = 24$, we construct 24 JRR pseudosamples with their complements by the Frankel method. The parent ratio and mean estimates $\hat{p}$ and $\overline{y}$, and the corresponding bias-corrected estimators given by (5.24) and (5.25) based on the pseudovalues $\hat{p}_h^p, \hat{p}_h^{pc}, \overline{y}_h^p$ and $\overline{y}_h^{pc}$ calculated from the pseudosamples and their complements, are first obtained. These are

$$\hat{p} = 0.3976,\ \hat{\overline{p}}^p = \sum_{k=1}^{24} \hat{p}_k^p/24 = 0.3972 \quad \text{and} \quad \hat{\overline{p}}^{pc} = \sum_{k=1}^{24} \hat{p}_k^{pc}/24 = 0.3980,$$

$$\overline{y} = 141.785,\ \hat{\overline{y}}^p = \sum_{k=1}^{24} \overline{y}_k^p/24 = 141.793 \quad \text{and} \quad \hat{\overline{y}}^{pc} = \sum_{k=1}^{24} \overline{y}_k^{pc}/24 = 141.777.$$

All three CHRON proportion estimates and SYSBP mean estimates are close. Next we calculate the JRR variance estimates. For a CHRON proportion estimator $\hat{p}$ the first variance estimate (5.20) is

$$\hat{v}_{1.jrr}(\hat{p}) = \sum_{h=1}^{24} (\hat{p}_h - 0.3976)^2 = 0.1099 \times 10^{-3},$$

and from the complement pseudosamples we obtain, using (5.21):

$$\hat{v}_{2.jrr}(\hat{p}) = \sum_{h=1}^{24} (\hat{p}_h^c - 0.3976)^2 = 0.1107 \times 10^{-3}.$$

The combined variance estimate (5.22) is thus

$$\hat{v}_{3.jrr}(\hat{p}) = (\hat{v}_{1.jrr}(\hat{p}) + \hat{v}_{2.jrr}(\hat{p}))/2 = 0.1103 \times 10^{-3}.$$

The second set (5.26)–(5.29) of JRR variance estimates is obtained by using the pseudovalues and the bias-corrected estimators. A counterpart of $\hat{v}_{1.jrr}$ is

$$\hat{v}_{4.jrr}(\hat{p}) = \sum_{h=1}^{24} (\hat{p}_h^p - 0.3972)^2 = 0.1060 \times 10^{-3},$$

and from the complement pseudosamples we have

$$\hat{v}_{5.jrr}(\hat{p}) = \sum_{h=1}^{24} (\hat{p}_h^{pc} - 0.3980)^2 = 0.1067 \times 10^{-3}.$$

The combined variance estimate is

$$\hat{v}_{6.jrr}(\hat{p}) = (\hat{v}_{4.jrr}(\hat{p}) + \hat{v}_{5.jrr}(\hat{p}))/2 = 0.1063 \times 10^{-3}.$$

From all the pseudosamples and their complements we obtain

$$\hat{v}_{7.jrr}(\hat{p}) = \sum_{h=1}^{24} (\hat{p}_h - \hat{p}_h^c)^2/4 = 0.1103 \times 10^{-3}.$$

The JRR variance estimates for the CHRON proportion estimator $\hat{p}$ are quite close, as expected. For the SYSBP mean estimator $\overline{y}$, we obtain the following JRR variance estimates:

$$\hat{v}_{1.jrr}(\overline{y}) = \sum_{h=1}^{24} (\overline{y}_h - 141.785)^2 = 0.2773,$$

$$\hat{v}_{2.jrr}(\overline{y}) = \sum_{h=1}^{24} (\overline{y}_h^c - 141.785)^2 = 0.2803,$$

$$\hat{v}_{3.jrr}(\overline{y}) = (\hat{v}_{1.jrr}(\overline{y}) + \hat{v}_{2.jrr}(\overline{y}))/2 = 0.2788,$$

$$\hat{v}_{4.jrr}(\overline{y}) = \sum_{h=1}^{24} (\overline{y}_h^{p} - 141.793)^2 = 0.2759,$$

$$\hat{v}_{5.jrr}(\overline{y}) = \sum_{h=1}^{24} (\overline{y}_h^{pc} - 141.777)^2 = 0.2789,$$

$$\hat{v}_{6.jrr}(\overline{y}) = (\hat{v}_{4.jrr}(\overline{y}) + \hat{v}_{5.jrr}(\overline{y}))/2 = 0.2774,$$

$$\hat{v}_{7.jrr}(\overline{y}) = \sum_{h=1}^{24} (\overline{y}_h - \overline{y}_h^{c})^2/4 = 0.2788.$$

For SYSBP, the JRR variance estimates of $\overline{y}$ are also very close. All the JRR variance estimators of a proportion estimator and a mean estimator provided closely related numerical results. Therefore, either practical or computational considerations can guide the selection of an appropriate JRR variance estimator. The jackknife technique is available in some software products for the analysis of complex surveys.

The BOOT Technique

Similar to the other sample reuse methods, the bootstrap can be used for variance approximation of a nonlinear estimator under a complex sampling design. The method, however, differs from BRR and JRR in many respects, e.g. the generation of pseudosamples is quite different. We consider the bootstrap technique for variance estimation of a ratio estimator under a two-stage stratified epsem design where a constant number of clusters (which may be greater than two) is drawn with replacement from each stratum. We adopt a simple version of the bootstrap, introduced in Rao and Wu (1988) as a *naive bootstrap*, for this kind of design, and call it the *BOOT technique*.

Let us assume that $m_h = a\ (\geq 2)$ clusters are drawn with replacement from each of the H strata. The number of sample clusters is thus $m = a \times H$. We construct the bootstrap pseudosamples in the following way:

Step 1. From the a sample clusters in stratum h, draw a simple random sample of size a with replacement. This is performed independently in each stratum. The resulting H simple random samples together constitute a *bootstrap sample* of m clusters.

Step 2. Repeating Step 1 K times, a total of K independent bootstrap samples are obtained.

It is important in Step 1 that the simple random samples in each stratum are drawn with replacement, and the stratum-wise samples are drawn independently.

So, a particular sample cluster in a stratum may be included in a bootstrap sample many (even a) times, or not at all.

We consider the BOOT technique for the estimation of the variance of the ratio estimator $\hat{r}$. A ratio estimator for a bootstrap sample k is denoted by $\hat{r}_k$ $(k = 1, \ldots, K)$. The mean of the bootstrap sample estimates $\hat{r}_k$ provides a *bootstrap estimator*

$$\overline{\hat{r}} = \sum_{k=1}^{K} \hat{r}_k / K. \tag{5.30}$$

A Monte Carlo variance estimator based on $\hat{r}_k$ and the bootstrap estimator (5.30) is first derived for the parent estimator $\hat{r}$:

$$\hat{v}_{mc}(\hat{r}) = \sum_{k=1}^{K} (\hat{r}_k - \overline{\hat{r}})^2 / K. \tag{5.31}$$

Unfortunately, this intuitively attractive variance estimator is unacceptable because it is not consistent for the variance of $\hat{r}$ and, moreover, it is not unbiased even for the variance of a linear estimator, as Rao and Wu (1988) have shown. But in the case considered, where a constant number of clusters is drawn from each stratum, an appropriately rescaled Monte Carlo variance estimator provides a consistent variance estimator for the parent estimator $\hat{r}$. Hence the first BOOT variance estimator is

$$\hat{v}_{1.boot}(\hat{r}) = \frac{a}{a-1} \hat{v}_{mc}(\hat{r}) = \frac{a}{a-1} \sum_{k=1}^{K} (\hat{r}_k - \overline{\hat{r}})^2 / K. \tag{5.32}$$

By using the parent estimator $\hat{r}$ in place of the bootstrap estimator, another variance estimator is obtained:

$$\hat{v}_{2.boot}(\hat{r}) = \frac{a}{a-1} \sum_{k=1}^{K} (\hat{r}_k - \hat{r})^2 / K. \tag{5.33}$$

It should be noticed that for the naive bootstrap there is no obvious solution to the scaling problem in the case in which the number of sample clusters per stratum varies. Rao and Wu (1988) derive a *rescaling bootstrap* for these cases, based on drawing simple random samples of size m_h (≥ 1) clusters with replacement from a stratum. With appropriate selection of m_h, different versions of the bootstrap are provided. Sitter (1992) proposes a generalization of this method, based on resampling without replacement rather than with replacement, and repeating this many times with replacement. Rao *et al.* (1992) redefine the rescaling bootstrap to be also suitable for variance estimation of non-smooth functions such as the median.

In the BOOT technique, to obtain variance estimation results with sufficient precision the number K of bootstrap samples should be large, preferably 500 to 1000. The technique thus requires large processing capabilities and can consume a lot of computer resources. In this, the BOOT technique is more obviously computer-intensive than BRR and JRR.

Example 5.4

The BOOT technique in the MFH Survey. We apply the BOOT technique for variance approximation of subpopulation proportion and mean estimators $\hat{p}$ (for CHRON) and $\overline{y}$ (for SYSBP), both considered as ratio estimators. The MFH Survey subgroup consists of 2699 males aged 30–64 years. In the MFH Survey design there are $H = 24$ strata each with $a = 2$ sample clusters, so each bootstrap sample constitutes of $m = 2 \times 24 = 48$ clusters. In the generation of the bootstrap samples we use the cluster-level data set. We obtain a bootstrap sample by drawing a simple random sample of two clusters with replacement, independently from each stratum. Thus, a cluster in a stratum can appear in a bootstrap sample either 0, 1 or 2 times so that the sample size from a stratum is always 2 clusters. Note that the number of such samples can become large; e.g. if we have 1000 bootstrap samples, a total of 24 000 independent samples of size 2 must be drawn. In this example, $K = 1000$ bootstrap samples.

An estimate $\hat{r}_k$ mimicking the parent estimator $\hat{r}$ is calculated from each of the K bootstrap samples. A bootstrap estimate is then calculated as an average of the $\hat{r}_k$. By using the $\hat{r}_k$, the bootstrap estimate and the parent estimate, we finally obtain BOOT variance estimates $\hat{v}_{1.boot}(\hat{r})$ and $\hat{v}_{2.boot}(\hat{r})$.

With $K = 1000$ bootstrap samples, the distribution of the bootstrap sample estimates for CHRON and SYSBP are displayed in Figure 5.1. The parent estimates and the bootstrap estimates (5.30) for CHRON proportion and SYSBP mean are

$$\hat{p} = 0.3976, \text{ and the bootstrap estimate is } \overline{\hat{p}} = 0.3973,$$
$$\overline{y} = 141.785, \text{ and the bootstrap estimate is } \overline{\hat{y}} = 141.783.$$

The BOOT variance estimates (5.32) and (5.33) for CHRON proportion $\hat{p}$ are, respectively

$$\hat{v}_{1.boot}(\hat{p}) = 2 \times \sum_{k=1}^{1000} (\hat{p}_k - 0.3973)^2/1000 = 0.1039 \times 10^{-3}$$

and

$$\hat{v}_{2.boot}(\hat{p}) = 2 \times \sum_{k=1}^{1000} (\hat{p}_k - 0.3976)^2/1000 = 0.1040 \times 10^{-3}.$$

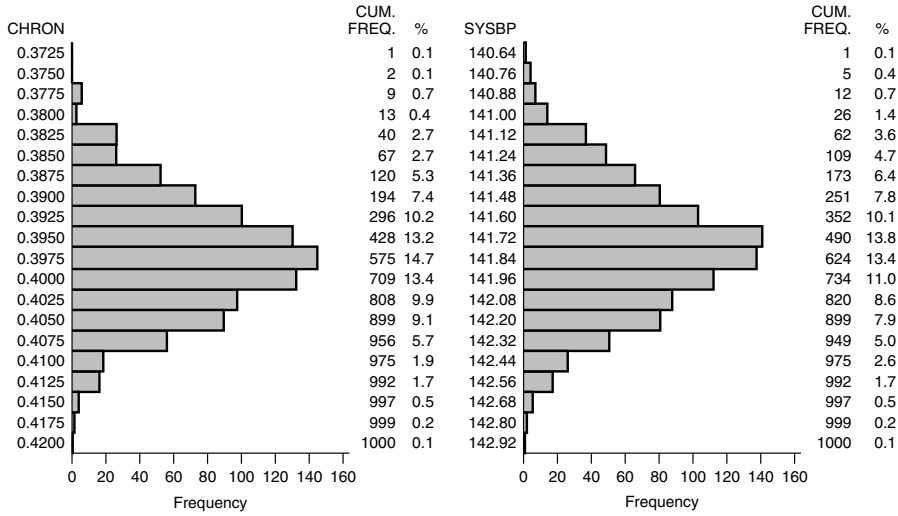


Figure 5.1 Bootstrap histograms for CHRON (a binary variable) and SYSBP (a continuous variable) from the bootstrap estimates $\hat{r}_k$ with $K = 1000$ bootstrap samples.

The BOOT variance estimates for SYSBP mean $\overline{y}$ are

$$\hat{v}_{1.boot}(\overline{y}) = 2 \times \sum_{k=1}^{1000} (\overline{y}_k - 141.783)^2/1000 = 0.2798$$

and

$$\hat{v}_{2.boot}(\overline{y}) = 2 \times \sum_{k=1}^{1000} (\overline{y}_k - 141.785)^2/1000 = 0.2798.$$

For a CHRON proportion estimator $\hat{p}$ and a SYSBP mean estimator $\overline{y}$, both BOOT variance estimates are approximately equal. As in the other reuse methods, any definite preference for the type of variance estimator has not been suggested. From a computational point of view, the estimator $\hat{v}_{2.boot}$ is simpler than $\hat{v}_{1.boot}$.

5.5 COMPARISON OF VARIANCE ESTIMATORS

The linearization method and sample reuse methods were used as basic approximation techniques for variance estimation of a nonlinear ratio estimator. It was assumed that the sample was from a two-stage epsem sampling design with at least two clusters drawn with replacement from each stratum. The linearization method was considered under a design with a varying number (≥ 2) of sample

clusters per stratum. Basic forms of the balanced half-samples (BRR) and jackknife repeated replications (JRR) techniques involved a design with exactly two sample clusters per stratum, and the number of strata is assumed to be large. Both the methods have been generalized for designs with a varying number (≥ 2) of sample clusters per stratum. The bootstrap technique was considered under a design in which a constant number (≥ 2) of clusters were drawn from each stratum. Also, the bootstrap has been generalized for the case of a varying number of sample clusters per stratum. Of the approximation methods, the bootstrap tends to require more computer resources. We next compare the numerical results obtained from the MFH Survey for variance approximation by the linearization and sample reuse techniques.

Comparison of Variance Estimates in the MFH Survey

Using linearization, BRR, JRR and BOOT techniques we estimated the variance of a subpopulation proportion estimator of a binary response CHRON (chronic morbidity), and the variance of a subpopulation mean estimator of a continuous response SYSBP (systolic blood pressure). Both estimators were ratio-type estimators for the MFH Survey subgroup that consisted of 2699 males aged 30–64 years. Detailed results were given in Examples 5.1–5.4. There were a total of 24 strata, each with two sample clusters in the MFH Survey sampling design, which therefore provides adequate data for demonstrating all the variance approximation methods. A cluster-level data set with 48 observations was used in all techniques.

Variance and design-effect estimates for a CHRON proportion $\hat{p}$ and a SYSBP mean $\overline{y}$ are displayed in Table 5.5. The design-effect estimator is of the form $\text{deff} = \hat{v}/\hat{v}_{srswr}$, where $\hat{v}$ is the variance estimator being considered and $\hat{v}_{srswr}$ is the variance estimator corresponding to simple random sampling with replacement.

For CHRON, the variance estimate from linearization, and the first three BRR and JRR estimates and the last one from the BRR and JRR techniques are all nearly equal. When compared with these estimates, the fourth, fifth and sixth BRR and JRR variance estimates, and both of the BOOT variance estimates, are somewhat smaller. Note that the linearization and the last BRR and JRR variance estimates (which could be taken as the most appropriate variance estimates) are equal. For SYSBP, all the BRR variance estimates are nearly equal, and the JRR estimates indicate larger variation. The BOOT variance estimates are somewhat larger than the others. For SYSBP, the linearization and the last BRR and JRR variance estimates are also nearly equal.

The design-effect estimates indicate a varying degree of intra-cluster correlation for CHRON and SYSBP. CHRON has noticeably less intra-cluster correlation than SYSBP. For SYSBP, the design-effect estimates indicate only a slight variation between techniques.

Table 5.5 Linearization, BRR, JRR, BOOT and SRSWR variance and design-effect estimates $\hat{v}$ and deff of a CHRON proportion estimate $\hat{p}$ and a SYSBP mean estimate $\overline{y}$ in the MFH Survey subgroup of 30–64-year-old males.

	Chronic morbidity		Systolic blood pressure	
Method	$10^{-3} \times \hat{v}(\hat{p})$	deff $(\hat{p})$	$\hat{v}(\overline{y})$	deff $(\overline{y})$
Linearization				
DES	0.1103	1.24	0.2788	2.06
Balanced repeated replications				
1	0.1104	1.24	0.2791	2.06
2	0.1103	1.24	0.2790	2.06
3	0.1103	1.24	0.2791	2.06
4	0.1052	1.18	0.2787	2.06
5	0.1056	1.19	0.2788	2.06
6	0.1054	1.19	0.2787	2.06
7	0.1103	1.24	0.2790	2.06
Jackknife repeated replications				
1	0.1099	1.24	0.2773	2.05
2	0.1107	1.25	0.2803	2.07
3	0.1103	1.24	0.2788	2.06
4	0.1060	1.19	0.2759	2.04
5	0.1067	1.20	0.2789	2.06
6	0.1063	1.20	0.2774	2.05
7	0.1103	1.24	0.2788	2.06
Bootstrap				
1	0.1039	1.17	0.2798	2.07
2	0.1040	1.17	0.2798	2.07
SRSWR	0.0888	1.00	0.1352	1.00

In conclusion, variance estimates of the ratio estimators obtained by the linearization, BRR, JRR and BOOT techniques do not differ significantly from each other, for both response variables. Therefore, software availability and other practical reasons might guide the selection of a technique in applications. For further training in the pseudoreplication methods, the reader is encouraged to use the facilities provided in the web extension of the book.

Other Properties of the Variance Approximation Methods

The variance approximation techniques based on linearization, BRR, JRR and the bootstrap have been evaluated in the literature by empirical investigations and

simulation studies, on more theoretical arguments. We briefly refer to some of the results.

Kish and Frankel (1974) empirically studied the relative performances of linearization, BRR and JRR under an epsem one-stage stratified design with two clusters drawn with replacement from each stratum. They showed first that for a linear estimator, the variance estimators coincided and were the same as a standard textbook variance estimator. Properties of the variance estimators were different for nonlinear estimators such as ratio estimators, regression coefficients and correlation coefficients. The linearization method provided the most stable variance estimates, whilst BRR gave the least stable, but none of the estimators gave an overall best performance when many criteria were considered. Kish and Frankel concluded that the linearization technique might be the best choice for ratio estimators, and sample reuse techniques for other nonlinear estimators.

Krewski and Rao (1981) showed that linearization, BRR and JRR have similar first-order asymptotic properties. Rao and Wu (1985) considered higher-order properties and showed that linearization and JRR provide equal second-order properties under a design in which two clusters are drawn with replacement from each stratum. Rao and Wu (1988) considered the bootstrap and showed that the first-order properties of their rescaling bootstrap variance estimator coincide with those of linearization, BRR and JRR. Second-order properties, however, differ. The rescaling bootstrap also indicated greater instability than either the linearization or the JRR. Rao *et al.* (1992) studied the performances of jackknife, BRR and bootstrap for variance estimation of the median and noticed no considerable differences between the methods.

5.6 THE OCCUPATIONAL HEALTH CARE SURVEY

In this section we describe the sampling design, data collection and properties of the available survey data of the Occupational Health Care Survey (OHC Survey). The sampling design of the OHC Survey is an example of stratified cluster sampling in which both one- and two-stage sampling are used. Thus, the OHC Survey sampling design is slightly more complex than that of the MFH Survey. Moreover, in the OHC Survey sampling design a large number of sample clusters are available, and the design produces noticeable clustering effects for several response variables. Therefore, this sampling design is very suitable for examining the effects of clustering in the analysis of complex surveys. The OHC Survey will be used for further examples given in Chapters 7 and 8.

In Finland, as in many industrialized countries, the provision of occupational health (OH) services is regulated by legislation. An Occupational Health Services Act came into force in 1979 to guide the development of OH services. All employers, with a few minor exceptions, would be required to provide OH services for their employees so that the activities would focus on the main work-related health hazards. Through the National Sickness Insurance Scheme, employers are

reimbursed by the Social Insurance Institution for a certain share of the costs of OH services. For employees, OH services are free of charge. Sample surveys have been carried out to evaluate the functioning of the OHC Act, with a major one, the OHC Survey, conducted in 1985.

Sampling Design

The OHC Survey can be characterized as a multi-purpose analytical sample survey similar to the MFH Survey. The OHC Survey was aimed at assessing implementation of the activities prescribed by the OHC Act, at discovering how well the essential goals of the legislation had been attained, and at defining how OH services could be further developed. The survey focused on establishments in all industries except farming and forestry, on the employers and employees, and on the units that provided the OH services for the sites surveyed. There were about 2 million employees and over 100 000 industrial establishments in the target populations.

In the study design, the industrial establishment was the primary unit of sampling and data collection. Because in Finland there are nationwide registers available for a sampling frame covering the target establishments, cluster sampling was a natural choice to be used with establishments as the clusters, i.e. primary sampling units. In contrast to the MFH sampling design, the principal motivation for cluster sampling in the OHC Survey was subject matter rather than cost efficiency.

Within the establishment sampling frame, the size of PSUs varied widely, from one-person workplaces to enterprises with a thousand or more workers. This property of varying cluster sizes should be taken into account when considering the person-level sample size for data collection. Therefore, the population of clusters was stratified by cluster size and by using two-stage sampling in strata that covered large sites. In addition to size, type of industry of establishment was used to form six explicit strata. One-stage sampling was used in strata covering establishments with a maximum of 100 employees; otherwise, two-stage sampling was used with approximately 50 employees sampled from each large site. This would produce an estimated total sample of about 17 000 employees in a sample of 1542 establishments. Stratum-wise allocation of the clusters, based on prior knowledge of their expected mean sizes, was carried out so that the employee sample would be nearly epsem, giving approximately equal inclusion probabilities for the employees. The sampling design is described in more detail in Lehtonen (1988).

Data Collection and Nonresponse

Structured questionnaires were used to collect data from employers, employees and OH units. During the data collection it turned out that a number of sample

establishments, mainly small ones, had closed down, and the final number of establishments for the appropriate questionnaire was 1362. The response rate was 88%. Furthermore, 82% (13 355) of the employees from the 1195 responding establishments completed the personnel questionnaire. Finally, 93% of the OH units of the responding establishments completed the appropriate questionnaire; this produced information on 760 out of a total of 816 establishments covered by OHC. The numbers of establishments and employees in the resulting survey data for each stratum are displayed in Table 5.6.

Analyses based on logit models indicated statistically significant variation in the response rates of the establishment questionnaire, depending on certain structural features of the establishments such as size, type of industry and organizational type. Predicted response rates for the appropriate questionnaire (based on a logit model with size, type of industry, organizational type and interaction of the two last mentioned as the model terms) are displayed in Figure 5.2. Small size, belonging to the construction industry, and having only a single site all increased the probability of nonresponse.

Nonresponse was quite low in large establishments and was independent of the type of industry or organizational type. It was also noted that establishments covered by OH services, and for which the regulations of the OHC Act were obligatory, responded most frequently to the appropriate questionnaire. Also, establishments for which the regulations of the Act were obligatory had an approximately equal response rate whether or not they were covered by OH services. Nonresponse was highest in those smallest single-site establishments that operated in the construction industry and were not covered by OH services.

Table 5.6 The number of establishments and employees by stratum in the OHC Survey data.

Stratum	Size	Number of		Average cluster sample size
		Establishments	Employees	
1	1–10	696	1730	2.5
2	11–100	176	4143	23.5
3	101–500	52	2396	46.1
4	501+	21	976	46.5
5	(all sizes)	109	1396	12.8
6	(all sizes)	141	2714	19.2
Total sample		1195	13 355	11.2

Type of industry:
Strata 1–4: All except those in strata 5 and 6
Stratum 5: Construction industry
Stratum 6: Public services

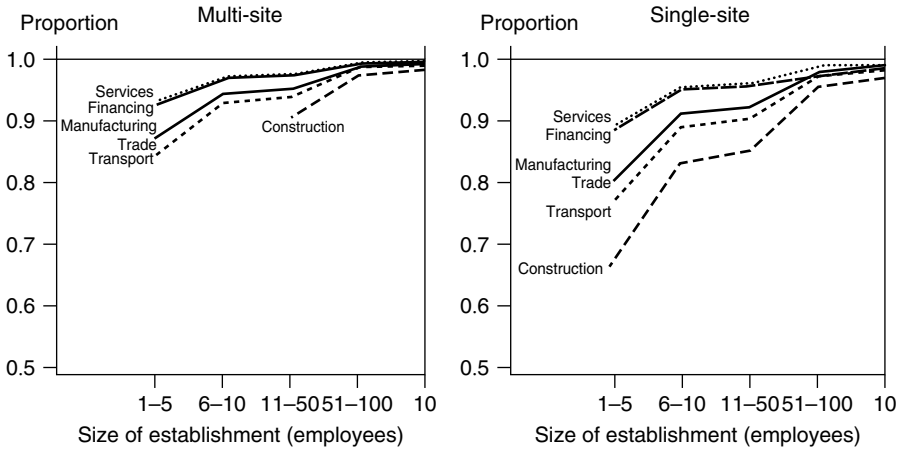


Figure 5.2 Predicted response rates in the establishment questionnaire (based on a logit model) by size and type of industry of establishment, in establishments of multi-site enterprises and in single-site establishments.

Weighting for nonresponse was required in the establishment-level analyses, for example, for the estimation of coverage of the OHC. The weight was constructed so that stratum-wise variation in inclusion probabilities of the PSUs was also compensated for. At the employee level the sampling design was nearly epsem, and the total number of employees at the small nonresponse establishments was relatively small. Therefore, adjustment for nonresponse in the element-level analyses was not so critical as at the cluster level. This was so, for example, in inferences concerning employee-level target populations on establishments covered by OH services.

Design Effects

A subgroup of establishments with a minimum of 10 employees will make up the OHC Survey data set used for demonstration purposes in examples. The data set includes a total of 250 clusters in 5 strata, and a total of 7841 employees. The data set can be regarded as approximately self-weighting. Cluster sample sizes in this subgroup vary from 10 to about 60 workers. Note that the subgroup is of a segregated classes type. These data, for selected response variables, are displayed in Table 5.7.

The number of sample clusters, i.e. establishments, is large (250) and this is favourable for covariance-matrix estimation. The sample establishments tend to be homogeneous with respect to certain subject-level response variables, resulting in positive intra-cluster correlations. For example, in a manufacturing

Table 5.7 The available OHC Survey data by sex and age of respondent, and proportions (%) of chronically ill persons (CHRON) and persons exposed to physical health hazards of work (PHYS), and the mean of the standardized first principal component of nine psychic (psychological or mental) symptoms (PSYCH).

		Sample		CHRON	PHYS	PSYCH
Sex	Age	*n*	%	%	%	Mean
Males		4485	57.2	29.3	46.0	−0.104
Females		3356	42.8	29.2	19.4	0.139
Males	15–24	504	6.4	15.5	52.8	−0.300
	25–34	1355	17.3	19.8	50.8	−0.160
	35–44	1453	18.5	27.1	42.9	−0.073
	45–54	847	10.8	44.2	41.9	−0.033
	55–64	326	4.2	61.3	39.3	0.102
Females	15–24	418	5.3	16.0	19.1	0.095
	25–34	993	12.7	18.9	18.9	0.132
	35–44	1002	12.8	26.5	17.9	0.104
	45–54	681	8.7	43.5	18.5	0.168
	55–64	262	3.3	61.8	29.4	0.301
Both sexes	15–24	922	11.8	15.7	37.5	−0.121
	25–34	2348	29.9	19.4	37.4	−0.036
	35–44	2455	31.3	26.9	32.7	−0.000
	45–54	1528	19.5	43.8	31.5	0.056
	55–64	588	7.5	61.6	34.9	0.191
Total sample		7841	100.0	29.2	34.6	0.000

firm, working conditions tend to be similar for most workers, these conditions being different from those of an office establishment, which in turn are also internally homogeneous. This produces design-effect estimates of means and proportions noticeably greater than one, especially for subject-level response variables measuring workplace-related matters such as physical or psycho-social working conditions. In some other variables, intra-cluster correlations were smaller, e.g. in variables describing overall psychic (psychological or mental) strain and psychosomatic symptoms. Design effects for selected response variables are displayed in Table 5.8.

The average design-effect estimates are noticeably large especially in response variables strongly associated with working conditions. The averages are closer to one in the variables that cannot be considered work-related. For further analyses, three response variables are selected: the variables PHYS (physical health hazards of work) and CHRON (chronic morbidity), which are binary, and the variable PSYCH (psychic strain), which is continuous. PHYS has strong intra-cluster correlation with a large overall design-effect estimate of 7.2. The

Table 5.8 Averages of design-effect estimates of proportion estimates of selected groups of binary response variables in the OHC Survey data set (number of variables in parentheses).

Study variable	Mean deff
Physical working conditions (12)	6.5
Psycho-social working conditions (11)	3.3
Psychosomatic symptoms (8)	2.0
Psychic symptoms (9)	1.8

overall design-effect estimates of CHRON and PSYCH are 1.8 and 2.0 respectively. Moreover, PHYS is apparently work-related; this is not as clear for CHRON and PSYCH.

5.7 LINEARIZATION METHOD FOR COVARIANCE-MATRIX ESTIMATION

Weighted Ratio Estimator

We previously considered the case of a single ratio estimator. A *vector* of ratio estimators consists of u ratio estimators, where $u \geq 2$ is the number of population subgroups called *domains*. The domains are formed by cross-classifying one or more categorical predictors such as sex, age group, socioeconomic factors, or regional variables. Our aim is to estimate consistently the domain ratio parameters and the corresponding covariance matrix of the ratio estimators under a given complex sampling design. For this, we construct a *weighted ratio estimator* to be used for the domain ratios. For a binary response variable, we work with weighted domain proportions, and for a continuous response variable we work with weighted domain means.

Let the population of N elements be divided into u non-overlapping subpopulations or domains. The unknown population ratio vector is a column vector denoted by $\mathbf{R} = (R_1, \ldots, R_u)'$. It consists of u domain ratio parameters $R_j = T_j/N_j$, where T_j denotes the population domain total of a response variable and N_j denotes the domain size, $\sum_{j=1}^{u} N_j = N$. In the binary case, the ratio parameter vector is denoted by $\mathbf{p} = (p_1, \ldots, p_u)'$, consisting of proportion parameters $p_j = N_{j1}/N_j$, where N_{j1} is the population total of a binary response variable in domain j. And in the continuous case, the parameter vector is denoted by $\overline{\mathbf{Y}} = (\overline{Y}_1, \ldots, \overline{Y}_u)'$, where $\overline{Y}_j$ are domain mean parameters $\overline{Y}_j = T_j/N_j$. A sample of n elements is drawn using stratified cluster sampling such that m_h clusters are drawn from each of the $h = 1, \ldots, H$ strata with a total $m = \sum_{h=1}^{H} m_h$ of sample clusters, where $H \geq 1$, $m \geq 2H$ and $m > u$. In two-stage cluster sampling, a sample of

n_{hi} elements is drawn from sample cluster i in stratum h, $\sum_{h=1}^{H}\sum_{i=1}^{m_h} n_{hi} = n$. If sampling is performed in one stage, all the elements of the selected sample clusters are taken in the element-level sample.

In complex surveys, epsem designs with an equal inclusion probability for each population element are often used because they are convenient for statistical analysis. We considered such designs in the previous sections of this chapter, and the MFH and OHC Survey sampling designs are taken as being epsem. In practice, however, element inclusion probabilities can vary between the strata, and, even in epsem designs, reweighting may be necessary to adjust for nonresponse to attain consistent estimation. Also to cover these cases, we derive a weighted ratio estimator, which is more generally applicable than that previously considered for epsem samples.

For a self-weighting data set, an epsem sampling design is required and unit nonresponse is considered ignorable. If the data set is not self-weighting, an appropriate weight variable should be generated for statistical analyses. A weight variable assigns a positive value for each element of the data set such that unequal element inclusion probabilities and nonresponse are adjusted. Basically, as shown in Chapter 2, the weight w_k for a sample element k is $w_k = 1/\pi_k$ i.e. the reciprocal of the inclusion probability. And in Chapter 4, a weight $w_k^* = 1/(\pi_k\hat{\theta}_k)$ was introduced, where $\hat{\theta}_k$ is an estimated response probability. In epsem designs, π_k is a constant π for all population elements. In *non-epsem* designs, unequal inclusion probabilities may arise, for example, due to non-proportional allocation. For nonresponse adjustment, the sample data set can be divided into a number of adjustment cells, and the response rate θ_c is assumed constant within cell c but is allowed to vary between the cells. The cells are formed using auxiliary variables, which are also available for nonresponse cases. When using poststratification, adjustment cells are formed by using auxiliary information on the population level (see Sections 3.3 and 5.1 and Chapter 4). Note that weight is a constant for all elements in a self-weighting data set because π_k and $\hat{\theta}_k$ are constants.

As shown in Section 5.1, there are two main approaches for a weight variable. In a descriptive survey in which the population total on a study variable is estimated, a weight variable is constructed such that the sum of all n element weights w_k^* provides a consistent estimate $\hat{N}$ of the population size N. This type of weighting was extensively used in Chapters 2 to 4. In analytical surveys, where such totals are rarely estimated, it is customary to rescale the weights so that their sum equals the size n of the available sample data set. Although either kind of a weight variable can be used in software available for survey analysis, rescaled weights w_k^{**}, that sum up to n, are often more convenient for statistical analyses requiring a weight variable.

When using weights w_k^*, a vector $\hat{\mathbf{r}} = (\hat{r}_1, \ldots, \hat{r}_u)'$ of combined ratio estimators is constructed consisting of domain ratio estimators $\hat{r}_j = \hat{t}_j/\hat{N}_j$, where $\hat{t}_j$ is a weighted total estimator of the population total T_j of the response variable in domain j and $\hat{N}_j$ is the weighted size of domain j, and $\sum_{j=1}^{u}\hat{N}_j = \hat{N}$, the sum of all

n sample weights. As a result, the weighted estimators t_j and $\hat{N}_j$ are consistent for the corresponding population analogues T_j and N_j, so the domain ratio estimator $\hat{r}_j$ is consistent for the domain ratio R_j in a given complex sampling design.

The weighted totals $\hat{t}_j$ and $\hat{N}_j$ in the previous domain ratio estimators $\hat{r}_j$ are scaled to sum to the population level. For analytical purposes, we rescale the weights so that they sum to n, the size of the sample data set. Thus, to derive an estimator $\hat{r}_j$ we use the scaled weighted analogues y_j and x_j of $\hat{t}_j$ and $\hat{N}_j$ such that $y_j = (n/\hat{N})\hat{t}_j$ and $x_j = (n/\hat{N})\hat{N}_j$ with $\sum_{j=1}^{u} x_j = n$. The domain ratio estimator $\hat{r}_j$ can thus be written in the form

$$\hat{r}_j = \frac{y_j}{x_j} = \frac{\sum_{h=1}^{H}\sum_{i=1}^{m_h} y_{jhi}}{\sum_{h=1}^{H}\sum_{i=1}^{m_h} x_{jhi}} = \frac{\sum_{h=1}^{H}\sum_{i=1}^{m_h}\sum_{k=1}^{x_{hi}} w_{jhik}^{**} y_{jhik}}{\sum_{h=1}^{H}\sum_{i=1}^{m_h}\sum_{k=1}^{x_{hi}} w_{jhik}^{**}}, \quad j = 1, \ldots, u, \tag{5.34}$$

where y_{jhi} is the weighted sample sum of the response variable for the elements falling in domain j in sample cluster i of stratum h, and x_{jhi} is the corresponding weighted domain sample size. The rescaled weights w_{jhik}^{**} in (5.34) therefore sum up to n.

For a binary response, the ratio estimator $\hat{\mathbf{r}}$ with elements of the form (5.34) is a proportion estimator vector denoted by $\hat{\mathbf{p}} = (\hat{p}_1, \ldots, \hat{p}_u)'$, which consists of domain ratio estimators $\hat{p}_j = y_j/x_j = \hat{n}_{j1}/\hat{n}_j$, where $\hat{n}_{j1}$ is the weighted sample sum of the binary response for sample elements belonging to the domain j and $\hat{n}_j$ is the weighted domain size such that $\sum_{j=1}^{u} \hat{n}_j = n$. Under an epsem design and, moreover, if the data set is self-weighting, a simple unweighted estimator $\hat{\mathbf{p}}^U = (\hat{p}_1^U, \ldots, \hat{p}_u^U)'$ of $\mathbf{p}$ is obtained, where $\hat{p}_j^U = n_{j1}/n_j$ is a consistent estimator of the domain parameter p_j, n_{j1} is the sample sum of the binary response in domain j and n_j is the corresponding domain sample size such that $\sum_{j=1}^{u} n_j = n$. In this case, $\hat{\mathbf{p}}$ and $\hat{\mathbf{p}}^U$ coincide. Note that if the data set is not self-weighting, the estimator $\hat{\mathbf{p}}^U$ is not consistent for $\mathbf{p}$.

For a continuous response variable, we denote the weighted ratio estimator vector $\overline{\mathbf{y}} = (\overline{y}_1, \ldots, \overline{y}_u)'$, where the domain sample means $\overline{y}_j = y_j/x_j$ are consistent for the corresponding population domain means $\overline{Y}_j = T_j/N_j$. The corresponding unweighted counterpart is $\overline{\mathbf{y}}^U = (\overline{y}_1^U, \ldots, \overline{y}_u^U)'$.

It may be noted that the data actually needed for the ratio estimators $\hat{r}_j$ consist of m cluster-level scaled weighted sample sums y_{jhi} and x_{jhi}. Indeed, the analysis of such data can be performed by using the cluster-level data set of size m and access to the element-level data set of size n is not necessarily required. In practice, however, when using software for survey analysis, the weighted sample sums y_j and x_j are estimated from an element-level data set using the rescaled element weights w_{jhik}^{**}.

Covariance-matrix Estimation

The unknown population covariance matrix $\mathbf{V}/n$ of the ratio estimator vector $\hat{\mathbf{p}}$ has u rows and u columns, thus it is a $u \times u$ matrix. $\mathbf{V}/n$ is symmetric such that the lower and upper triangles of the matrix are identical. Variances of the domain ratio estimators are placed on the main diagonal of $\mathbf{V}/n$ and covariances of the corresponding domain ratio estimators on the off-diagonal part of the matrix. There is a total of $u \times (u+1)/2$ distinct parameters in $\mathbf{V}/n$ that need to be estimated.

The variance and covariance estimators $\hat{v}_{des}(\hat{r}_j)$ and $\hat{v}_{des}(\hat{r}_j, \hat{r}_l)$, being respectively the diagonal and off-diagonal elements of a consistent covariance-matrix estimator $\hat{\mathbf{V}}_{des}$ of the asymptotic covariance matrix $\mathbf{V}/n$ of the ratio estimator vector $\hat{\mathbf{r}} = (\hat{r}_1, \ldots, \hat{r}_u)'$, are derived using the linearization method considered in Section 5.3. The variance and covariance estimators of the sample sums y_j and x_j in a variance estimator $\hat{v}_{des}(\hat{r}_j)$ of $\hat{r}_j = y_j/x_j$, and the covariance estimators of the sample sums y_j, y_l, x_j and x_l in the covariance estimators $\hat{v}_{des}(\hat{r}_j, \hat{r}_l)$ of $\hat{r}_j$ and $\hat{r}_l$ in separate domains, are straightforward generalizations of the corresponding variance and covariance estimators given in Section 5.3 for the variance estimator of a single ratio estimator $\hat{r}$. We therefore do not show these formulae.

Like the scalar case, the variance and covariance estimators of $\hat{r}_j$ and $\hat{r}_l$ are based on the with-replacement assumption and the variation accounted for is the between-cluster variation. This causes bias in the estimates, but the bias can be assumed to be negligible if the first-stage sampling fraction is small.

The variance and covariance estimators of y_j, x_j, y_l and x_l are finally collected into the corresponding $u \times u$ covariance-matrix estimators $\hat{\mathbf{V}}_{yy}$, $\hat{\mathbf{V}}_{xx}$ and $\hat{\mathbf{V}}_{yx}$. Using these estimators, the design-based covariance-matrix estimator of $\hat{\mathbf{r}}$ based on the linearization method is given by

$$\begin{aligned}\hat{\mathbf{V}}_{des} = \text{diag}(\hat{\mathbf{r}})(&\mathbf{Y}^{-1}\hat{\mathbf{V}}_{yy}\mathbf{Y}^{-1} + \mathbf{X}^{-1}\hat{\mathbf{V}}_{xx}\mathbf{X}^{-1} \\ &- \mathbf{Y}^{-1}\hat{\mathbf{V}}_{yx}\mathbf{X}^{-1} - \mathbf{X}^{-1}\hat{\mathbf{V}}_{xy}\mathbf{Y}^{-1})\text{diag}(\hat{\mathbf{r}}),\end{aligned} \tag{5.35}$$

where

$\text{diag}\,(\hat{\mathbf{r}}) = \text{diag}\,(\hat{r}_1, \ldots, \hat{r}_u) = \text{diag}\,(y_1/x_1, \ldots, y_u/x_u)$
$\mathbf{Y} = \text{diag}\,(\mathbf{y}) = \text{diag}\;\,(y_1, \ldots, y_u)$
$\mathbf{X} = \text{diag}\,(\mathbf{x}) = \text{diag}(x_1, \ldots, x_u)$
$\hat{\mathbf{V}}_{yy}$ is the covariance-matrix estimator of the sample sums y_j and y_l
$\hat{\mathbf{V}}_{xx}$ is the covariance-matrix estimator of the sample sums x_j and x_l
$\hat{\mathbf{V}}_{yx}$ is the covariance-matrix estimator of the sums y_j and x_l, and
$\hat{\mathbf{V}}_{xy} = \hat{\mathbf{V}}'_{yx}$

and the operator 'diag' generates a diagonal matrix with the elements of the corresponding vector as the diagonal elements and with off-diagonal elements equal to zero. Note that in a linear case, all elements of the covariance-matrix estimators $\hat{\mathbf{V}}_{xx}$, $\hat{\mathbf{V}}_{yx}$ and $\hat{\mathbf{V}}_{xy}$ are zero.

In the estimation of the elements of $\hat{\mathbf{V}}_{des}$, at least two clusters are assumed to be drawn with replacement from each of the H strata. In the special case of $m_h = 2$ clusters routinely used in survey sampling, the estimators can be simplified in a manner similar to that done in Section 5.3.

As a simple example, let the number of domains be $u = 2$. The elements of the covariance-matrix estimator

$$\hat{\mathbf{V}}_{des} = \begin{bmatrix} \hat{v}_{des}(\hat{r}_1) & \hat{v}_{des}(\hat{r}_1, \hat{r}_2) \\ \hat{v}_{des}(\hat{r}_2, \hat{r}_1) & \hat{v}_{des}(\hat{r}_2) \end{bmatrix}$$

are the following:
Variance estimator:

$$\hat{v}_{des}(\hat{r}_j) = \hat{r}_j^2(y_j^{-2}\hat{v}(y_j) + x_j^{-2}\hat{v}(x_j) - 2(y_j x_j)^{-1}\hat{v}(y_j, x_j)), \quad j = 1, 2.$$

Covariance estimator:

$$\begin{aligned}\hat{v}_{des}(\hat{r}_1, \hat{r}_2) = \hat{r}_1\hat{r}_2((y_1y_2)^{-1}\hat{v}(y_1, y_2) + (x_1x_2)^{-1}\hat{v}(x_1, x_2) \\ - (y_1x_2)^{-1}\hat{v}(y_1, x_2) - (y_2x_1)^{-1}\hat{v}(y_2, x_1)).\end{aligned}$$

The estimator $\hat{v}_{des}(\hat{r}_2, \hat{r}_1)$ is equal to $\hat{v}_{des}(\hat{r}_1, \hat{r}_2)$ because of symmetry of $\hat{\mathbf{V}}_{des}$. If the estimators $\hat{r}_j$ are taken as linear estimators, then the denominators x_j are assumed fixed. In this case, the variance and covariance estimates $\hat{v}(x_j)$ and $\hat{v}(y_j, x_j)$ are zero, and $\hat{v}_{des}(\hat{r}_j) = \hat{v}(y_j)/x_j^2$. And for a binary response in the binomial case, this estimator reduces to $\hat{v}_{bin}(\hat{p}_j) = \hat{p}_j(1 - \hat{p}_j)/n_j$.

It is important to note that $\hat{\mathbf{V}}_{des}$ is distribution-free so that it requires no specific distributional assumptions about the sampled observations. This allows an estimate $\hat{\mathbf{V}}_{des}$ to be nondiagonal. The nondiagonality of $\hat{\mathbf{V}}_{des}$ is because the ratio estimators $\hat{r}_j$ and $\hat{r}_l$ from distinct domains can have nonzero correlations. In contrast, the binomial covariance-matrix estimators considered in this section have zero correlation by definition.

One source of nonzero correlation of the estimators $\hat{r}_j$ and $\hat{r}_l$ from separate domains comes from the clustering of the sample. Varying degrees of correlation can be expected depending on the type of the domains. If the domains cut smoothly across the sample clusters, distinct members in a given sample cluster may fall in separate domains j and l such as cross-classes like demographic or related factors. Large correlations can then be expected if the clustering effect is noticeable. In contrast, if the domains are totally segregated in such a way that all members of a given sample cluster fall in the same domain, zero correlations of distinct estimates $\hat{r}_j$ and $\hat{r}_l$ are obtained. This happens if the predictors used in forming the domains are cluster-specific unlike cross-classes where factors are essentially individual-specific. If, for example, households are clusters, typical cluster-specific factors are net income of the household and family size, whereas age and sex of a family member are individual-specific. Mixed-type domains, often met in practice, are intermediate, so that nonzero correlations are present in some dimensions of the table with zero correlations in the others.

Detecting Instability

The covariance-matrix estimator (5.35) is consistent for the asymptotic covariance matrix $\mathbf{V}/n$ under the given complex sampling design so that, with a fixed cluster sample size, it is assumed to converge to $\mathbf{V}/n$ by increasing the number m of sample clusters. But with small m, an estimate $\hat{\mathbf{V}}_{des}$ can become unstable, i.e. near-singular. This can also happen if the number of domains u is large, which may require the estimation of several hundred distinct variance and covariance terms. The instability of a covariance-matrix estimate causes numerical problems when the inverse of the matrix is formed, which can severely disturb the reliability of testing and modelling procedures.

A near-singularity or *instability problem* is present if the degrees of freedom f for the estimation of the asymptotic covariance matrix $\mathbf{V}/n$ are small. For standard complex sampling designs, f can be taken as the number of sample clusters less the number of strata, i.e. $f = m - H$. A stable $\hat{\mathbf{V}}_{des}$ can be expected if f is large relative to the number u of domains or, more specifically, relative to the residual degrees of freedom of the model to be fitted. In practice, instability problems are not expected if a large number of sample clusters are available, and if u is also much smaller than m.

The statistic *condition number* can be used as a measure of instability of $\hat{\mathbf{V}}_{des}$. It is defined as the ratio $\text{cond}(\hat{\mathbf{V}}_{des}) = \hat{\lambda}_{\max}/\hat{\lambda}_{\min}$, where $\hat{\lambda}_{\max}$ and $\hat{\lambda}_{\min}$ are the largest and smallest eigenvalues of $\hat{\mathbf{V}}_{des}$ respectively. If this statistic is large, e.g. in hundreds or thousands, an instability problem is present. If the statistic is small, e.g. less than 50, no serious instability problems can be expected. Unfortunately, this statistic is not a routine output in software products from survey analysis. In the following table, condition numbers of $\hat{\mathbf{V}}_{des}$ with various values of u are displayed for the proportion estimator vector of the binary response variable CHRON (chronic morbidity) from the MFH and OHC Survey designs. The domains for each survey are formed by the sex of respondent and equal-sized age groups.

No. of domains	MFH	OHC
4	6.5	2.8
8	10.6	3.5
12	39.8	3.6
20	421.5	5.6
24	423 684	6.6
40	n.a.	9.9

n.a. not available

Note that in the MFH Survey $f = 24$, and in the OHC Survey $f = 245$. Therefore, in the MFH Survey, the largest possible value of u is 24, and with this value the corresponding $\hat{\mathbf{V}}_{des}$ becomes very unstable. With values of u less than 12 the

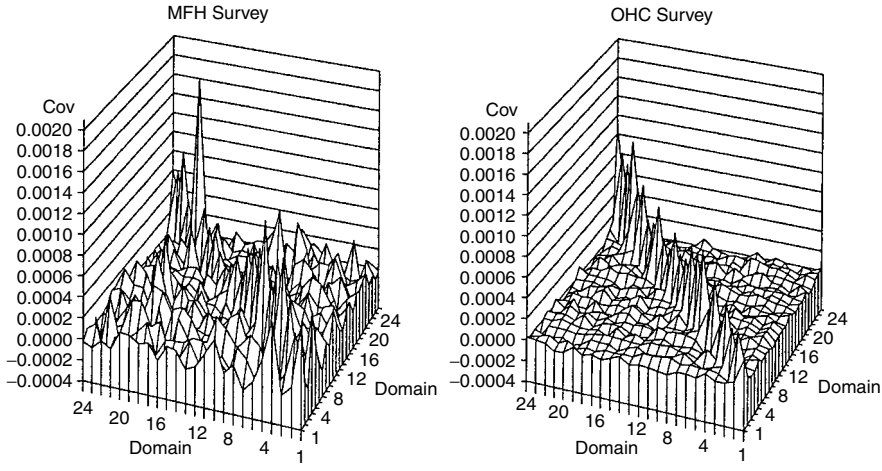


Figure 5.3 The covariance-matrix estimates $\hat{\mathbf{V}}_{des}$ of $u = 24$ domain proportion estimates of CHRON in the MFH and OHC Survey designs.

estimate remains quite stable. In the OHC Survey, condition numbers slightly increase with increasing u, but $\hat{\mathbf{V}}_{des}$ indicates stability with all values of u. These properties of the covariance-matrix estimates $\hat{\mathbf{V}}_{des}$ can also be depicted graphically. In Figure 5.3, the estimates $\hat{\mathbf{V}}_{des}$ for CHRON proportions with $u = 24$ domains from the MFH and OHC Survey designs are displayed. For the MFH Survey, the instability in $\hat{\mathbf{V}}_{des}$ is indicated by high 'peaks' in the off-diagonal part of the matrix. The stability of $\hat{\mathbf{V}}_{des}$ in the OHC Survey design is also clearly seen.

Design-effects Matrix Estimator

For a *design-effects matrix* estimator, we derive the binomial covariance-matrix estimator of a proportion estimator vector. A design-effects matrix is obtained using the binomial and the corresponding design-based covariance-matrix estimators. Design-effect estimators taken from the diagonal of the design-effects matrix are used to derive the covariance-matrix estimators that account for *extra-binomial variation*.

For the construction of a design-effects matrix estimator we need not only the design-based covariance-matrix estimator of the proportion vector but also the binomial counterpart. For a binary response, we assume a binomial sampling model for a proportion vector $\hat{\mathbf{p}}$ so that the weighted number of successes in each domain j is assumed to be generated by a binomial distribution and the generation processes are assumed independent between the u domains. The covariance-matrix estimator $\hat{\mathbf{V}}_{bin}(\hat{\mathbf{p}})$ of a proportion estimator $\hat{\mathbf{p}}$ is a diagonal matrix with

diagonal elements derived from the binomial distribution, given by

$$\hat{v}_{bin}(\hat{p}_j) = \hat{p}_j(1 - \hat{p}_j)/\hat{n}_j, \quad j = 1, \ldots, u. \tag{5.36}$$

For the unweighted proportion vector $\hat{\mathbf{p}}^U$, the corresponding estimate, denoted by $\hat{\mathbf{V}}_{bin}(\hat{\mathbf{p}}^U)$, is obtained by using (rescaled) element weights equal to one. It should be emphasized that, in the denominator of the binomial variance estimate (5.36) the weighted number of observations $\hat{n}_j$ is used, i.e. an expected sample size for the jth domain. An observed domain sample size n_j could be used in the denominator instead of the expected one.

Using the design-based covariance-matrix estimator $\hat{\mathbf{V}}_{des}(\hat{\mathbf{p}})$ and the binomial counterpart $\hat{\mathbf{V}}_{bin}(\hat{\mathbf{p}})$, the corresponding design-effects matrix estimator is derived for the domain proportion estimator vector $\hat{p}$, given by

$$\hat{\mathbf{D}} = \hat{\mathbf{V}}_{bin}^{-1}(\hat{\mathbf{p}})\hat{\mathbf{V}}_{des}(\hat{\mathbf{p}}), \tag{5.37}$$

where $\hat{\mathbf{V}}_{bin}^{-1}$ is the inverse of $\hat{\mathbf{V}}_{bin}$. The design-effect estimators $\hat{d}_j$ of $\hat{p}_j$ are the diagonal elements of the design-effects matrix estimator, hence the name *design-effects matrix*. The eigenvalues $\hat{\delta}_j$ of the design-effects matrix are often called the *generalized design-effects*. The sum of the design-effects estimates equals the sum of the eigenvalues, whose sum can be obtained from the sum of the diagonal elements of $\hat{\mathbf{D}}$, i.e. its trace. And the design-effect estimates and the corresponding eigenvalues are equal only in the special case where the estimate $\hat{\mathbf{V}}_{des}$ is also diagonal. All this holds in the case where the first covariance-matrix estimate in (5.37) is a diagonal matrix, such as $\hat{\mathbf{V}}_{bin}$. But in more complicated situations with proportions, where this is not true, the design-effects are not the diagonal elements of $\hat{\mathbf{D}}$ nor is the sum of design-effects equal to the sum of the eigenvalues. These more complicated design-effects matrices are sometimes called *generalized design-effects matrices* and will be discussed in Chapters 7 and 8.

The design-effect estimators of the proportion estimators $\hat{p}_j$ are of the form

$$\hat{d}_j = \hat{v}_{des}(\hat{p}_j)/\hat{v}_{bin}(\hat{p}_j), \quad j = 1, \ldots, u, \tag{5.38}$$

where the variance estimators $\hat{v}_{des}$ are diagonal elements of $\hat{\mathbf{V}}_{des}$. The design-effect estimates $\hat{d}_j$ measure the extra-binomial variation in the proportion estimates $\hat{p}_j$ due to the effect of clustering. Extra-binomial variation is present if design-effect estimates are greater than one.

If in the binomial variance estimate in (5.38) an observed domain sample size is used instead of an expected one, different design-effect estimates can be obtained. This is especially so if expected and observed domain sample sizes, $\hat{n}_j$ and n_j, deviate considerably, as can happen, e.g. due to non-proportionate sample allocation. Thus, design-effect estimates for subgroup proportion estimates calculated with a certain software package can differ from those obtained using

another. Obviously, in self-weighting samples both approaches should yield equal design-effect estimates.

It should be noted that, in the design-effects matrix estimator (5.37) only the contribution of the clustering is accounted for, because a binomial covariance-matrix estimator of the consistent weighted proportion estimator vector is used. By using in (5.37) a binomial covariance-matrix estimator of the unweighted proportion estimator vector instead of that of the weighted proportion estimator vector, all the contributions of complex sampling on covariance-matrix estimation are reflected, such as unequal inclusion probabilities, clustering and adjustment for nonresponse. Obviously, both approaches give similar design-effect matrix estimates when working with self-weighting samples. If adopting as a rule the use of a consistent proportion estimator $\hat{\mathbf{p}}$, then working with weighted observations, and thus with (5.37) would be reasonable. Then, the crucial role of adjusting for the clustering effect in the analysis of complex surveys would also be emphasized. However, the calculation of the deff matrix estimate by using both versions of the binomial covariance-matrix estimate can be useful in assessing the contribution of weighting to the design effects.

Example 5.5

Covariance-matrix and design-effects matrix estimation with the linearization method. Using the OHC Survey data we carry out a detailed calculation of the covariance-matrix estimate $\hat{\mathbf{V}}_{des}$ of a proportion estimate $\hat{\mathbf{p}}$ of the binary response PHYS (physical health hazards of work), and of a mean estimate $\overline{\mathbf{y}}$ of the continuous response PSYCH (the first standardized principal component of nine psychic symptoms), in the simple case of $u = 2$ domains formed by the variable sex. $\hat{\mathbf{V}}_{des}$ is thus a 2×2 matrix, and the domains are of a cross-class type. A part of the data set needed for the covariance-matrix estimation is displayed in Table 5.9. Note that these data are cluster-level, consisting of $m = 250$ clusters in five strata. Thus, the degrees of freedom $f = 245$. The employee-level sample size is $n = 7841$.

The ratio estimator is $\hat{\mathbf{r}} = (\hat{r}_1, \hat{r}_2)' = (y_1/x_1, y_2/x_2)'$, where $\hat{r}_1$ and $\hat{r}_2$ are given by (5.34). For the binary response PHYS, we denote the ratio estimator as $\hat{\mathbf{p}} = (\hat{p}_1, \hat{p}_2)'$, and for the continuous response PSYCH $\overline{\mathbf{y}} = (\overline{y}_1, \overline{y}_2)'$. The following figures for PHYS are calculated from Table 5.9.

Sums of the cluster-level sample sums $y_{jhi} (= y_{ji})$ and $x_{jhi} (= x_{ji})$:

$$\hat{n}_{11} = y_1 = 2061 \quad \text{and} \quad \hat{n}_1 = x_1 = 4485 \text{ (males)},$$
$$\hat{n}_{21} = y_2 = 650 \quad \text{and} \quad \hat{n}_2 = x_2 = 3356 \text{ (females)}.$$

Proportion estimates for PHYS, i.e. the elements of $\hat{\mathbf{p}} = (\hat{p}_1, \hat{p}_2)'$:

$$\hat{p}_1 = y_1/x_1 = 2061/4485 = 0.4595 \text{ (males)},$$

Table 5.9 Cluster-level sample sums y_{1i} (males) and y_{2i} (females) of the response variables PHYS and PSYCH with the corresponding cluster sample sizes x_{1i} (males) and x_{2i} (females) in sample clusters $i = 1, \ldots, 250$ in two domains formed by sex (the OHC Survey).

Stratum	Cluster	PHYS		PSYCH			
h	i	y_{1i}	y_{2i}	y_{1i}	y_{2i}	x_{1i}	x_{2i}
2	1	11	3	−0.1434	−0.0322	36	22
2	2	18	4	−0.1925	0.1867	57	21
2	3	4	5	0.0045	0.3674	9	15
2	4	2	2	0.7135	−0.3679	12	15
2	5	1	0	−0.1681	0.1235	27	8
2	6	1	0	−0.2673	0.1504	19	21
2	7	9	4	0.0099	0.2099	23	27
2	8	4	2	0.3681	0.0155	16	31
2	9	0	0	−0.5033	0.0755	6	6
2	10	3	0	−0.3176	−0.2516	8	8
2	11	2	7	0.9746	0.1903	6	67
2	12	7	3	−0.3361	0.5572	22	31
2	13	4	1	−0.2329	−0.2181	9	7
2	14	0	0	−0.2032	0.5893	13	16
2	15	1	23	0.4137	0.2565	4	56
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
6	245	14	2	0.1984	−0.4271	23	7
6	246	2	1	−0.1049	0.3905	7	7
6	247	4	7	−0.2961	0.5018	7	13
6	248	0	1	−0.8073	0.9278	3	9
6	249	2	0	0.0006	−0.3484	16	13
6	250	13	1	−0.1273	−0.1466	26	4
Total sample		2061	650	−26.7501	33.7983	4485	3356

and

$$\hat{p}_2 = y_2/x_2 = 650/3356 = 0.1937 \text{ (females)}.$$

We next construct the diagonal 2×2 matrices $\text{diag}(\hat{\mathbf{p}})$, $\mathbf{Y}$ and $\mathbf{X}$ for the calculation of the estimate $\hat{\mathbf{V}}_{des}$ for the PHYS proportion estimator $\hat{\mathbf{p}}$:

$$\text{diag}(\hat{\mathbf{p}}) = \begin{bmatrix} 0.4595 & 0 \\ 0 & 0.1937 \end{bmatrix}, \quad \mathbf{Y} = \begin{bmatrix} 2061 & 0 \\ 0 & 650 \end{bmatrix}$$

and

$$\mathbf{X} = \begin{bmatrix} 4485 & 0 \\ 0 & 3356 \end{bmatrix}.$$

The covariance-matrix estimates $\hat{\mathbf{V}}_{yy}$, $\hat{\mathbf{V}}_{xx}$ and $\hat{\mathbf{V}}_{yx}$, also obtained from the cluster-level data displayed in Table 5.9, are the following:

$$\hat{\mathbf{V}}_{yy} = \begin{bmatrix} 15\,722.50 & -130.45 \\ -130.45 & 3261.71 \end{bmatrix},$$

$$\hat{\mathbf{V}}_{xx} = \begin{bmatrix} 34\,560.23 & -7315.43 \\ -7315.43 & 34\,099.04 \end{bmatrix},$$

and

$$\hat{\mathbf{V}}_{yx} = \begin{bmatrix} 18\,973.88 & -5907.69 \\ -1098.11 & 6051.14 \end{bmatrix} = \hat{\mathbf{V}}'_{xy}.$$

By using these matrices we finally calculate for PHYS proportions the covariance-matrix estimate $\hat{\mathbf{V}}_{des}$ given by (5.35). Hence we have

$$\hat{\mathbf{V}}_{des} = \begin{bmatrix} \hat{v}_{des}(\hat{p}_1) & \hat{v}_{des}(\hat{p}_1, \hat{p}_2) \\ \hat{v}_{des}(\hat{p}_2, \hat{p}_1) & \hat{v}_{des}(\hat{p}_2) \end{bmatrix} = 10^{-4} \begin{bmatrix} 2.775 & 0.576 \\ 0.576 & 1.951 \end{bmatrix}.$$

For example, using the estimates calculated, the variance estimate $\hat{v}_{des}(\hat{p}_1)$ is obtained as

$$\begin{aligned} \hat{v}_{des}(\hat{p}_1) = 0.4595^2 \times (2061^{-2} \times 15\,722.50 + 4485^{-2} \times 34\,560.23 \\ - 2 \times (2061 \times 4485)^{-1} \times 18\,973.88) = 0.2775 \times 10^{-3}. \end{aligned}$$

Correlation of $\hat{p}_1$ and $\hat{p}_2$ is 0.25, which is quite large and indicates that the domains actually constitute cross-classes. The condition number of $\hat{\mathbf{V}}_{des}$ is $\text{cond}(\hat{\mathbf{V}}_{des}) = 1.9$, indicating stability of the estimate owing to a large f and small u.

For PSYCH, the following figures are calculated from Table 5.9.
Sums of the cluster-level sample sums y_{jhi} and x_{jhi}:

$$y_1 = -26.7501 \quad \text{and} \quad x_1 = 4485 \text{ (males)},$$
$$y_2 = \quad 33.7983 \quad \text{and} \quad x_2 = 3356 \text{ (females)}.$$

Mean estimates for PSYCH, i.e. the elements of $\overline{\mathbf{y}} = (\overline{y}_1, \overline{y}_2)'$:

$$\overline{y}_1 = y_1/x_1 = -0.1008 \text{ (males)},$$

and

$$\overline{y}_2 = y_2/x_2 = 0.1347 \text{ (females)}.$$

The diagonal 2×2 matrices diag($\overline{\mathbf{y}}$), $\mathbf{Y}$ and $\mathbf{X}$ are constructed in the same way as for PHYS. The covariance-matrix estimate $\hat{\mathbf{V}}_{xx}$ is equal to that for PHYS, and the covariance-matrix estimates $\hat{\mathbf{V}}_{yy}$ and $\hat{\mathbf{V}}_{yx}$ are:

$$\hat{\mathbf{V}}_{yy} = \begin{bmatrix} 6765.34 & 1036.34 \\ 1036.34 & 6585.20 \end{bmatrix},$$

$$\hat{\mathbf{V}}_{yx} = \begin{bmatrix} -3139.98 & 2129.01 \\ -2051.46 & 2259.73 \end{bmatrix} = \hat{\mathbf{V}}'_{xy}.$$

By using these matrices we calculate for PSYCH means the covariance-matrix estimate $\hat{\mathbf{V}}_{des}$:

$$\hat{\mathbf{V}}_{des} = \begin{bmatrix} \hat{v}_{des}(\overline{y}_1) & \hat{v}_{des}(\overline{y}_1, \overline{y}_2) \\ \hat{v}_{des}(\overline{y}_2, \overline{y}_1) & \hat{v}_{des}(\overline{y}_2) \end{bmatrix} = 10^{-4} \begin{bmatrix} 3.223 & 0.427 \\ 0.427 & 5.856 \end{bmatrix}.$$

Results from the design-based covariance-matrix estimation for PHYS proportions and PSYCH means including the standard-error estimates $\text{s.e}_{des}(\hat{r}_j)$ are displayed below.

		PHYS		PSYCH		
j	Domain	$\hat{p}_j$	$\text{s.e}_{des}(\hat{p}_j)$	$\overline{y}_j$	$\text{s.e}_{des}(\overline{y}_j)$	$\hat{n}_j$
1	Males	0.460	0.0167	−0.1008	0.0180	4485
2	Females	0.194	0.0140	0.1347	0.0242	3356
Total sample		0.346	0.0144	0.0000	0.0158	7841

Variance and covariance estimates $\hat{\mathbf{V}}_{yy}$, $\hat{\mathbf{V}}_{xx}$ and $\hat{\mathbf{V}}_{yx}$ can be calculated using the cluster-level data set displayed in Table 5.9 by suitable software for correlation analysis. The matrix operations in the formula of $\hat{\mathbf{V}}_{des}$ can be executed by any suitable software for matrix algebra. In practice, however, it is convenient to estimate $\hat{\mathbf{V}}_{des}$ using an element-level data set using appropriate software for survey analysis. Generally, in the case of u domains formed by several categorical predictors, a linear ANOVA model can be used by fitting, with an appropriate sampling design option, for the response variable, a full-interaction model excluding the intercept. The model coefficients are then equal to the domain proportion or mean estimates, and the covariance-matrix estimate of the model coefficients provides the covariance-matrix estimate $\hat{\mathbf{V}}_{des}$ of the proportions or means.

We next calculate the design-effects matrix. For this, a binomial covariance-matrix estimate is needed.

For PHYS, by computing the elements of the binomial covariance-matrix estimate

$$\hat{\mathbf{V}}_{bin}(\hat{\mathbf{p}}) = \begin{bmatrix} \hat{v}_{bin}(\hat{p}_1) & 0 \\ 0 & \hat{v}_{bin}(\hat{p}_2) \end{bmatrix} = \begin{bmatrix} \hat{p}_1(1-\hat{p}_1)/\hat{n}_1 & 0 \\ 0 & \hat{p}_2(1-\hat{p}_2)/\hat{n}_2 \end{bmatrix}$$

of the proportion vector $\hat{\mathbf{p}}$ we obtain

$$\hat{p}_1(1-\hat{p}_1)/\hat{n}_1 = 0.4595(1-0.4595)/4485 = 0.0000554 \text{ (males)},$$

and

$$\hat{p}_2(1-\hat{p}_2)/\hat{n}_2 = 0.1937(1-0.1937)/3356 = 0.0000465 \text{ (females)}.$$

Inserting these variance estimates in $\hat{\mathbf{V}}_{bin}$ we have

$$\hat{\mathbf{V}}_{bin}(\hat{\mathbf{p}}) = 10^{-4}\begin{bmatrix} 0.554 & 0 \\ 0 & 0.465 \end{bmatrix}.$$

It is important to note that the covariance-matrix estimate $\hat{\mathbf{V}}_{bin}$ is diagonal because the proportion estimates $\hat{p}_1$ and $\hat{p}_2$ are assumed to be uncorrelated. The effect of clustering is not accounted for, even in the variance estimates, in the estimate $\hat{\mathbf{V}}_{bin}$. Therefore, with positive intra-cluster correlation, the binomial variance estimates $\hat{v}_{bin}(\hat{p}_j)$ tend to be underestimates of the corresponding variances. This appears when calculating the design-effects matrix estimate $\hat{\mathbf{D}} = \hat{\mathbf{V}}_{bin}^{-1}\hat{\mathbf{V}}_{des}$ of the estimate $\hat{\mathbf{p}}$:

$$\begin{aligned}\hat{\mathbf{D}}(\hat{\mathbf{p}}) &= \begin{bmatrix} 18\,058.295 & 0 \\ 0 & 21\,489.421 \end{bmatrix} \times 10^{-4}\begin{bmatrix} 2.775 & 0.576 \\ 0.576 & 1.951 \end{bmatrix} \\ &= \begin{bmatrix} 5.01 & 1.04 \\ 1.24 & 4.19 \end{bmatrix}.\end{aligned}$$

The design-effect estimates $\hat{d}_j$ on the diagonal of $\hat{\mathbf{D}}$ are thus

$$\hat{d}(\hat{p}_1) = \hat{v}_{des}(\hat{p}_1)/\hat{v}_{bin}(\hat{p}_1) = 0.0002775/0.0000554 = 5.01 \text{ (males)},$$

and

$$\hat{d}(\hat{p}_2) = \hat{v}_{des}(\hat{p}_2)/\hat{v}_{bin}(\hat{p}_2) = 0.0001951/0.0000465 = 4.19 \text{ (females)}.$$

These estimates are quite large, indicating a strong clustering effect for the response PHYS. This results in severe underestimation of standard errors of the estimates $\hat{p}_j$ when the binomial covariance-matrix estimate $\hat{\mathbf{V}}_{bin}$ is used. In addition to the design-effect estimates, the eigenvalues of the design-effect matrix, i.e. the generalized design effects, can be calculated. These are $\hat{\delta}_1 = 5.81$ and $\hat{\delta}_2 = 3.39$. It may be noted that the sum of the design-effect estimates is 9.20, which is equal to the sum of the eigenvalues. The mean of the design-effect estimates is 4.60, which indicates a strong average clustering effect over the sex groups. However, the mean is noticeably smaller than the overall design-effect estimate $\hat{d} = 7.2$ for the proportion estimate $\hat{p}$ calculated from the whole sample. This is due to

the property of design-effect estimates that, when compared against the overall design-effect estimate, they tend to get smaller in cross-class-type domains.

Estimation results for PHYS proportions are collected below.

j	Domain	$\hat{p}_j$	s.e_{des}	s.e_{bin}	$\hat{d}_j$	$\hat{n}_j$
1	Males	0.460	0.0167	0.0074	5.01	4485
2	Females	0.194	0.0140	0.0068	4.19	3356
Total sample		0.346	0.0144	0.0054	7.17	7841

5.8 CHAPTER SUMMARY AND FURTHER READING

Summary

Proper estimation of the variance of a ratio estimator is important in the analysis of complex surveys. First, variance estimates are needed to derive standard errors and confidence intervals for nonlinear estimators such as a ratio estimator. The estimation of the variance of ratio mean and ratio proportion estimators was carried out under an epsem two-stage stratified cluster-sampling design, where the sample data set was assumed self-weighting so that adjustment for nonresponse was not necessary. The demonstration data set from the modified sampling design of the Mini-Finland Health Survey (MFH Survey) fulfilled these conditions.

A ratio-type estimator $\hat{r} = y/x$ was examined for the estimation of the subpopulation mean and proportion in the important case of a subgroup of the sample whose size x was not fixed by the sampling design. Therefore, the denominator quantity x in $\hat{r}$ is a random variable, involving its own variance and covariance with the numerator quantity y. In addition to the variance of y, these variance and covariance terms contributed to the variance estimator of a ratio estimator calculated with the linearization method. This method was considered in depth because of its wide applicability in practice and popularity in software products for survey analysis.

We also introduced alternative methods for variance estimation of a ratio estimator based on sample reuse methods. The techniques of balanced half-samples (BRR) and jackknife (JRR) are traditional sample reuse methods, but the bootstrap (BOOT) has been applied for complex surveys only recently. Being computer-intensive, they differ from the linearization technique but are, as such, readily applicable for different kinds of nonlinear estimators. With-replacement sampling of clusters was assumed for all the approximation methods. With this assumption, the variability of a ratio estimate was evaluated using the between-cluster variation only, leading to relatively simple variance estimators. The design effect was used extensively as a measure of the contribution of the clustering on

a variance estimate, relative to the variance estimate based on simple random sampling with replacement.

The MFH Survey sampling design was selected for variance estimation because of its simplicity: there were exactly two sample clusters in each stratum in the modified sampling design. A subgroup of the MFH Survey data set covering 30–64-year-old males was used with all the variance approximation methods. This specific subgroup was chosen instead of the entire MFH Survey sample because the total sample size was fixed by the sampling design, but for the subpopulation considered the sample size was a random variate, thus providing a good target for demonstrating variance estimation with approximative methods. The selected subgroup constitutes a cross-classes-type domain mimicking properly all essential properties of the MFH Survey sampling design such as inclusion of elements from all of the 24 strata and 48 sample clusters. This would not be the case if, for example, a regional subgroup were chosen where only a part of the strata and sample clusters would be covered.

The variance approximation methods provided similar results in variance estimation of a proportion estimator of a binary response variable CHRON (chronic morbidity), which was a slightly intra-cluster correlated variable, and for a mean estimator of a continuous response variable SYSBP (systolic blood pressure) having stronger intra-cluster correlation. Because no theoretical arguments are available for choosing between the approximative variance estimators, technical factors such as software availability often guide the selection of an appropriate method in practice.

Several domain ratios, collected in a vector of ratios, were estimated using appropriate element weights in a combined ratio estimator derived for each domain. This produced consistent estimation of the ratios under a non-epsem complex sampling design. Use of the linearization method gave consistent estimation of the covariance matrix of the weighted domain ratio estimator vector. It was demonstrated that positive intra-cluster correlation of a response variable not only increases the variance estimates but can also introduce nonzero correlations between ratio estimates from separate domains; the asymptotically valid covariance-matrix estimator was derived to account for the extra variation and nonzero correlations. The estimator was essentially nondiagonal with nonzero off-diagonal covariance terms that occurred especially when working with cross-classes-type domains. This kind of a covariance-matrix estimator is needed for asymptotically valid modelling procedures with logit and linear models.

A covariance-matrix estimate calculated by the linearization method might be unstable in such small-sample situations where the number of sample clusters is small. Instability can cause problems in standard-error estimation and in testing and modelling procedures. Techniques are available for detecting instability, based, for example, on a statistic condition number and on graphical inspection of a covariance-matrix estimate. For a design-effects matrix estimator, a binomial covariance-matrix estimator of the consistent (weighted) domain proportion estimator vector was constructed. This kind of a design-effects matrix estimator is

primarily intended to account for intra-cluster correlation in testing and modelling procedures, and will be extensively used in Chapters 7 and 8. By using a binomial covariance-matrix estimator of an unweighted proportion estimator vector, a different design-effects matrix estimator would be obtained accounting for all the other contributions of complex sampling on covariance-matrix estimation such as weighting procedures. We demonstrate empirically both approaches in Sections 9.3 and 9.4. It should be noticed that different definitions of a design effect can be employed in software products for survey analysis, leading to different design-effect estimates from the same data set. Therefore, care should be taken to avoid misinterpretation.

Further Reading

In-depth consideration of the estimation of variance of a ratio, and other nonlinear estimators, can be found in Wolter (1985). Supplementary sources on the topic, in addition to those already mentioned, are Kalton (1983) and Verma *et al.* (1980). Thorough discussion on the concept of design effect is given in Kish (1995). Jackknife technique and the bootstrap are discussed in Shao and Tu (1995). Rao and Shao (1993) and Yung and Rao (2000) address the jackknife technique for variance estimation. Rao (1999) reviews many of the advances in variance estimation under complex sampling.

The estimation of the asymptotic covariance matrix of a domain ratio estimator vector is considered in Skinner *et al.* (1989). Smoothed estimates for unstable situations are derived in Singh (1985), Kumar and Singh (1987), Morel (1989) and Lehtonen (1990). The method of effective sample sizes is introduced in Scott (1986) and applied in Rao and Scott (1992). Brier (1980), Williams (1982) and Wilson (1989) consider accounting for extra-binomial variation using the beta-binomial sampling model. The role of weighting for unequal inclusion probabilities and for adjustment for nonresponse in the analysis of complex surveys has deserved its considerable attention in the literature. Important contributions are by Little (1991, 1993), Kish (1992), Pfeffermann (1993) and Pfeffermann *et al.* (1998).

6

Model-Assisted Estimation for Domains

In this chapter, we examine the estimation for population subgroups or domains. Regional areas constructed by administrative criteria, such as county or municipality, are typical domains or *domains of interest.* The population also can be grouped into domains by demographic criteria, such as sex and age group, as in a social survey. In a business survey, enterprises are often grouped into domains according to the type of industry. Further, elements can be assigned into domains by demographic criteria within regional areas. In all these instances, *estimation for domains*, or *domain estimation*, refers to the estimation of population quantities, such as totals, for the desired population subgroups. Estimation of domain totals will be discussed in the context of design-based estimation, which is the main approach of the book. In practice, design-based estimation is mainly used for domains whose sample size is reasonably large. For small domains (with a small sample size in a domain), methods falling under the headline of *small area estimation* are often used. In Section 6.1, we outline the framework and basic principles of domain estimation. We also summarize the operational steps of a domain estimation procedure. Section 6.2 introduces two important concepts, estimator type and model choice, in the context of domain estimation. Selected estimators and models are worked out and illustrated in Section 6.3. Section 6.4 includes an empirical examination of properties of some estimators of domain totals based on Monte Carlo experiments. Summary and further reading is in Section 6.5.

6.1 FRAMEWORK FOR DOMAIN ESTIMATION

We focus on the estimation of population totals for domains in a descriptive survey. The estimation of domain totals is discussed from a design-based perspective, with the use of auxiliary information. According to Särndal *et al.* (1992), the framework

Practical Methods for Design and Analysis of Complex Surveys Risto Lehtonen and Erkki Pahkinen
 ISBN: 0-470-84769-7

is called *model-assisted*. The reason for incorporating auxiliary data in a domain estimation procedure is obvious: with strong auxiliary data it is possible to obtain better accuracy for domain estimates, when compared to an estimation procedure not using auxiliary data. Thus, this chapter extends the treatment of model-assisted estimation introduced in Section 3.3.

Different types of auxiliary data can be used in model-assisted estimation. In Section 3.3, we used population-level aggregates of auxiliary variables. Here, we also employ unit-level auxiliary data for model-assisted estimation for domains. These data are incorporated in a domain estimation procedure by unit-level statistical models. This is possible if we make the following technical assumptions: (1) register data (such as population census register, business register, different administrative registers) are available as frame populations and sources of auxiliary data, (2) registers contain unique identification keys that can be used in merging at micro-level data from registers and sample surveys (see Figure 1.1 in Chapter 1). Obviously, access to micro-merged register and survey data involves much flexibility for a domain estimation procedure. This view has been adopted, for example, in Särndal (2001) and Lehtonen *et al.* (2003). Much of the material of this chapter are based on these sources.

The methods specific to small-area estimation include a variety of model-dependent techniques such as synthetic (SYN) estimators, composite estimators, EBLUP (empirical best linear predictor) estimators and various Bayesian techniques, and techniques developed in the context of demography and disease mapping. The monograph by J.N.K. Rao (2003) provides a comprehensive treatment of model-dependent small-area estimation and discusses design-based methodologies for the estimation for domains as well. Other materials include, for example, Schaible (1996), Lawson *et al.* (1999), and Ghosh (2001), who discusses especially empirical and hierarchical Bayes techniques.

Basic Principles

Let us introduce our basic notation for population quantities and sample-specific quantities in the context of domain estimation. The finite population is again denoted by $U = \{1, 2, \ldots, k, \ldots, N\}$ and, in domain estimation, we consider a set of mutually exhaustive subgroups of the population denoted $U_1, \ldots, U_d, \ldots, U_D$ (note that in this chapter we use exclusively a subscript d for domains of interest). We assume that the population U can be used as a sampling frame. This implies that U is available as a computerized data set, for example, a population register, or a register of business firms. We therefore also assume that the frame population U contains (in addition to the 'labels' k of the population elements) values for certain additional variables for all elements $k \in U$ (where the symbol '$\in$' refers to the inclusion of an element in a set of elements). These variables are unique element-identification (ID) keys, domain membership indicators, stratum membership indicators and the auxiliary z-variables.

Denote by y the variable of interest and by Y_k its unknown population value for unit k. The target parameters are the set of domain totals, $T_d = \sum_{k \in U_d} Y_k$, $d = 1, \ldots, D$, where summation is over all population elements k belonging to domain U_d (for simplicity, we use this notation throughout this chapter). Auxiliary information is essential for building accurate domain estimators, and increasingly so when the sample size of domains get smaller. Let $\mathbf{z}_k = (z_{1k}, \ldots, z_{jk}, \ldots, z_{Jk})'$ be the auxiliary variable vector of dimension $J \geq 1$. The value $\mathbf{z}_k$ is assumed to be known for every element $k \in U$. In a survey on individuals, $\mathbf{z}_k$ may specify known data about a person k, such as age, sex, taxable income and other continuous or qualitative variable values. In a business survey, $\mathbf{z}_k$ may indicate the turnover, or the total number of staff, for business firm k. It is important to emphasize that we assume the auxiliary z-data to be at the micro-level, that is, a value is assigned for each population element in the frame register. This is for flexibility, because the data can be then aggregated at higher levels of the population, such as at the domain or stratum level, if desired. Indeed, for some estimators, it suffices to know the population totals $T_{dz_1}, \ldots, T_{dz_j}$ of the auxiliary variables z_j for each domain of interest. In the model-fitting phase, we often assume that a constant value 1 is assigned as the first element in a vector $\mathbf{z}_k$.

For unique identification of domain membership for each population element, we define $\boldsymbol{\delta}_k = (\delta_{1k}, \ldots, \delta_{dk}, \ldots, \delta_{Dk})'$ to be the *domain indicator vector* for unit k, such that $\delta_{dk} = 1$ for all elements $k \in U_d$, and $\delta_{dk} = 0$ for all elements $k \notin U_d$, $d = 1, \ldots, D$. An indicator vector $\boldsymbol{\tau}_k$ for *stratum identification* for population element k is constructed in a similar manner: $\tau_{hk} = 1$ for all $k \in U_h$, $h = 1, \ldots, H$, and $\tau_{hk} = 0$ otherwise, where U_h refers to stratum h and H is the number of strata. Thus, a total of D domain indicator variables and H stratum indicator variables are assumed in the population frame.

A probability sample s of size n is drawn from U using a sampling design $p(s)$ such that an inclusion probability π_k is assigned to unit k. The corresponding sampling weights are $w_k = 1/\pi_k$. Measurements y_k of the response variable y are obtained for the sampled elements $k \in s$. We assume that a unique element ID key is included in sample s making it possible to micro-merge these data with the frame register U.

The domain samples are $s_d = U_d \cap s$, $d = 1, \ldots, D$. A domain is defined *unplanned*, if the domain sample size n_{s_d} is not fixed in the sampling design. This is the case in which the desired domain structure is not a part of the sampling design. Thus, the *domain sample sizes are random quantities* introducing an increase in the variance estimates of domain estimators. In addition, an extremely small number (even zero) of sample elements in a domain can be realized in this case, if the domain size in the population is small. For *planned* domains on the other hand, the *domain sample sizes are fixed in advance by stratification*. Stratified sampling in connection with a suitable allocation scheme is often used in practical applications.

A certain domain structure for a stratified sample of n elements can be illustrated, for example, as in Table 6.1. In the table setting, an unplanned domain structure

Table 6.1 Planned and unplanned domain structures in a stratified sample of n elements.

Unplanned domains	Strata (planned domains)						
	1	2	...	h	...	H	Sum
1	$n_{s_{11}}$	$n_{s_{12}}$	...	$n_{s_{1h}}$	...	$n_{s_{1H}}$	n_{s_1}
2	$n_{s_{21}}$	$n_{s_{22}}$	...	$n_{s_{2h}}$	...	$n_{s_{2H}}$	n_{s_2}
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
d	$n_{s_{d1}}$	$n_{s_{d2}}$	⋯	$n_{s_{dh}}$	⋯	$n_{s_{dH}}$	n_{s_d}
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
D	$n_{s_{D1}}$	$n_{s_{D2}}$	...	$n_{s_{Dh}}$	...	$n_{s_{DH}}$	n_{s_D}
Sum	n_1	n_2	...	n_h	...	n_H	n

Sample sizes n_{s_d}, $d = 1, \ldots, D$, for *unplanned* domains are not fixed in advance and thus are random variables.
Stratum sample sizes n_h, $h = 1, \ldots, H$ are fixed in the sampling design. Thus, the strata are defined as *planned* domains.
Cell sample sizes $n_{s_{dh}}$ are random variables in both cases.

cuts across the strata, a situation that is common in practice. In other types of structures, strata and domains can be nested such that a stratum contains several unplanned domains (for example, regional sub-areas within larger areas) or the strata themselves constitute the domains. The latter case represents a planned domain structure. Singh *et al.* (1994) illustrates the benefits of the planned domain approach for domain estimation. They presented compromise sample allocation schemes for the Canadian labour force survey to satisfy reliability requirements at the provincial level as well as at sub-provincial level. However, for practical reasons, it is usually not possible to define all desired domain structures as strata.

For the estimation for domains, it is advisable to apply the planned domains approach when possible, by defining the most important domains of interest as strata and to use a suitable allocation scheme in the sampling design, such as power or Bankier allocation (see the next example). It is also beneficial to use a large overall sample size to avoid small expected domain sample sizes if an unplanned domain approach is used. And in the estimation phase, it is often useful to incorporate strong auxiliary data into the estimation procedure by carefully chosen models and estimators of domain totals (see Example 6.2 and Section 6.4).

Example 6.1

Impact of sampling design in estimation for domains: the cases of unplanned and planned domain structures. Problems may be encountered when working with an unplanned domain structure, because small domain samples can be obtained

for domains with a small population size, if the overall sample size is not large, involving imprecise estimation. For example, if the sample has been drawn with simple random sampling without replacement, then the expected sample size in a domain would be $E(n_{s_d}) = n \times (N_d/N)$, thus corresponding to the proportional allocation in stratified sampling. An alternative is based on the *planned domain* structure, where the domains are defined as strata. Then, more appropriate allocation schemes can be used. In this example, the allocation scheme is based on *power allocation* (see Section 3.1). In power or Bankier allocation, the sample is allocated to the domains on the basis of information on the coefficient of variation of the response variable y in the domains and on the possibly known domain totals T_{dz} of an auxiliary variable z. We use a simplified version of power allocation in a hypothetical situation in which the coefficients of variation $\mathrm{C.V}_{dy} = S_{dy}/\overline{Y}_d$ of the response variable y are known in all domains, where S_{dy} and $\overline{Y}_d$ are the population standard deviation and the population mean of y in domain d, respectively.

In power allocation, the domain sample sizes are given by

$$n_{d,pow} = n \times \frac{T_{dz}^a \times \mathrm{C.V}_{dy}}{\sum_{d=1}^{D} T_{dz}^a \times \mathrm{C.V}_{dy}},$$

where the coefficient a refers to the desired power (typical choices are 0, 0.5 or 1). Here we have chosen $a = 0$ for simplicity. Thus, information on coefficients of variation is only used.

We illustrate the methodology by selecting an SRSWOR sample ($n = 392$ persons) from the Occupational Health Care Survey (OHC) data set ($N = 7841$ persons) and estimating the total number of chronically ill persons in the $D = 30$ domains constructed. In the population, the sizes of the domains vary with a minimum of 81 persons and a maximum of 517 persons. The results for the allocation of the sample by proportional allocation (corresponding to an unplanned domain structure) and by power allocation (corresponding to a planned domain structure) are shown in Table 6.2. The domain totals of the number of chronically ill persons are estimated by a Horvitz–Thompson (HT) estimator $\hat{t}_{dHT} = \sum_{k \in s_d} w_k y_k$. The stability of the estimators is measured by the population coefficient of variation of an estimator of a domain total, given by $\mathrm{C.V}(\hat{t}_{dHT}) = \mathrm{S.E}\,(\hat{t}_{dHT})/T_d$.

The results show that SRSWOR sampling produces a large variation in the expected domain sample size: the average domain sample size is 13, the minimum sample size is 4 and the maximum is 26. On the other hand, power allocation smoothes considerably the variation in domain sample size: the minimum domain sample size is now 10 and the maximum is 17. The percentage coefficient of variation varies much in the case of SRSWOR. For example, the difference between the smallest and largest coefficient of variation is over 60% points. In power

Table 6.2 Allocation schemes for a sample of $n = 392$ elements for $D = 30$ domains of the OHC Survey data set. Calculation of the expected domain sample size $E(n_{s_d})$ under an SRSWOR design and realized domain sample size n_d under a stratified SRSWOR design with power allocation ($a = 0$), and the corresponding coefficients of variation (%) of a Horvitz–Thompson estimator $\hat{t}_{dHT}$.

		Domain sample size		Coefficient of variation C.V (%) of HT estimators of domain totals	
Domain		**Unplanned domain structure** Expected under SRSWOR	**Planned domain structure** Realized under stratified SRSWOR (power allocation)	**Unplanned domain structure** SRSWOR	**Planned domain structure** Stratified SRSWOR (power allocation)
d	N_d	$E(n_{s_d})$	n_d	C.V($\hat{t}_{dHT}$)	C.V($\hat{t}_{dHT}$)
10	81	4	11	84.10	38.88
20	101	5	12	78.41	40.54
18	129	6	13	72.69	42.38
3	133	7	15	81.04	45.63
8	141	7	16	81.03	46.54
30	146	7	15	74.80	45.03
21	153	8	12	62.87	41.15
23	156	8	11	57.65	39.05
16	165	8	13	64.94	43.19
1	181	9	17	75.90	48.78
11	187	9	14	63.52	44.52
6	188	9	13	60.37	43.22
28	194	10	10	50.52	38.69
24	200	10	13	58.68	43.39
22	242	12	10	44.27	38.30
15	252	13	14	55.68	45.50
7	292	15	17	60.34	50.06
4	295	15	15	53.92	47.04
13	305	15	13	46.00	43.04
12	311	16	12	44.50	42.38
5	323	16	16	53.50	48.23
25	339	17	11	40.57	41.03
2	352	18	14	46.80	45.74
26	364	18	11	38.87	40.88
29	365	18	11	38.25	40.45
9	366	18	14	45.99	45.85
17	426	21	12	36.67	41.62
14	447	22	13	37.95	43.37
19	490	24	11	33.60	41.22
27	517	26	10	30.68	39.34
Sum	7841	392	392		

allocation, the difference is reduced to 12% points. Thus, power allocation tends to smooth the variation in the coefficient of variation such that large coefficients are considerably decreased. However, the coefficients of variation of estimated domain totals tend to be quite large; this is mainly due to the small overall sample size.

The progression in coefficients of variation can be illustrated graphically. In Figure 6.1, the coefficients of variation have been plotted against domain size in population. The curve for the HT estimator obtained for coefficients of variation under SRSWOR shows clear decrease with increasing domain size. For power allocation, the curve is clearly stabilized.

To continue the specification of the setting for domain estimation, our further technical assumption is as follows. We assume that after data collection from the selected sample and preparation of the final sample data set, denoted by $s(y)$, the population frame U and the sample measurements $s(y)$ can be micro-merged using the unique element ID keys that are available in both data sources. Completing this procedure we have obtained an enhanced frame register data set that includes the auxiliary z-data and stratum and domain indicator variables for all population elements, amended with y-measurements for the elements belonging to the sample.

We have now completed the technical preparations for conducting an estimation for the domains. The operational steps in a domain estimation procedure, given in general terms, are summarized in Box 6.1.

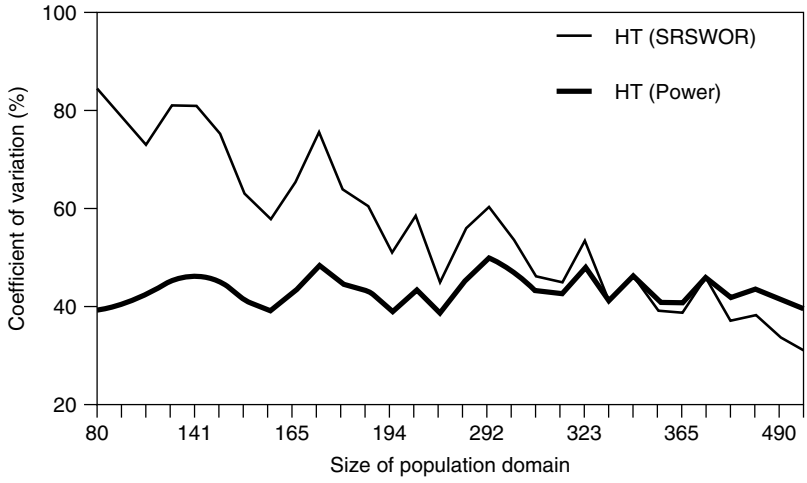


Figure 6.1 Coefficient of variation (%) of Horvitz–Thompson estimator of domain total under SRSWOR sampling (corresponding to the unplanned domain structure) and stratified SRSWOR sampling with power allocation ($a = 0$) (corresponding to the planned domain structure).

BOX 6.1 ***Operational steps in a domain estimation procedure***

Step 1: Construction of frame population Construction of the frame population $U = \{1, 2, \ldots, k, \ldots, N\}$ of N elements containing unique element ID keys, domain indicator vectors $\boldsymbol{\delta}_k$, stratum indicator vectors $\boldsymbol{\tau}_k$, inclusion probabilities π_k for drawing of an n element sample with sampling design $p(s)$, and the vectors $\mathbf{z}_k$ of auxiliary z-data, for all elements k in U.

Step 2: Sampling and measurement Sample selection by using the design $p(s)$ and measurement of the values of the response variable y, and the construction of the sample data set $s(y)$, including the element ID keys, observed values y_k and sampling weights $w_k = 1/\pi_k$, for all elements $k \in s$.

Step 3: Frame population revisited Construction of a combined data set by micro-merging the frame population U and the sample data set $s(y)$ by using the element ID keys.

Step 4: Model choice and model fitting The choice of the model, specification of model parameters and effects, model fitting using the sample data set and model validation and diagnostics. On the basis of the fitted model, calculation of fitted values $\hat{y}_k$ for all population elements $k \in U$ and residuals $\hat{e}_k = y_k - \hat{y}_k$ for all elements $k \in s\,(y)$, the sample data set.

Step 5: Choice of estimator of domain totals and estimation for domains Supply of fitted values, residuals and weights in the chosen estimator for domain totals. Basically, estimators of domain totals labeled 'model-dependent' use the fitted values $\hat{y}_k$, $k \in U$, and the estimators of domain totals labeled 'model-assisted' use the fitted values $\hat{y}_k$, $k \in U$, and in addition, the residuals $\hat{e}_k$ and the weights w_k, $k \in s$.

Step 6: Variance estimation and diagnostics Choice of an appropriate variance estimator. Calculation of standard error estimates and coefficients of variation.

In Table 6.3, we summarize in a hypothetical situation, the progression in the population frame data set that occurs when the operations in Steps 1 to 4 of Box 6.1 are implemented for a domain estimation procedure. Because the vectors $\mathbf{z}_k = (z_{1k}, \ldots, z_{Jk})'$ of auxiliary z-variables are assumed to be known for every population element, including sampled and nonsampled elements, the vector $\mathbf{T}_z = (T_{z_1}, \ldots, T_{z_J})'$ with $T_{z_j} = \sum_{k \in U} z_{jk}$, $j = 1, \ldots, J$, of population totals of auxiliary z-variables is known. Also, domain totals $T_{dz_j} = \sum_{k \in U_d} z_{jk}$, $d = 1, \ldots, D$ and $j = 1, \ldots, J$, can be calculated for each z-variable, because the domain indicators are assumed to be known for all $k \in U$. The sample membership

Table 6.3 Execution of Steps 1, 3 and 4 of Box 6.1 in a domain estimation procedure (hypothetical situation).

	Step 1: Construction of the frame population U				**Step 3**: Merging of the frame population U and the sample data set $s(y)$			**Step 4**: Calculation of fitted y-values and residuals	
Element ID	Domain ID vectors $\boldsymbol{\delta}'_k$	Stratum ID vectors $\boldsymbol{\tau}'_k$	Inclusion probability π_k	Auxiliary z-vectors $\mathbf{z}'_k$	Sampling weight w_k	Sample membership indicator I_k	Study variable y_k	Fitted values $\hat{y}_k$	Residuals $\hat{e}_k$
1	$\boldsymbol{\delta}'_1$	$\boldsymbol{\tau}'_1$	π_1	$\mathbf{z}'_1$	0	0	...	$\hat{y}_1$	...
2	$\boldsymbol{\delta}'_2$	$\boldsymbol{\tau}'_2$	π_2	$\mathbf{z}'_2$	0	0	...	$\hat{y}_2$	...
3	$\boldsymbol{\delta}'_3$	$\boldsymbol{\tau}'_3$	π_3	$\mathbf{z}'_3$	w_3	1	y_3	$\hat{y}_3$	$\hat{e}_3$
4	$\boldsymbol{\delta}'_4$	$\boldsymbol{\tau}'_4$	π_4	$\mathbf{z}'_4$	0	0	...	$\hat{y}_4$	...
5	$\boldsymbol{\delta}'_5$	$\boldsymbol{\tau}'_5$	π_5	$\mathbf{z}'_5$	w_5	1	y_5	$\hat{y}_5$	$\hat{e}_5$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
k	$\boldsymbol{\delta}'_k$	$\boldsymbol{\tau}'_k$	π_k	$\mathbf{z}'_k$	w_k	1	y_k	$\hat{y}_k$	$\hat{e}_k$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
N	$\boldsymbol{\delta}'_N$	$\boldsymbol{\tau}'_N$	π_N	$\mathbf{z}'_N$	0	0	...	$\hat{y}_N$	...

... Nonsampled element.

indicator variable I is created for the whole population data set such that $I_k = 1$ if $k \in s$, zero otherwise. Obviously, the sum of the indicator variable over the population is n, the sample size. In the model-fitting phase, the fitted values $\hat{y}_k$ are calculated for all N elements $k \in U$. On the other hand, the residuals $\hat{e}_k = y_k - \hat{y}_k$ can be calculated for the sampled elements $k \in s$ only. It is also important to emphasize that the fitted values $\{\hat{y}_k; k \in U\}$ calculated by a given model differ from one model specification to another. This will be apparent in the next section in which models and estimators of domain totals are treated in more detail.

6.2 ESTIMATOR TYPE AND MODEL CHOICE

Important phases in a model-assisted domain estimation procedure are the selection of the type of the estimator of a total, the choice of the auxiliary variables to be used, the formulation of the model for the incorporation of the auxiliary data into the estimation procedure, the model-fitting phase and the derivation of variance estimators for the selected domain total estimators (see Box 6.1). In this section, we consider these phases in a more technical manner.

Estimator Type

We first discuss two concepts, estimator type and model choice, making the basis for the construction of an estimator of the population totals for domains of interest.

The concept *estimator type* refers to the explicit structure of the selected estimator of the domain totals. There are two main types of estimators discussed in this chapter. These are the *generalized regression (GREG) estimator* and the *synthetic (SYN) estimator*. The main conceptual difference in these estimators is that GREG estimators use models as assisting tools, whereas SYN estimators rely exclusively on the model used. Thus, GREG estimators are *model-assisted* and SYN estimators are *model-dependent*. The main consequence of this differing role of a model is that a GREG estimator of a domain total is constructed to be design unbiased (or approximately so) irrespective of the 'truth' of the model. This is a benefit of GREG estimators. However, a GREG estimator can be very unstable if the sample size in a domain becomes small. On the other hand, the bias of a SYN estimator depends heavily on a correct model specification. If the model is severely misspecified, a SYN estimator can involve substantial design bias. If, on the other hand, the model is correctly specified or nearly so, then the bias of a SYN estimator can be small.

In a typical large-scale survey conducted, for example, by a national statistical agency, some domains of interest are large enough, and the auxiliary information strong enough, so that the GREG-type estimators will be sufficiently precise. But for a small domain the variance of a GREG estimator can become unacceptably large, and in this case, the variance of a SYN estimator can be much smaller. Better precision of SYN estimators for small domains favours their use, in particular, for small-area estimation (recall that 'small area' refers to the situation in which the attained sample size in a given domain, or 'area', is small, or very small, even zero).

To summarize the main theoretical properties of the estimator types, GREG estimators are constructed to be design unbiased; the SYN estimators usually are not. Variance of the GREG estimator can be large for a small domain, that is, if the domain sample size is small, causing poor precision. The SYN estimator is usually design biased; its bias does not approach zero with increasing sample size; its variance is usually smaller than that of GREG; this holds especially for small domains. The accuracy, measured by the mean squared error MSE, of a SYN estimator can be poor even in the case of a small variance, if the bias is substantial.

Model Choice

The concept *model choice* refers to the specification of the relationship of the study variable y with the auxiliary predictor variables $z_1, \ldots, z_J$, as reflected by the structure of the constructed model. Model choice has two aspects, the *mathematical form* of the model and the *specification of the parameters and effects* in the model. For example, when working with a continuous study variable, a *linear model formulation* is usually appropriate. For binary or polytomous study variables, one might make a choice for a nonlinear model, such as a binomial or multinomial logistic model. For example, for a binary study variable, a *logistic model formulation* is arguably an improvement on a linear model type, because the fitted y-values

under the former will necessarily fall in to the unit interval, which is not always true for a linear model.

The second aspect of model choice is the specification of the parameters and effects in the model. Some of these may be defined at the fully aggregated population level, others at the level of the domain (domain-specific parameters), yet others at some intermediate level. We will separate a *fixed-effects model formulation* and a *mixed model formulation*. A fixed-effects model can involve population-level or domain-specific fixed effects, or effects specified on an intermediate level. In a mixed model, there are domain-specific *random effects* in addition to the fixed effects. Using a mixed model type, we can introduce stochastic effects that recognize domain differences.

To summarize, the chosen model specifies a hypothetical relationship between the variable of interest, y, and the predictor variables, $z_1, \ldots, z_J$, and makes assumptions about its perhaps complex error structure. Fixed-effects models can often be satisfactory, but mixed models offer additional possibilities for flexible modelling. For every specified model, we can derive one GREG estimator and one SYN estimator, by observing the respective construction principles. However, fixed-effects models have been more common in model-assisted estimators, whereas mixed models have most often been used in model-dependent estimators.

By combining these two aspects of an estimator for domain totals, estimator type and model choice, we get a two-dimensional arrangement of estimators. To illustrate this, we have included in Table 6.4 a number of selected estimators. There are six model-dependent SYN-type estimators and six design-based GREG-type estimators in the table. Each of the six rows corresponds to a different model choice. A *population model* (P-model; rows 1 and 2) is one whose only parameters are fixed effects defined at the population level; it contains no domain-specific parameters. A *domain model* (D-model) is one having at least some of its parameters or effects defined at the domain level. These are fixed effects for rows 3 and 4 and

Table 6.4 Classification of estimators for domain totals by model choice and estimator type.

Model choice			Estimator type	
Specification of model effects	Level of aggregation	Functional form	Model-dependent	Design-based model-assisted
Fixed-effects models	Population models	Linear	SYN-P	GREG-P
		Logistic	LSYN-P	LGREG-P
	Domain models	Linear	SYN-D	GREG-D
		Logistic	LSYN-D	LGREG-D
Mixed models including fixed and random effects	Domain models	Linear	MSYN-D	MGREG-D
		Logistic	MLSYN-D	MLGREG-D

random effects for rows 5 and 6. 'Linear' and 'logistic' refer to the mathematical forms. In Example 6.2 and Section 6.4, we will consider in more detail a number of these estimators.

6.3 CONSTRUCTION OF ESTIMATORS AND MODEL SPECIFICATION

Construction of Estimators of Domain Totals

The estimators of domain totals are constructed in the following three phases (according to Steps 4 and 5 in Box 6.1):

1. The parameters of the designated model are estimated using the sample data set $s(y) = \{(y_k, \mathbf{z}_k); k \in s\}$.
2. Using the estimates of the model parameters and the population vectors $\mathbf{z}_k$, the fitted value $\hat{y}_k$ is computed for every population element k, including elements belonging to the sample and also elements that are not sampled.
3. For obtaining an estimate $\hat{t}_d$ of the total T_d in domain d, the fitted values, $\{\hat{y}_k; k \in U\}$, and the sample observations, $\{y_k; k \in s\}$, are incorporated in the respective formulas for the GREG and SYN estimators.

We will illustrate the domain estimation procedure in the context of linear models. Consider a *fixed-effects linear model* specification such that $y_k = \mathbf{z}'_k\boldsymbol{\beta} + \varepsilon_k$, where $\boldsymbol{\beta}$ is an unknown parameter vector requiring estimation, and ε_k are the residual terms. The model fit yields the estimate $\hat{\boldsymbol{\beta}}$. The supply of fitted values given by $\hat{y}_k = \mathbf{z}'_k\hat{\boldsymbol{\beta}}$ is computed for all elements $k \in U$. Similarly, for a *linear mixed model* involving domain-specific random effects in addition to the fixed effects, the model specification is $y_k = \mathbf{z}'_k(\boldsymbol{\beta} + \mathbf{u}_d) + \varepsilon_k$, where $\mathbf{u}_d$ is a vector of *random effects* defined at the domain level. Using the estimated parameters, fitted values given by $\hat{y}_k = \mathbf{z}'_k(\hat{\boldsymbol{\beta}} + \hat{\mathbf{u}}_d)$ are computed for all $k \in U$. In more general terms, models used in the construction of GREG- and SYN-type estimators of domain totals are special cases of *generalized linear mixed models*, such as a mixed linear model and a logistic model (see e.g. McCulloch and Searle 2001; Dempster *et al.* 1981).

The fitted values $\{\hat{y}_k; k \in U\}$ differ from one model specification to another. For a given model specification, an estimator of domain total $T_d = \sum_{k \in U_d} y_k$ has the following structure for the two basic estimator types:

$$\text{Synthetic estimator:} \qquad \hat{t}_{dSYN} = \sum_{k \in U_d} \hat{y}_k \tag{6.1}$$

$$\text{Generalized regression estimator:} \qquad \hat{t}_{dGREG} = \sum_{k \in U_d} \hat{y}_k + \sum_{k \in s_d} w_k(y_k - \hat{y}_k) \tag{6.2}$$

where $w_k = 1/\pi_k$, $s_d = s \cap U_d$ is the part of the full sample s that falls in to domain U_d, and $d = 1, \ldots, D$.

Note that $\hat{t}_{dSYN}$ uses the fitted values given by the estimated model, and thus relies on the 'truth' of the model and, therefore, can be biased. On the other hand, $\hat{t}_{dGREG}$ has a second term that aims at protecting against possible model misspecification. Note also that in the case in which there are no sample elements in a domain, $\hat{t}_{dGREG}$ reduces to $\hat{t}_{dSYN}$ for that domain. A Horvitz–Thompson estimator $\hat{t}_{dHT} = \sum_{k \in s_d} w_k y_k$ is often calculated as a reference when assessing the benefits from the more complex estimators.

Model Specification

Let us first discuss *fixed-effects linear models*. Let $\mathbf{z}_k = (1, z_{1k}, \ldots, z_{jk}, \ldots, z_{Jk})'$ be a $(J+1)$-dimensional vector containing the values of $J \geq 1$ predictor variables z_j, $j = 1, \ldots, J$. This vector is used to create the predicted values $\hat{y}_k$, $k \in U$, in the estimators (6.1) and (6.2).

1. *Fixed-effects P-models.* The estimators SYN-P and GREG-P build on the model specification

$$y_k = \beta_0 + \beta_1 z_{1k} + \cdots + \beta_J z_{Jk} + \varepsilon_k = \mathbf{z}'_k \boldsymbol{\beta} + \varepsilon_k \tag{6.3}$$

for $k \in U$, where $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_J)'$ is a vector of fixed effects defined for the whole population. Owing to this property, we call (6.3) the *fixed-effects P-model*. If y-data were observed for the whole population, we could compute the generalized least-squares (GLS) estimator of $\boldsymbol{\beta}$ given by

$$\mathbf{B} = \left(\sum_{k \in U} \mathbf{z}_k \mathbf{z}'_k / c_k \right)^{-1} \sum_{k \in U} \mathbf{z}_k y_k / c_k, \tag{6.4}$$

where the c_k are specified positive weights. With no significant loss of generality, we specify these to be of the form $c_k = \boldsymbol{\lambda}' \mathbf{z}_k$ for $k \in U$, where the $(J+1)$-vector $\boldsymbol{\lambda}$ does not depend on k. As a further simple specification, we can set $c_k = 1$ for all k, and (6.4) reduces to an ordinary least-squares (OLS) estimator. In practice, a weighted least-squares (WLS) estimate for (6.4) is calculated on the observed sample data, yielding

$$\hat{\mathbf{b}} = \left(\sum_{k \in s} w_k \mathbf{z}_k \mathbf{z}'_k \right)^{-1} \sum_{k \in s} w_k \mathbf{z}_k y_k, \tag{6.5}$$

where $w_k = 1/\pi_k$ is the sampling weight of unit k. The resulting predicted values are given by

$$\hat{y}_k = \mathbf{z}'_k \hat{\mathbf{b}}, \quad k \in U. \tag{6.6}$$

By incorporating predicted values $\hat{y}_k$ into (6.1) and (6.2), we obtain the corresponding SYN-P and GREG-P estimators. Note that using a P-model for a given domain d, y-values from other domains also contribute to the predicted values incorporated in an estimator SYN-P and GREG-P for that domain. For this reason, the estimators $\hat{t}_{dSYN-P}$ and $\hat{t}_{dGREG-P}$, using a fixed-effects P-model type, are called *indirect* estimators.

2. *Fixed-effects D-models.* The estimators SYN-D and GREG-D are built with the same predictor vector $\mathbf{z}_k$, but with a different model specification allowing a fixed-effects vector $\boldsymbol{\beta}_d$ separately for every domain, so that

$$y_k = \mathbf{z}'_k\boldsymbol{\beta}_d + \varepsilon_k \tag{6.7}$$

for $k \in U_d$, $d = 1, \ldots, D$, or equivalently,

$$y_k = \sum_{d=1}^{D} \delta_{dk}\mathbf{z}'_k\boldsymbol{\beta}_d + \varepsilon_k \tag{6.8}$$

for $k \in U$, where δ_{dk} is the domain indicator of unit k, defined by $\delta_{dk} = 1$ for all $k \in U_d$, and $\delta_{dk} = 0$ for all $k \notin U_d$, $d = 1, \ldots, D$. Model (6.7) is called the *fixed-effects D-model*. Again, if the model (6.7) could be fitted to the data for the whole subpopulation U_d, the GLS estimator of β_d would be

$$\mathbf{B}_d = \left(\sum_{k\in U_d} \mathbf{z}_k\mathbf{z}'_k/c_k\right)^{-1} \sum_{k\in U_d} \mathbf{z}_k y_k/c_k, \quad d = 1, \ldots, D. \tag{6.9}$$

In practice, the fit must be based on the observed sample data in domain d. Setting again $c_k = 1$ for all k, the following WLS estimator can be used:

$$\hat{\mathbf{b}}_d = \left(\sum_{k\in s_d} w_k\mathbf{z}_k\mathbf{z}'_k\right)^{-1} \sum_{k\in s_d} w_k\mathbf{z}_k y_k, \quad d = 1, \ldots, D. \tag{6.10}$$

The resulting predicted values are given by

$$\hat{y}_k = \mathbf{z}'_k\hat{\mathbf{b}}_d \tag{6.11}$$

for $k \in U_d$; $d = 1, \ldots, D$. By incorporating predicted values $\hat{y}_k$ from (6.11) into (6.1) and (6.2), we obtain the corresponding SYN-D and GREG-D estimators. For a given domain d, y-values are used from that domain only in the model fitting and in the calculation of the predicted values incorporated in an estimator SYN-D and GREG-D in that domain. Thus, the estimators $\hat{t}_{dSYN-D}$ and $\hat{t}_{dGREG-D}$, using a fixed-effects D-model type, are called *direct* estimators. Note that because of the

specification $c_k = \boldsymbol{\lambda}'\mathbf{z}_k = 1$, we have $\sum_{k\in s_d} w_k(y_k - \hat{y}_k) = 0$. Consequently, SYN-D and GREG-D are identical, that is, $\hat{t}_{dSYN-P} = \hat{t}_{dGREG-P}$ for every sample s, when using the fixed-effects D-model specification.

3. *Mixed D-models.* The estimators MSYN-D and MGREG-D build on a two-level linear model, called the *mixed linear D-model*, involving fixed as well as random effects recognizing domain differences,

$$y_k = \beta_0 + u_{0d} + (\beta_1 + u_{1d})z_{1k} + \cdots + (\beta_J + u_{Jd})z_{Jk} + \varepsilon_k = \mathbf{z}'_k(\boldsymbol{\beta} + \mathbf{u}_d) + \varepsilon_k \tag{6.12}$$

for $k \in U_d$, $d = 1, \ldots, D$. Each coefficient is the sum of a fixed component and a domain-specific random component: $\beta_0 + u_{0d}$ for the intercept and $\beta_j + u_{jd}$, $j = 1, \ldots, J$ for the slopes. The components of $\mathbf{u}_d = (u_{0d}, u_{1d}, \ldots, u_{Jd})'$ represent deviations from the coefficients of the fixed-effects part of the model,

$$y_k = \beta_0 + \beta_1 z_{1k} + \cdots . + \beta_J z_{Jk} + \varepsilon_k = \mathbf{z}'_k\boldsymbol{\beta} + \varepsilon_k, \tag{6.13}$$

which agrees with (6.3). More generally, we can have that only some of the coefficients in (6.12) are treated as random, so that, for some j, $u_{jd} = 0$ for every domain d. A simple special case of (6.12), commonly used in practice, is the one that includes domain-specific random intercepts u_{0d} as the only random terms, given by $y_k = \beta_0 + u_{0d} + \beta_1 z_{1k} + \cdots + \beta_J z_{Jk} + \varepsilon_k$. We insert the resulting fitted y-values

$$\hat{y}_k = \mathbf{z}'_k(\hat{\boldsymbol{\beta}} + \hat{\mathbf{u}}_d) \tag{6.14}$$

into (6.1) to obtain the two-level MSYN-D estimator. Inserting the fitted values (6.14) into (6.2), we obtain the two-level MGREG-D estimator, introduced by Lehtonen and Veijanen (1999). A two-level D-model (6.12) can be fitted, for example, by estimating the variance components by maximum likelihood (ML) or restricted maximum likelihood (REML) and the fixed effects by GLS given these variance estimates; for details see, for example, Goldstein (2002) and McCulloch and Searle (2001). In estimating a mixed D-model, an assumption is usually made that the random effects follow a joint normal distribution. Note, however, that the assumption of normality is not necessary to obtain approximate unbiasedness for the resulting MGREG-D estimator.

Alternative options are available for the estimation of the design variance for estimators (6.1) and (6.2) of domain totals. When working with *planned domains*, where the domain sample sizes n_d are fixed in the stratified sampling design and, for example, the samples are drawn with SRSWOR in each stratum, approximate variance estimators presented in Section 3.3 for regression estimation can be used separately for each domain. In this setting, a sample of n_d elements is drawn from the population of N_d elements in domain d, and the weights are $w_k = N_d/n_d$ for

all $k \in U_d$. For example, for the GREG estimator (6.2), an approximate variance estimator is given by

$$\hat{v}_{srs}(\hat{t}_{dGREG}) = N_d^2 \left(1 - \frac{n_d}{N_d}\right) \left(\frac{1}{n_d}\right) \sum_{k \in s_d} \frac{(\hat{e}_k - \bar{\hat{e}}_d)^2}{n_d - 1}, \tag{6.15}$$

where the residuals are $\hat{e}_k = y_k - \hat{y}_k$, $k \in s_d$, and $\bar{\hat{e}}_d = \sum_{k \in s_d} \hat{e}_k / n_d$ is the mean of the residuals in domain d, $d = 1, \ldots, D$. It is obvious that in the SRSWOR case in which the weights are constants, for a direct estimator the sum of residuals in each domain is zero. But for other designs, and for an indirect estimator, the sum can differ from zero.

In an *unplanned domain* case, the extra variation due to a random domain sample size n_{s_d} should be accounted for. Let us consider the case of SRSWOR with n elements drawn from the population of N elements. The sampling fraction is n/N and the weights are $w_k = N/n$ for all k. By denoting $y_{dk} = \delta_{dk} y_k$ and $\hat{e}_{dk} = y_{dk} - \hat{y}_k$, $d = 1, \ldots, D$, where the domain membership indicator was given by $\delta_{dk} = 1$ if $k \in U_d$, zero otherwise, we obtain an approximate variance estimator given by

$$\hat{v}_{srs}(\hat{t}_{dGREG}) = N^2 \left(1 - \frac{n}{N}\right) \left(\frac{1}{n}\right) \sum_{k \in s} \frac{(\hat{e}_{dk} - \bar{\hat{e}}_d)^2}{n - 1}. \tag{6.16}$$

Note that also elements outside the domain d contribute to the variance estimate, because $\hat{e}_{dk} = -\hat{y}_k$ for elements $k \notin U_d$ and $k \in s$. An alternative approximate variance estimator is given by

$$\hat{v}_{srs}(\hat{t}_{dGREG}) = N^2 \left(1 - \frac{n}{N}\right) \left(\frac{1}{n}\right) p_d \sum_{k \in s_d} \frac{(\hat{e}_k - \bar{\hat{e}}_d)^2}{n_d - 1} \left(1 + \frac{q_d}{\text{c.v}_{d\hat{e}}^2}\right), \tag{6.17}$$

$d = 1, \ldots, D$, where $p_d = n_d/n$ and $q_d = 1 - p_d$, and $\text{c.v}_{d\hat{e}} = \hat{s}_{d\hat{e}} / \bar{\hat{e}}_d$ is the sample coefficient of variation of residuals in domain d with $\hat{s}_{d\hat{e}}$ as the sample standard deviation of residuals in domain d. The estimator (6.17) corresponds to the variance estimator commonly used under Bernoulli sampling (see Example 2.2).

Let us consider in more detail the choice of a model and the construction of an estimator of the total in the context of ratio estimation and regression estimation for domains. In Section 3.3 the estimation of the total T for the whole population was discussed. There, the auxiliary information assumed to be known at the whole-population level was the total T_z of the auxiliary variable z, and the assisting fixed-effects linear regression model was of the form $y_k = \beta_0 + \beta_1 z_k + \varepsilon_k$, $k \in U$, given by (6.3). The ratio estimator of the population total was given in Section 3.3 by $\hat{t}_{rat} = T_z \times \hat{t}/\hat{t}_z$, and the regression estimator by $\hat{t}_{reg} = \hat{t} + \hat{b}_1(T_z - \hat{t}_z)$, where $\hat{t}$ and $\hat{t}_z$ are SRSWOR estimators of totals T and T_z, respectively and the estimate $\hat{b}_1$ is a sample-based OLS estimate of the finite-population regression coefficient B_1.

For the estimation of domain totals T_d these ratio and regression estimators can be used, but more complex model types can also be introduced, including model types (6.3), (6.7) and (6.12) described above.

Consider a continuous response variable y, whose total T_d is to be estimated for a number of domains of interest U_d, $d = 1, \ldots, D$. Assuming one auxiliary variable z, for example, the following assisting models can be postulated.

1. Fixed-effects P-model for y_k, $k \in U$:
   (1a) $y_k = \beta_0 + \varepsilon_k$ Common intercept model
   (1b) $y_k = \beta_1 z_k + \varepsilon_k$ Common slope model
   (1c) $y_k = \beta_0 + \beta_1 z_k + \varepsilon_k$ Common intercept and slope model.
2. Fixed-effects D-model for y_k, $k \in U_d$, $d = 1, \ldots, D$:
   (2a) $y_k = \beta_{0d} + \varepsilon_k$ Domain-specific intercepts model
   (2b) $y_k = \beta_{1d} z_k + \varepsilon_k$ Domain-specific slopes model
   (2c) $y_k = \beta_{0d} + \beta_{1d} z_k + \varepsilon_k$ Domain-specific intercepts and slopes model.
3. Mixed D-model for y_k, $k \in U_d$, $d = 1, \ldots, D$:
   (3a) $y_k = \beta_{0d} + \varepsilon_k = \beta_0 + u_{0d} + \varepsilon_k$ Domain-specific random intercepts model
   (3b) $y_k = \beta_{0d} + \beta_1 z_k + \varepsilon_k = \beta_0 + u_{0d} + \beta_1 z_k + \varepsilon_k$ Domain-specific random intercepts and common slope model.

Models (1b) and (2b) can be used in ratio estimation for domains and models (1c) and (2c) in regression estimation. It is obvious that indirect SYN and GREG estimators are obtained with model specification (1) and (3), and model type (2) gives direct SYN and GREG estimators.

For example, using the P-model (1b), a SYN estimator (6.1) for domain totals T_d is given by

$$\hat{t}_{dSYN-P} = \sum_{k \in U_d} \hat{y}_k = \sum_{k \in U_d} \hat{b}_1 z_k = T_{dz} \hat{b}_1 = T_{dz} \times \hat{t}_{HT} / \hat{t}_{zHT}, \quad d = 1, \ldots, D, \tag{6.18}$$

resembling the ratio estimator $\hat{t}_{rat}$ for the whole population, but in $\hat{t}_{dSYN-P}$, domain totals T_{dz} are used instead of the overall total T_z. The estimator for the population slope B_1 is

$$\hat{b}_1 = \frac{\sum_{k \in s} w_k y_k}{\sum_{k \in s} w_k z_k} = \frac{\hat{t}_{HT}}{\hat{t}_{zHT}},$$

which is the ratio of two HT estimators, $\hat{t}_{HT}$ and $\hat{t}_{zHT}$, of totals of the study variable y and auxiliary variable z respectively. These total estimates are calculated at the whole-population level and, thus, the estimator of domain totals is *indirect*. While using y-values from the whole sample, the estimator $\hat{t}_{dSYN-P}$ aims at *borrowing strength* from the other domains.

A SYN estimator (6.18) using a type (1b) model can be biased. The bias of $\hat{t}_{dSYN-P}$ is approximated by

$$\text{BIAS}(\hat{t}_{dSYN-P}) = E(\hat{t}_{dSYN-P}) - T_d \doteq -T_{dz}(B_{1d} - B_1),$$

where $B_{1d} = \sum_{k\in U_d} y_k / \sum_{k\in U_d} z_k$ is the domain-specific slope, $d = 1, \ldots, D$, and $B_1 = \sum_{k\in U} y_k / \sum_{k\in U} z_k$ is the slope for the whole population. For a given domain, the bias is negligible if the domain slope closely approximates the population slope. But a substantial bias can be encountered if this condition does not hold.

The corresponding indirect GREG estimator (6.2) for domain totals T_d is given by

$$\begin{aligned}\hat{t}_{dGREG-P} &= \sum_{k\in U_d} \hat{y}_k + \sum_{k\in s_d} w_k(y_k - \hat{y}_k) = \hat{t}_{dSYN-P} + \sum_{k\in s_d} w_k(y_k - \hat{b}_1 z_k) \\ &= \hat{t}_{dHT} + \frac{\hat{t}_{HT}}{\hat{t}_{zHT}}(T_{dz} - \hat{t}_{dzHT}) \qquad (6.19)\end{aligned}$$

mimicking the regression estimator for the whole population, but the underlying model is different. Note that an attempt to 'borrow strength' also holds for the indirect GREG estimator.

The *direct* SYN and GREG estimators of type (2b) use y-values from the given domain only. The estimators are obtained by replacing $\hat{b}_1$ by domain-specific counterparts $\hat{b}_{1d}$ given by

$$\hat{b}_{1d} = \frac{\sum_{k\in s_d} w_k y_k}{\sum_{k\in s_d} w_k z_k} = \frac{\hat{t}_{dHT}}{\hat{t}_{dzHT}}, \quad d = 1, \ldots, D,$$

where $\hat{t}_{dHT}$ and $\hat{t}_{dzHT}$ are HT estimators of totals T_d and T_{dz} at the domain level. The direct SYN estimator $\hat{t}_{dSYN-D}$ hence is

$$\hat{t}_{dSYN-D} = \sum_{k\in U_d} \hat{y}_k = \sum_{k\in U_d} \hat{b}_{1d} z_k = T_{dz}\hat{b}_{1d} = T_{dz} \times \hat{t}_{dHT}/\hat{t}_{dzHT}, \quad d = 1, \ldots, D. \quad (6.20)$$

For this model specification, the direct GREG counterpart $\hat{t}_{dGREG-D}$ coincides with the SYN estimator because the second term in GREG estimator (6.2) vanishes.

Let us consider the relative properties of the estimators (6.18) and (6.20) with respect to bias, precision and accuracy. First, the indirect estimator $\hat{t}_{dSYN-P}$ given by (6.18) is biased, and the bias can be substantial if the model assumption does not hold in a given domain. The direct counterpart $\hat{t}_{dSYN-D}$ given by (6.20), which coincides with the GREG estimator $\hat{t}_{dGREG-D}$, is almost design unbiased, irrespective of the validity of the model assumption. The variance of the indirect estimator (6.18) is of the order n^{-1} and thus can be small even in a small domain if the total sample size n is large. On the other hand, the variance of the direct

estimator (6.20) is of the order n_d^{-1} and becomes large when the sample size n_d in domain d is small. Thus, there is a trade-off between bias and precision, depending on the validity of the model assumption and the domain sample size. Using the mean squared error, $\mathrm{MSE}(\hat{t}_d) = V(\hat{t}_d) + \mathrm{BIAS}^2(\hat{t}_d)$, we can conclude the following. In small domains, the indirect estimator (6.18) can be more accurate than the direct counterpart (6.20) because the variance of (6.20) can be very large. But for large domains (with large domain sample size), the direct estimator can be more accurate, because the squared bias of (6.18) can dominate. This holds especially if the model assumption is violated (this trade-off is examined in more detail, for example, in Lehtonen *et al.* 2003).

In Example 6.2, we study selected estimators for domain totals for a single SRSWOR sample drawn from the OHC Survey data set. In Section 6.4, we examine in more detail the relative properties (bias and accuracy) of the synthetic and generalized regression estimators under different model choices. There, the methods are investigated by Monte Carlo simulation techniques, where a large number of independent SRSWOR samples are drawn from a fixed population.

Example 6.2

Estimation of domain totals by design-based methods under SRSWOR sampling. We illustrate the domain estimation methodology by selecting an SRSWOR sample ($n = 1960$ persons) from the OHC Survey data set ($N = 7841$ persons) and estimating the total number of chronically ill persons in the $D = 30$ domains constructed. In the population, the sizes of the domains vary with a minimum of 81 persons and a maximum of 517 persons. The domain proportion of chronically ill persons varies from 18 to 39%, and the overall proportion is 29%. The intra-domain correlation of being chronically ill (binary response) and the age (in years) varies from 0.08 to 0.55; the overall correlation is 0.28.

In the sampling procedure, we consider the domains as unplanned type. Thus, the domain sample sizes are not fixed in the sampling design but are random variates. A Horvitz-Thompson estimator is first calculated. Auxiliary data are then incorporated into the estimation procedure by using the model-assisted GREG estimator given by (6.2). Values of the auxiliary variable z are measurements of age, being available for all persons in the OHC data set, which we, for this example, assume to constitute the population of interest. Therefore, in this hypothetical situation the domain totals T_d of the study variable y also are known for all domains $d = 1, \ldots, D$, and can be used when comparing the estimates of domain totals.

A simple model (1b) from Example 6.2, given by $y_k = \beta \times z_k + \varepsilon_k$, postulates a uniform ratio $R = T/T_z (= 7.778 \times 10^{-3})$ for all domains. Thus, a GREG estimator built on this P-model is of indirect type. On the basis of the SRSWOR sample of $n = 1960$ elements, an estimate of the ratio R is $\hat{r} = \hat{t}_{HT}/\hat{t}_{zHT} = 7.651 \times 10^{-3}$, where $\hat{t}_{HT} (= 2252.3)$ is the HT estimator of the total T of the study variable y and $\hat{t}_{zHT} (= 294357.5)$ is that of the total T_z of the auxiliary variable z. The predicted y-values are calculated by $\hat{y}_k = \hat{r} \times z_k, k = 1, \ldots, 7841$. Alternative

expressions of the estimators are summarized in (6.21). There, the sampling weights are $w_k = N/n = 7841/1960 = 4.001$, T_{dz} are the known domain totals of the auxiliary variable z and $\hat{t}_{dzHT} = \sum_{k \in s_d} w_k z_k$ are the corresponding HT estimates.

$$\begin{aligned} \hat{t}_{dHT} &= \textstyle\sum_{k \in s_d} w_k y_k = N/n \sum_{k \in s_d} y_k \\ \hat{t}_{dGREG-P} &= \textstyle\sum_{k \in U_d} \hat{y}_k + \sum_{k \in s_d} w_k (y_k - \hat{y}_k) = \hat{t}_{dHT} + \hat{r}(T_{dz} - \hat{t}_{dzHT}), \end{aligned} \tag{6.21}$$

where s_d (with n_d elements) and U_d (with N_d elements) are the sets of the sample and the population elements belonging in domain d respectively and $d = 1, \ldots, D$. Note that the corresponding indirect synthetic estimator is $\hat{t}_{dSYN-P} = \sum_{k \in U_d} \hat{y}_k = T_{dz} \times \hat{r}$, which is based on the same simple model as the GREG estimator.

In the examination of the accuracy, we use the estimated standard error $\text{s.e}(\hat{t}_d)$ and percentage coefficient of variation $\text{c.v}(\hat{t}_d)\% = 100 \times \text{s.e}(\hat{t}_d)/\hat{t}_d$ of an estimator $\hat{t}_d$. The variance estimators used are as follows:

$$\begin{aligned} \hat{v}_{srs}(\hat{t}_{dHT}) &= N^2 \left(1 - \frac{n}{N}\right) \left(\frac{1}{n}\right) p_d \hat{s}^2_{dy} \left(1 + \frac{q_d}{\text{c.v}^2_{dy}}\right), \text{ and} \\ \hat{v}_{srs}(\hat{t}_{dGREG-P}) &= N^2 \left(1 - \frac{n}{N}\right) \left(\frac{1}{n}\right) p_d \hat{s}^2_{d\hat{e}} \left(1 + \frac{q_d}{\text{c.v}^2_{d\hat{e}}}\right), \end{aligned} \tag{6.22}$$

where $p_d = n_d/n, q_d = 1 - p_d$, variance estimators are $\hat{s}^2_{dy} = \sum_{k \in s_d} (y_k - \overline{y}_d)^2/(n_d - 1)$ and $\hat{s}^2_{d\hat{e}} = \sum_{k \in s_d} (\hat{e}_k - \overline{\hat{e}}_d)^2/(n_d - 1)$, estimated coefficients of variation are $\text{c.v}_{dy} = \hat{s}_{dy}/\overline{y}_d$ and $\text{c.v}_{d\hat{e}} = \hat{s}_{d\hat{e}}/\overline{\hat{e}}_d$, where $\overline{y}_d = \sum_{k \in s_d} y_k/n_d$ and $\overline{\hat{e}}_d = \sum_{k \in s_d} \hat{e}_k/n_d$, and residuals are $\hat{e}_k = y_k - \hat{r} \times z_k$.

In the realized sample, domain sample sizes vary from 24 to 132 elements and the mean size is 65. The situation thus is realistic for design-based estimation for domain totals. We first examine the average performance of the Horvitz-Thompson estimator $\hat{t}_{dHT}$ and the indirect GREG estimator $\hat{t}_{dGREG-P}$. In the first part of Table 6.5, a simple average measure $|\overline{\hat{t}} - \overline{T}|/\overline{T}$ of absolute relative difference is calculated in three domain sample size classes, where $\overline{\hat{t}}$ is the mean of the estimated domain totals $\hat{t}_d$ and $\overline{T}$ is the mean of the true values T_d in a given size class. Absolute relative differences of the HT and GREG estimates tend to decrease with increasing domain sample size, and for a given size class, the figures closely coincide. The realized domain sample size and coefficient of variation have a clear association for GREG and HT estimators: sample coefficients of variation tend to decrease with increasing domain sample size, as is indicated in the average coefficient of variation figures given in the second part of Table 6.5. On average, estimated coefficients of variation are smaller for the GREG estimator.

Domain-wise point estimates, standard errors and coefficients of variation for the 30 domains are given in Table 6.6 in which the domains are sorted by the domain sample size. When compared to the HT estimator $\hat{t}_{dHT}$, use of auxiliary information by the model-assisted GREG estimator $\hat{t}_{dGREG-P}$ clearly improves

Table 6.5 Average absolute relative difference and average coefficient of variation of Horvitz–Thompson and GREG estimates by domain sample size class.

	Average absolute relative difference (%)		Average coefficient of variation (%)	
Size class	HT estimator	GREG estimator	HT estimator	GREG estimator
−39	10.6	10.2	30.8	24.7
40–79	2.0	3.4	23.5	19.8
80–	3.2	3.7	16.0	13.6
All	1.8	1.7	23.0	19.0

accuracy. In all 30 domains, estimated standard errors of the GREG estimator are smaller than those of the HT estimator. In most domains, estimated coefficients of variation are smaller for the GREG estimator, as expected.

Let us complete the example by considering briefly the relationship of the GREG estimator and the corresponding model-dependent indirect SYN estimator $\hat{t}_{dSYN-P} = T_{dz} \times \hat{r}$ in the context of the realized sample. By the expression (6.21) for the GREG estimator, we obtain for example in the first domain ($n_1 = 41$):

$$\begin{aligned}\hat{t}_{1GREG-P} &= \sum_{k \in U_1} \hat{y}_k + \sum_{k \in s_1} w_k (y_k - \hat{y}_k) \\ &= 45.43 + 4.001 \times (-0.5974) = 43.04,\end{aligned}$$

where the sum of predicted values $\hat{y}_k$ in the first domain is calculated as $\sum_{k \in U_1} \hat{y}_k = T_{1z} \times \hat{r} = 5937 \times 0.0076515 = 45.43$. This is the synthetic estimate $\hat{t}_{1SYN-P}$ for the first domain. And, for example, for domain $d = 19$ ($n_{19} = 115$) we obtain $\hat{t}_{19GREG-P} = 160.00$ and $\hat{t}_{19SYN-P} = 138.09$, whereas the true value is $T_{19} = 165$. The bias-adjustment term of the GREG estimator thus happens to adjust successfully the bias of the SYN estimator for these domains. But this does not necessarily hold for all domains. In fact, the GREG estimator is more successful than the SYN estimator in 17 out of 30 domains because in several domains, the bias correction affects to a correct direction but too strongly. In the estimation of the accuracy of the SYN estimator, an estimated mean squared error (MSE) should be used because the SYN estimator is not design unbiased. We will consider the relationship of the GREG and SYN estimators for domain totals in more detail in Section 6.4 and further, in the web extension of the book.

6.4 FURTHER COMPARISON OF ESTIMATORS

In this section, we examine further the properties of model-dependent estimators and model-assisted estimators for domain totals using Monte Carlo simulation methods. For this exercise, we again use the OHC Survey data set. To examine empirically the theoretical properties (bias and accuracy) of the different SYN and

Table 6.6 Estimates of the total number of chronically ill persons in domains calculated for an SRSWOR sample ($n = 1960$) from the OHC data set. Domain sample sizes n_d, domain sizes N_d, population totals T_d, and point estimates, standard error estimates and coefficient of variation estimates (%) for HT and GREG estimators, by domain sample size class.

Domain				Estimate of total		Standard error		Coefficient of variation (%)	
d	n_d	N_d	T_d	$\hat{t}_{dHT}$	$\hat{t}_{dGREG}$	s.e$(\hat{t}_{dHT})$	s.e$(\hat{t}_{dGREG})$	c.v$(\hat{t}_{dHT})$	c.v$(\hat{t}_{dGREG})$
Domain sample size n_d < 40									
20	24	101	31	32.0	31.6	9.77	7.13	30.5	22.5
10	26	81	27	32.0	25.6	10.83	8.05	33.8	31.5
18	26	129	36	20.0	27.2	7.60	6.95	38.0	25.5
23	31	156	57	44.0	53.2	10.82	9.10	24.6	17.1
8	35	141	29	24.0	24.5	8.57	7.88	35.7	32.2
30	36	146	34	32.0	33.8	9.86	8.56	30.8	25.3
3	37	133	29	36.0	32.6	10.77	8.73	29.9	26.8
16	37	165	45	52.0	54.8	12.14	9.15	23.3	16.7
Domain sample size 40 ≤ n_d < 80									
1	41	181	33	40.0	43.0	10.80	9.15	27.0	21.3
21	43	153	48	64.0	55.3	14.55	10.93	22.7	19.8
6	45	188	52	24.0	26.6	8.51	7.67	35.5	28.9
28	51	194	74	88.0	85.4	16.61	11.65	18.9	13.6
24	53	200	55	56.0	55.7	13.21	11.06	23.6	19.9
22	57	242	96	112.0	115.0	17.79	13.08	15.9	11.4
15	58	252	61	60.0	66.4	13.20	11.90	22.0	17.9
11	59	187	47	52.0	39.5	13.30	10.89	25.6	27.6
13	69	305	89	80.0	88.5	15.10	12.86	18.9	14.5
12	73	311	95	56.0	65.9	12.85	11.40	22.9	17.3
4	76	295	65	68.0	68.1	14.39	12.17	21.2	17.9
7	78	292	52	40.0	36.3	11.09	10.17	27.7	28.0
Domain sample size n_d ≥ 80									
2	84	352	86	76.0	78.6	14.95	13.49	19.7	17.2
5	86	323	66	76.0	70.5	15.31	13.62	20.1	19.3
26	89	364	124	124.0	126.0	19.07	15.72	15.4	12.5
29	90	365	128	124.0	124.5	19.12	15.10	15.4	12.1
25	91	339	114	112.0	101.6	18.68	14.81	16.7	14.6
17	99	426	139	176.0	183.3	22.11	16.72	12.6	9.1
9	103	366	89	88.0	79.3	16.66	13.82	18.9	17.4
19	115	490	165	152.0	160.0	20.81	17.13	13.7	10.7
14	116	447	130	136.0	128.4	20.31	16.28	14.9	12.7
27	132	517	197	176.0	173.8	22.94	17.51	13.0	10.1
All	1960	7841	2293	2252.3	2254.8	69.42	66.88	3.1	3.0

GREG estimators for domains, we make the following conventions. First, similarly as in Example 6.2, we consider the OHC data set as a frame population of size 7841 elements, such that the necessary auxiliary data are included at micro-level in the data set. Secondly, we construct for the population frame data set a domain structure involving 60 domains in total. This is because we want to consider also domains with a small sample size. Finally, we will draw a large number of independent SRSWOR samples of 1000 elements from the constructed artificial frame population under an unplanned domain structure. We study the bias and accuracy of estimators on the basis of the average figures calculated over the simulated samples.

We assume (according to the principles presented in Box 6.1) that the constructed OHC frame population of $N = 7841$ persons and $D = 60$ domains includes unique identification keys, domain membership indicators, inclusion probabilities for all elements $k \in U$ for a SRSWOR sample of $n = 1000$ elements and values of the auxiliary z-variable age (in years). The binary response variable y to be measured from the sample elements is chronic illness (value 0: No, 1: Yes).

P-models and D-models are used for the indirect SYN and GREG estimators based on linear models of the general form $y_k = \beta_0 + u_{0d} + \beta_1 z_k + \varepsilon_k$. In the mixed D-model case, model parameters are estimated by restricted maximum likelihood (REML) and generalized least squares (GLS), and predictions $\hat{y}_k = \hat{\beta}_0 + \hat{u}_{0d} + \hat{\beta}_1 z_k$, $k \in U$, are calculated. For a fixed-effects P-model, estimation is based on ordinary least squares (OLS), and predictions are calculated as $\hat{y}_k = \hat{b}_0 + \hat{b}_1 z_k$, $k \in U$. Residuals are calculated as $\hat{e}_k = y_k - \hat{y}_k$, $k \in s$, in both cases. By micro-merging these data in the frame population U (see Table 6.3), the data are successfully completed for domain estimation.

Domain totals to be estimated are given by

$$T_d = \sum_{k \in U_d} Y_k, \quad d = 1, \ldots, D.$$

The indirect estimators to be used are the following:

$$\hat{t}_{dSYN} = \sum_{k \in U_d} \hat{y}_k, \quad d = 1, \ldots, D \text{ (synthetic estimator), and}$$

$$\hat{t}_{dGREG} = \sum_{k \in U_d} \hat{y}_k + \sum_{k \in s_d} w_k (y_k - \hat{y}_k), \; d = 1, \ldots, D$$

(generalized regression estimator).

In these formulas, the predicted values $\hat{y}_k$, $k \in U$, and observed y-data y_k, sampling weights w_k and residuals $\hat{e}_k$, $k \in s$, provide the materials for the calculation of estimates for domain totals. The indirect estimators use fixed-effects P-models and mixed D-models. For the synthetic estimators $\hat{t}_{dSYN-P}$ and

$\hat{t}_{dMSYN-D}$, only the predictions $\hat{y}_k$ are used. And for the GREG estimators $\hat{t}_{dGREG-P}$ and $\hat{t}_{dMGREG-D}$, predicted values $\hat{y}_k$, observed y-data y_k, sampling weights w_k and residuals $\hat{e}_k = y_k - \hat{y}_k$ are used. In the SRSWOR case considered here, the weights $w_k = N/n$ are constants, and the sum of residuals over the whole sample data set is $\sum_{k \in s} \hat{e}_k = 0$. Note that this does not necessarily hold for the domains because we work with indirect estimators of domain totals.

We compare the bias and accuracy of the various estimators by using estimates $\hat{t}_d(s_v)$ from the K repeated Monte Carlo samples $s_v; v = 1, 2, \ldots, K$. For each domain $d = 1, \ldots, D$, the following Monte Carlo summary measures of bias and accuracy are computed. We use two measures of accuracy, the relative root mean squared error (RRMSE) and the median absolute relative error (MdARE), because for a binary response variable there is sometimes a difference in the conclusions drawn from the two measures.

(i) Absolute relative bias (ARB), defined as the ratio of the absolute value of bias to the true value:

$$\left| \frac{1}{K} \sum_{v=1}^{K} \hat{t}_d(s_v) - T_d \right| / T_d.$$

(ii) Relative root mean squared error (RRMSE), defined as the ratio of the root MSE to the true value:

$$\sqrt{\frac{1}{K} \sum_{v=1}^{K} (\hat{t}_d(s_v) - T_d)^2} / T_d.$$

(iii) Median absolute relative error (MdARE) is defined as follows. For each simulated sample $s_v; v = 1, 2, \ldots, K$, the absolute relative error is calculated and a median is taken over the K samples in the simulation:

$$\underset{\text{over } v = 1, \ldots, K}{\text{Median}} \{|\hat{t}_d(s_v) - T_d|/T_d\}.$$

A summary of the features of the experimental design used in this simple exercise is given in Table 6.7.

A summary of the results for the simple models (1a) and (2a) is presented in Part A of Table 6.8 and for the more complex models (1b) and (2b) in Part B of the table. The results indicate that the bias, measured by the average of absolute relative bias ARB, of the GREG estimators GREG-P and MGREG-D is negligible for all models and in all size classes. The bias for the SYN-type estimators varies with the model choice. The bias of SYN-P is substantial for the extremely simple fixed-effects P-model (1a), and the bias decreases when the more realistic fixed-effects model (1b) is used. A similar effect is noticed for the mixed models (2a) and (2b), which provides the smallest bias figures for SYN estimators. Especially

Table 6.7 Summary of technical details of Monte Carlo experiments.

Population:

OHC Survey frame population of size $N = 7841$ persons

Sample size: $n = 1000$ persons

Number of domains: $D = 60$ areas

Number of simulated samples:

$K = 500$ independent SRSWOR samples (unplanned domain structure)

Response variable y:

Chronic illness (binary; $0 = \text{No}, 1 = \text{Yes}$)

Auxiliary z**-data**:

Domain membership indicators

Age (in years)

Models:

(1a) Linear fixed-effects P-model with intercept only:

$y_k = \beta_0 + \varepsilon_k$

(1b) Linear fixed-effects P-model with age as the predictor:

$y_k = \beta_0 + \beta_1 z_k + \varepsilon_k$

(2a) Linear mixed D-model with random intercepts:

$y_k = \beta_0 + u_{0d} + \varepsilon_k$

(2b) Linear mixed D-model with age as the predictor:

$y_k = \beta_0 + u_{0d} + \beta_1 z_k + \varepsilon_k$

Target parameters:

Domain totals T_d of chronically ill people, $d = 1, \ldots, 60$

Estimators of domain totals:

SYN estimators:

$\hat{t}_{dSYN-P}$ using a linear fixed-effects P-model

$\hat{t}_{dMSYN-D}$ using a two-level linear D-model

GREG estimators:

$\hat{t}_{dGREG-P}$ using a linear fixed-effects P-model

$\hat{t}_{dMGREG-D}$ using a two-level linear D-model

Measures of performance:

Averages calculated over domain size classes of:

ARB Absolute relative bias

RRMSE Relative root mean squared error

MdARE Median absolute relative error

in small domains, the accuracy is better for SYN estimators when compared to GREG estimators, in all model types and with both measures RRMSE and MdARE. But as soon as the domain sample size increases, the difference in accuracy tends to decrease.

The results in Table 6.8 also indicate that the model improvement, that is, moving from a 'weak' model towards a 'stronger' model, is much more prominent for SYN-type estimators than for GREG-type estimators. Note that for this estimation exercise we needed an access to the micro-merged frame population and sample data set. An access to these data is provided by the web extension of the book.

6.5 CHAPTER SUMMARY AND FURTHER READING

In this chapter, we concentrated on design-based model-assisted estimation for domains. This approach is frequently used, for example, in the production of official statistics. We made several assumptions for the treatment of estimation for domain totals. In particular, we assumed that in a given statistical infrastructure, registers

Table 6.8 Simulation results for SYN and GREG estimators for domain totals of chronically ill people with different model choices ($K = 500$ independent SRSWOR samples with $n = 1000$ elements in each).

A. Fixed-effects P-model $y_k = \beta_0 + \varepsilon_k$ and mixed D-model $y_k = \beta_0 + u_{0d} + \varepsilon_k$.

		Average over domains of					
Estimator	Domain sample size class	Domain total in population	Estimate of domain total	Absolute relative bias ARB%	Relative root MSE RRMSE%	Median absolute relative error MdARE%	Domain sample size
SYN-P	0–10	17.5	13.7	36.9	37.4	37.0	5.6
	11–20	37.0	34.4	50.3	50.7	50.3	14.1
	21–	62.4	78.8	43.6	44.2	43.6	32.4
	All	38.2	41.2	43.5	44.0	43.5	16.9
MSYN-D	0–10	17.5	14.9	25.1	33.0	27.9	5.6
	11–20	37.0	35.7	22.7	33.3	25.0	14.1
	21–	62.4	66.3	11.6	26.0	17.4	32.4
	All	38.2	38.2	20.0	30.9	23.6	16.9
GREG-P	0–10	17.5	17.5	2.4	55.2	39.5	5.6
	11–20	37.0	37.0	1.6	40.7	27.8	14.1
	21–	62.4	62.4	1.1	31.1	20.8	32.4
	All	38.2	38.2	1.7	42.8	29.7	16.9
MGREG-D	0–10	17.5	17.3	2.6	53.5	38.9	5.6
	11–20	37.0	37.0	1.9	39.5	27.3	14.1
	21–	62.4	62.5	1.1	30.3	20.2	32.4
	All	38.2	38.2	1.9	41.5	29.1	16.9
B. Fixed-effects P-model $y_k = \beta_0 + \beta_1 z_k + \varepsilon_k$ and mixed D-model $y_k = \beta_0 + u_{0d} + \beta_1 z_k + \varepsilon_k$.							
SYN-P	0–10	17.5	18.0	27.0	28.1	27.1	5.6
	11–20	37.0	36.6	19.6	20.8	19.7	14.1
	21–	62.4	62.0	12.1	13.9	12.5	32.4
	All	38.2	38.1	19.8	21.2	20.0	16.9
MSYN-D	0–10	17.5	18.0	25.9	27.5	26.4	5.6
	11–20	37.0	36.6	17.7	20.2	18.5	14.1
	21–	62.4	62.1	9.7	14.4	11.6	32.4
	All	38.2	38.2	18.1	20.9	19.1	16.9
GREG-P	0–10	17.5	17.5	2.7	53.0	38.5	5.6
	11–20	37.0	37.0	1.4	38.9	26.5	14.1
	21–	62.4	62.5	1.1	30.0	20.2	32.4
	All	38.2	38.2	1.8	41.0	28.7	16.9
MGREG-D	0–10	17.5	17.5	2.7	52.8	38.4	5.6
	11–20	37.0	37.0	1.5	38.8	26.4	14.1
	21–	62.4	62.5	1.0	29.8	20.2	32.4
	All	38.2	38.2	1.8	40.8	28.6	16.9

are available as frame populations and sources of micro-level and aggregate-level auxiliary data, and unique identification keys are available in order to merge the data from a sample survey with data from a statistical register. We believe that fulfilling these conditions can provide much flexibility for sampling design and estimation for domains. For example, the data can then be aggregated at higher levels of the population if desired. The use of unit-level data and unit-level modelling can be beneficial for both design-based model-assisted estimation and model-dependent estimation for domains. It appeared that careful and realistic modelling is especially important in model-dependent estimation for domains. This was demonstrated by a small-scale simulation study. The materials discussed in the examples of this chapter will be worked out further in the web extension of the book.

In practice, design-based model-assisted estimation is most often used for domains whose sample size is reasonably large. For small domains, methods of small-area estimation are used instead. For the estimation for domains, it is recommended to define, if possible, the intended domains as strata in the sampling phase, and to use a suitable allocation scheme, such that a reasonably large sample size is attained for all domains. And in the estimation phase it is advisable to incorporate strong auxiliary data into the estimation procedure by using carefully chosen models.

Supplementing the references mentioned earlier in this chapter, design-based model-assisted estimation for domains is discussed, for example, in Estevao *et al.* (1995) and Estevao and Särndal (1999). Lehtonen and Veijanen (1998) discuss nonlinear GREG estimators, such as a multinomial logistic GREG estimator.

In addition to Rao (2003), model-dependent methods for small area estimation are presented in Ghosh and Rao (1994) and Rao (1999). You and Rao (2002) discuss pseudo EBLUP estimators involving survey weights. Underlying models and their features is a prominent theme in recent literature (Ghosh *et al.* 1998; Marker 1999; Moura and Holt 1999; Prasad and Rao 1999; Feder *et al.* 2000). There is extensive recent literature on small area estimation from a Bayesian point of view, including empirical Bayes and hierarchical Bayes techniques (Datta *et al.* 1999; Ghosh and Natarajan 1999; You and Rao 2000). Some recent publications relate frequentist and Bayesian approaches in small area estimation (Singh *et al.* 1998). Valliant *et al.* (2000) discuss small-area estimation under a prediction approach.

7

Analysis of One-way and Two-way Tables

One-way and two-way frequency tables commonly occur in the analysis of complex surveys. Such tables are formed by tabulating the available survey data by a categorical variable or by cross-classifying two categorical variables with the aim being to test the hypotheses of goodness of fit, homogeneity or independence. For example, goodness of fit of the age distribution of the MFH Survey subgroup of 30–64-year-old males can be studied relative to the respective population age distribution. Or the OHC Survey data set may be tabulated by sex of respondent and a binary response variable CHRON (chronic morbidity) in a 2×2 table, with a null hypothesis of homogeneity of CHRON proportions in males and females stated. Further, we may consider an independence hypothesis of response variables CHRON and a categorical variable formed by classifying PSYCH (first principal component of psychic—psychological or mental—symptoms) into a number of classes. Under simple random sampling, valid inferences for these hypotheses can be based on a standard Pearson chi-squared test statistic. But with more complex designs, the testing procedures are more complicated because of clustering effects.

For homogeneity and independence hypotheses on an $r \times c$ frequency table from simple random sampling, the Pearson test statistic is asymptotically chi-squared with $(r-1)(c-1)$ degrees of freedom. But this standard asymptotic property is not valid for a frequency table from a complex survey based on cluster sampling. Positive intra-cluster correlation of the variables used in forming the table causes the test to be overly liberal relative to nominal significance levels. Therefore, the observed values of the test statistic can be too large, which can lead to erroneous inferences.

For valid inferences in complex surveys, certain corrections to the Pearson test statistic have been suggested such as Rao–Scott adjustments or, alternatively, test statistics such as the Wald test statistic can be used, which automatically account

Practical Methods for Design and Analysis of Complex Surveys Risto Lehtonen and Erkki Pahkinen
 ISBN: 0-470-84769-7

for the clustering. Both approaches are demonstrated with an introductory example for a simple goodness-of-fit test in Section 7.1. The goodness-of-fit test is further considered in Section 7.2. The basics of testing for two-way tables are presented in Section 7.3. In Section 7.4, test statistics for a homogeneity hypothesis in a two-way table are examined, and in Section 7.5, a test of independence of two categorical variables is considered. The OHC and MFH Surveys involving clustered designs, described in Chapter 5, are used in the examples.

7.1 INTRODUCTORY EXAMPLE

Binomial Test and Effective Sample Size

Let us consider a hypothetical example of a simple goodness-of-fit test, basically originating from Sudman (1976), also illustrated in Rao and Thomas (1988), but applied here for the OHC Survey setting. A sample of $m = 50$ clusters is drawn from a large population of clusters which are industrial establishments. Let us assume that in each sample cluster $i = 1, \ldots, 50$, there are $n_i = 20$ employees. The element sample size is thus $n = 1000$. Given appropriate data under this sampling design, one might want to study whether the coverage of occupational health care (OHC), i.e. the unknown population proportion p of workers having access to occupational health (OH) services, is 80% based on prior knowledge from the previous year. The null hypothesis $H_0 : p = p_0 = 0.8$ can thus be stated. Let the significance level for this test be chosen as $\alpha = 5\%$.

A survey estimate $\hat{p} = n_1/n = 0.84$ is obtained, where $n_1 = 840$ is the number of sample workers having access to OH services. The binomial test is chosen, to be referred to the standard normal $N(0, 1)$ distribution, with a large-sample test statistic

$$Z = |\hat{p} - p_0| \Big/ \sqrt{p_0(1 - p_0)/n}, \tag{7.1}$$

where the denominator is the standard error of the estimate $\hat{p}$ under the null hypothesis. We calculate the value of Z with an assumption of simple random sampling with replacement and also using a design-based approach that takes the clustering into account. In this simple case, the standard error of $\hat{p}$, needed for the calculation of an observed value of Z, is, for both approaches, based on a binomial assumption but with different sample sizes.

In a test based on the assumption of simple random sampling, we ignore the clustering and use the actual sample size $n = 1000$ in the standard error formula. The observed value of the test statistic (7.1) is hence

$$Z_{bin} = |\hat{p} - p_0| \Big/ \sqrt{p_0(1 - p_0)/1000} = 3.162 > Z_{0.025} = 1.96,$$

where $\sqrt{0.8(1 - 0.8)/1000} = 0.0126$ is the corresponding standard error of $\hat{p}$. The result obviously suggests rejecting the null hypothesis when compared against the appropriate critical value from a standard $N(0, 1)$ distribution.

It appeared that if an establishment is covered by OHC, then each worker at that site has equal access to OH services, which is an important piece of information that was ignored in the previous test. In fact, taking more than one person from a sample establishment does not increase our knowledge of the coverage of OHC at that site. Therefore, the effective sample size is $\overline{n} = 50$ in contrast to the assumed 1000 in the previous test. Recall that the concept of effective sample size refers to the size of a simple random sample, which gives an equally precise estimate for an unknown parameter p as that given by a sample of $n = 1000$ persons from the actual cluster sample design.

By using the effective sample size, we have for a design-based test,

$$Z_{des} = |\hat{p} - p_0| \Big/ \sqrt{p_0(1 - p_0)/50} = 0.707,$$

where $\sqrt{0.8(1 - 0.8)/50} = 0.0566$, which is much larger than the corresponding standard error from the previous test. Therefore, the observed value of Z_{des} is smaller than that of Z_{bin}, and our test now suggests that the null hypothesis should not be rejected. We shall next study this example in a slightly more general setting and introduce alternative test statistics in which the effect of clustering can be successfully removed.

Pearson Test Statistic and Rao–Scott Adjustment

The binomial test statistic Z_{bin} appeared to be liberal when compared to the design-based counterpart Z_{des}. This is because, with Z_{bin}, the clustering is not taken into account. Let us examine the asymptotic behaviour of the test statistic Z_{bin} more closely by constructing the corresponding *Pearson test statistic* X^2_{P}. For this, the following frequency table is used, where n_j are the observed cell frequencies and p_{0j} are the hypothesized cell proportions:

j	n_j	p_{0j}
1	840	0.8
2	160	0.2
All	1000	1.0

In a finite-population framework, let the unknown cell proportions be $p_j = N_j/N$, on the basis of a population of N elements, where N_j is the number of population elements in cell j. The p_j can also be taken as the unknown cell probabilities under a superpopulation framework. The Pearson test statistic for the simple goodness-of-fit hypothesis $H_0 : p_j = p_{0j}$, $j = 1, 2$, is given by

$$X^2_{\mathrm{P}} = \sum_{j=1}^{2} (n_j - np_{0j})^2/(np_{0j}) = n \sum_{j=1}^{2} (\hat{p}_j - p_{0j})^2/p_{0j}, \tag{7.2}$$

where the proportions $\hat{p}_j = n_j/n$ are estimates of the parameters p_j with n_j being the sample value of N_j. In the case of two cells, $\hat{p}_2 = 1 - \hat{p}_1$ and $p_{02} = 1 - p_{01}$, and an analogy exists between the statistics Z_{bin} and X^2_{P}:

$$X^2_{\mathrm{P}} = n\sum_{j=1}^{2}(\hat{p}_j - p_{0j})^2/p_{0j} = (\hat{p} - p_0)^2/(p_0(1-p_0)/n) = Z^2_{bin},$$

where $\hat{p} = \hat{p}_1$ and $p_0 = p_{01}$. With two cells, there is one degree of freedom for the goodness-of-fit test statistic X^2_{P} because of one constraint (the proportions must sum up to one), and no parameters need to be estimated.

Rao and Scott (1981) have given general results about the asymptotic distribution of the Pearson test statistic X^2_{P}. With two cells, the test statistic X^2_{P} is asymptotically distributed as a random variate dW, where W is distributed as a chi-squared random variate χ^2_1 with one degree of freedom, and d denotes the design effect of the proportion estimate $\hat{p}$. The design effect can be obtained from $d = V_{des}(\hat{p})/V_{bin}(\hat{p})$, where $V_{des}(\hat{p}) = p_0(1-p_0)/\overline{n}$ is the design variance of the estimate $\hat{p}$, $\overline{n}$ denotes the effective sample size, and $V_{bin}(\hat{p}) = p_0(1-p_0)/n$ is the standard binomial variance counterpart. Hence, in this case, the design effect reduces to $d = n/\overline{n}$, which also confirms that the effective sample size is $\overline{n} = n/d$.

If the sample of employees had actually been drawn with simple random sampling directly from the employee population, we would have $d = 1$ because V_{des} and V_{bin} would then be equal. In this case, for two cells, the Pearson test statistic X^2_{P} would be asymptotically chi-squared with one degree of freedom. But if the sample is actually drawn under cluster sampling, positive intra-cluster correlation gives a design effect d greater than one. Owing to this, the statistic X^2_{P} is no longer asymptotically chi-squared with the appropriate degrees of freedom.

Being now aware of the consequences of positive intra-cluster correlation on the asymptotic distribution of the Pearson test statistic X^2_{P}, the next step is to derive a valid testing procedure. Because, in general, accounting for intra-cluster correlation cannot be incorporated in the formula for X^2_{P}, an external correction to X^2_{P} must be made. For this purpose, first note that the asymptotic expectation of X^2_{P} is $E(X^2_{\mathrm{P}}) = d$, which under positive intra-cluster correlation is greater than the nominal expected value of one. Since $E(X^2_{\mathrm{P}}/d) = E(\chi^2_1) = 1$, we can construct a simple *Rao–Scott correction* to X^2_{P} by dividing the observed value of the test statistic by the design effect. The resulting test statistic adjusted for the clustering effect is given by

$$X^2_{\mathrm{P}}(d) = X^2_{\mathrm{P}}/d \tag{7.3}$$

and is asymptotically chi-squared with one degree of freedom in the case of two cells.

An analogous adjustment can be made to the corresponding *likelihood ratio* (LR) *test statistic* X^2_{LR} of goodness of fit, which in the case of two cells is

$$X^2_{\mathrm{LR}} = 2n \sum_{j=1}^{2} \hat{p}_j \log(\hat{p}_j / p_{0j}). \tag{7.4}$$

Under simple random sampling, the statistic X^2_{LR} is also asymptotically chi-squared with one degree of freedom when the null hypothesis is true. For clustered designs, the corresponding adjusted test statistic is

$$X^2_{\mathrm{LR}}(d) = X^2_{\mathrm{LR}}/d, \tag{7.5}$$

which is asymptotically chi-squared with one degree of freedom.

We next compute the values of the Pearson and LR test statistics, with their Rao–Scott adjustments, for the OHC Survey setting. For the adjustments, the observed design effect is required, and this is

$$d = V_{des}(\hat{p})/V_{bin}(\hat{p}) = 0.0032/0.00016 = 20,$$

which can also be calculated as $d = n/\overline{n} = 1000/50 = 20$.

For the Pearson test statistic, we obtain

$$X^2_{\mathrm{P}} = (0.84 - 0.80)^2/(0.80 \times 0.20/1000) = 10.00$$

with a p-value of 0.0016. The value of the Rao–Scott corrected Pearson test statistic is hence

$$X^2_{\mathrm{P}}(d) = X^2_{\mathrm{P}}/d = Z^2_{bin}/d = 3.162^2/20 = 10.00/20 = 0.50,$$

which has a p-value of 0.4795. It can be noticed also that $Z^2_{des} = 0.707^2 = 0.50$, i.e. $Z^2_{des} = X^2_{\mathrm{P}}(d)$ as expected. For the LR test statistic and the corresponding Rao–Scott correction, we obtain

$$X^2_{\mathrm{LR}} = 2 \times 1000 \times (0.84 \times \log(0.84/0.80) + 0.16 \times \log(0.16/0.20)) = 10.56,$$

with a p-value of 0.0012, and

$$X^2_{\mathrm{LR}}(d) = X^2_{\mathrm{LR}}/d = 10.560/20 = 0.528$$

with a p-value of 0.4675.

The observed design effect $d = 20$ is unusually large since the positive intra-cluster correlation is complete. The intra-cluster correlation coefficient is thus

$\rho_{int} = 1$, calculated from the equation $d = 1 + (\overline{m} - 1)\rho_{int}$, where $\overline{m} = 20$ is the average cluster size. In practice, intra-cluster correlations are usually positive but less than one, and design-effect estimates $\hat{d}$ are correspondingly greater than one. A typical $\hat{d}$ is less than 3, corresponding to an estimated positive intra-cluster correlation coefficient $\hat{\rho}_{int} < 0.1$ with $\overline{m} = 20$.

Neyman and Wald Test Statistics

As an alternative to the Pearson test statistic, a *Neyman test statistic* X^2_{N} of a simple goodness-of-fit hypothesis can be calculated. In the case of two cells, it reduces to

$$X^2_{\mathrm{N}} = n\sum_{j=1}^{2}(\hat{p}_j - p_{0j})^2/\hat{p}_j = (\hat{p} - p_0)^2/(\hat{p}(1-\hat{p})/n), \tag{7.6}$$

which differs from the Pearson statistic since the estimated proportions $\hat{p}_j$ are inserted in the denominator in place of the hypothetical ones, p_{0j}. With simple random sampling, the Neyman test statistic is asymptotically chi-squared with one degree of freedom in the case of two cells. But under cluster sampling the Neyman test statistic should be adjusted in a similar manner to that used for the Pearson test statistic. The Rao–Scott adjusted Neyman test statistic is hence

$$X^2_{\mathrm{N}}(\hat{d}) = X^2_{\mathrm{N}}/\hat{d} = \hat{d}^{-1}(\hat{p} - p_0)^2/(\hat{p}(1-\hat{p})/n). \tag{7.7}$$

The estimated design effect is calculated by the formula $\hat{d} = \hat{v}_{des}(\hat{p})/\hat{v}_{bin}(\hat{p})$, where $\hat{v}_{des}$ is the design-based variance estimate of $\hat{p}$ corresponding to the actual sampling design and $\hat{v}_{bin}$ is the binomial counterpart.

We next calculate the values of the Neyman test statistic and its Rao–Scott correction. For this, the estimated design effect is used. The design-based variance estimate of $\hat{p}$ is first obtained:

$$\hat{v}_{des}(\hat{p}) = \sum_{i=1}^{m}(\hat{p}_i - \hat{p})^2/(m(m-1)) = \sum_{i=1}^{50}(\hat{p}_i - 0.84)^2/(50 \times 49) = 0.002743,$$

where m is the number of sample clusters, $\hat{p}_i$ is the coverage of OHC in sample cluster i and $\hat{p}$ is the corresponding estimate in the whole sample. It should be noted that $\hat{p}_i$ is either zero or one. A design-effect estimate can be calculated using a binomial variance estimate, which is

$$\hat{v}_{bin}(\hat{p}) = \hat{p}(1-\hat{p})/n = 0.000134,$$

giving an estimated design effect $\hat{d} = 0.002743/0.000134 = 20.4$. Alternatively, the design effect can be estimated as $\hat{d} = \hat{v}_{des}(\hat{p})/V_{bin}(\hat{p}) = 17.1$.

The observed value of the Neyman test statistic is

$$X^2_{\mathrm{N}} = (0.84 - 0.80)^2/(0.84 \times 0.16/1000) = 11.90$$

with a p-value of 0.0006. For the Rao–Scott corrected Neyman test statistic, we obtain

$$X^2_{\mathrm{N}}(\hat{d}) = X^2_{\mathrm{N}}/\hat{d} = 11.9/20.4 = 0.583$$

with a p-value of 0.4451. Note that the observed values of the Neyman test statistic and the corresponding Rao–Scott adjustment are somewhat larger than the values of the Pearson statistic and its Rao–Scott adjustment.

The Neyman test statistic X^2_{N} is a special case of the *Wald* (1943) *test statistic* of goodness of fit. The Wald statistic differs from the Pearson, LR and Neyman test statistics by automatically accounting for intra-cluster correlation. This can be seen in the formula of the design-based Wald statistic, which in the case of two cells reduces to

$$X^2_{des} = (\hat{p} - p_0)^2/\hat{v}_{des}, \tag{7.8}$$

where $\hat{v}_{des}$ is the design-based variance estimate of $\hat{p}$. The statistic X^2_{des} is asymptotically chi-squared with one degree of freedom in the cluster-sampling design considered, without any auxiliary corrections. For a simple random sample, the variance estimator $\hat{v}_{bin}$ is used in (7.8) in place of $\hat{v}_{des}$ and so the Neyman test statistic X^2_{N} and the resulting Wald statistic, denoted by X^2_{bin}, coincide. Obviously, for a clustered design, X^2_{bin} also requires an adjustment similar to that of the Neyman statistic.

When calculating the value of the design-based Wald statistic, we obtain

$$X^2_{des} = (0.84 - 0.80)^2/0.002743 = 0.583,$$

which is equal to the value of the Rao–Scott corrected Neyman statistic, as expected. This demonstrates the flexibility of the Wald statistic. Using an appropriate variance estimate reflecting the complexities of the sampling design, we have an asymptotically valid test statistic without any auxiliary corrections. This can be seen as an obvious advantage over the Rao–Scott corrected statistics, but, as we shall see later, in more general cases when working with more than two cells, there are certain drawbacks to the design-based Wald statistic caused by possible instability in the variance estimates in some small-sample situations.

Finally, we display the test results from the test statistics (7.2)–(7.8) below:

Test statistic	df	Observed value	*p*-value
Pearson			
X^2_{P}	1	10.00	0.0016
$X^2_{\mathrm{P}}(d)$ (adjusted)	1	0.500	0.4795
Likelihood ratio			
X^2_{LR}	1	10.56	0.0012
$X^2_{\mathrm{LR}}(d)$ (adjusted)	1	0.528	0.4675
Neyman			
$X^2_{\mathrm{N}}(= X^2_{bin})$	1	11.90	0.0006
$X^2_{\mathrm{N}}(\hat{d})$ (adjusted)	1	0.583	0.4451
Wald			
X^2_{des}	1	0.583	0.4451

The two main approaches to accounting for the clustering effect in the test statistics demonstrated in this example, namely the Rao–Scott adjusting methodology used for the Pearson, likelihood ratio and Neyman test statistics, and the design-based Wald statistic, are readily applicable for more general one-way tables, and for two-way tables where the number of rows and columns is greater than two. We next consider a more general case for a simple goodness-of-fit test and give details of alternative test statistics. Then, the tests for a homogeneity hypothesis and a hypothesis of independence are considered for a two-way table. In the testing procedures, we will concentrate on the design-based Wald statistic and on various Rao–Scott adjustments to the Pearson and Neyman test statistics.

7.2 SIMPLE GOODNESS-OF-FIT TEST

A valid testing procedure for a goodness-of-fit hypothesis in the case of more than two cells is more complicated than the simple case of two cells. This is true both for the design-based Wald statistic and for the Rao–Scott adjustments to the Pearson and Neyman test statistics. We next discuss these testing procedures in some detail.

The design-based Wald statistic provides a natural testing procedure for a simple goodness-of-fit hypothesis since it is generally asymptotically correct in complex surveys. The Wald statistic can be expected to work adequately in practice if a large number of sample clusters are present, which is the case, for example, in the OHC Survey. But the test statistic can suffer from problems of instability if the number of sample clusters is too small. Then, observed values of the statistic

can be obtained, which are too large. Fortunately, effects of instability on the test statistic can be reduced by an F-correction. Another generally asymptotically valid testing procedure is based on a second-order Rao–Scott adjustment to the Pearson and Neyman test statistics. It is important to be able to obtain a full design-based covariance-matrix estimate for both these testing procedures, and this presupposes access to the element-level data set.

There are situations we come across in practice where there is no access to the element-level data set. For example, in secondary analyses on published tables, an estimate of the full design-based covariance matrix is rarely provided. Therefore, a Wald statistic, or a second-order Rao–Scott adjustment, cannot be used. But certain approximative first-order adjustments are possible if appropriate design-effect estimates are reported. Although adjustments based on these design-effect estimates are asymptotically valid only under special conditions, in many situations they can be used as a better alternative to the uncorrected Pearson or Neyman test statistics.

A goodness-of-fit hypothesis for $u \geq 2$ cells can be written as $H_0 : p_j = p_{0j}$, $j = 1, \ldots, u$, where $p_j = N_j/N$ are the unknown cell proportions and p_{0j} are the hypothesized cell proportions. The null hypothesis can be conveniently written, using the corresponding vectors, as $H_0 : \mathbf{p} = \mathbf{p}_0$, where $\mathbf{p} = (p_1, \ldots, p_{u-1})'$ is the vector of the unknown cell proportions and $\mathbf{p}_0 = (p_{01}, \ldots, p_{0,u-1})'$ is the vector of the hypothesized proportions. The consistently estimated vector of cell proportions, based on a sample of n elements, is denoted by $\hat{\mathbf{p}} = (\hat{p}_1, \ldots, \hat{p}_{u-1})'$, where $\hat{p}_j = \hat{n}_j/n$. The $\hat{n}_j$ are scaled weighted cell frequencies accounting for unequal element inclusion probabilities and adjustment for nonresponse, such that $\sum_{j=1}^{u} \hat{n}_j = n$ (see Chapter 5). The $\hat{p}_j$ are ratio estimators if n is not fixed in advance, typically when working with a population subgroup as is assumed here. Note that only $u - 1$ elements are included in each of the vectors $\mathbf{p}$, $\mathbf{p}_0$ and $\hat{\mathbf{p}}$ because the proportions are constrained to sum up to one, thus, for example, $\hat{p}_u = 1 - \sum_{j=1}^{u-1} \hat{p}_j$.

Design-based Wald Statistic

A design-based Wald statistic X^2_{des} of the simple goodness-of-fit hypothesis was previously introduced for the case of two cells with clustered sampling designs as an alternative to the adjusted Pearson statistic. In the case of more than two cells, the design-based Wald statistic of goodness of fit is slightly more complicated:

$$X^2_{des} = (\hat{\mathbf{p}} - \mathbf{p}_0)' \hat{\mathbf{V}}^{-1}_{des} (\hat{\mathbf{p}} - \mathbf{p}_0), \tag{7.9}$$

where $\hat{\mathbf{V}}_{des}$ denotes a consistent covariance-matrix estimator of the true covariance matrix $\mathbf{V}/n$ of the proportion estimator vector $\hat{\mathbf{p}}$. An estimate $\hat{\mathbf{V}}_{des}$ can be obtained by the linearization method; the sample reuse methods, such as the jackknife, can also be used. The statistic X^2_{des} is asymptotically chi-squared with $u - 1$ degrees

of freedom if the null hypothesis is true, thus providing a valid testing procedure for complex surveys. In practice, X^2_{des} can be expected to work reasonably if the number of sample clusters is large and the number of cells is relatively small, because then we can expect a stable estimate $\hat{\mathbf{V}}_{des}$. Note that the statistic (7.8) is a special case of the statistic (7.9).

Unstable Situations

If there is a small number m of sample clusters available, an instability problem in the estimate $\hat{\mathbf{V}}_{des}$ may be encountered because there may only be a few degrees of freedom $f = m - H$ for the estimate. Consequences of instability of an estimate $\hat{\mathbf{V}}_{des}$ to the Wald statistic X^2_{des} can be severe, making the statistic overly liberal. One of the most widely used techniques to overcome instability is to make a degrees-of-freedom correction to the Wald statistic, giving rise to a new statistic that is assumed F-distributed. There are two alternative F-corrected Wald statistics. The first one is given by

$$F_{1.des} = \frac{f - u + 2}{f(u-1)} X^2_{des}, \tag{7.10}$$

which is treated as an F-distributed random variate with $u - 1$ and $f - u + 2$ degrees of freedom, and the second is

$$F_{2.des} = X^2_{des}/(u-1), \tag{7.11}$$

which is in turn referred to the F-distribution with $u - 1$ and f degrees of freedom. Note that if $u = 2$, both corrections reproduce the original statistic. The effect of an F-correction to X^2_{des} can be easily seen in the case of just two cells. If f is small, then a p-value for X^2_{des} from the F-distribution with one and f degrees of freedom is larger than that from the chi-squared distribution with one degree of freedom, but when f increases the difference vanishes. Thus, the corrections are ineffective if f is large. But for a small f, they can effectively correct the liberality in the uncorrected Wald statistic; this is true also where $u > 2$.

Thomas and Rao (1987) provide comparative results of the performances of various test statistics of a simple goodness of fit under instability, based on simulation. Although they noticed that the F-corrected Wald statistic $F_{1.des}$ did not indicate overall best performance relative to its competitors, it behaved relatively well in standard situations where instability was not very severe. The F-corrected Wald statistics are widely applied in practice and are also implemented in software products for survey analysis.

Pearson Test Statistic and Rao–Scott Adjustments

As noted in the introductory example, test statistics based on an assumption of simple random sampling require adjustments for the clustering effects to meet the

desired asymptotic properties. Let us first consider the Pearson test statistic X_P^2. The statistic can be compactly written in a matrix form

$$X_P^2 = n\sum_{j=1}^{u}(\hat{p}_j - p_{0j})^2/p_{0j} = n(\hat{\mathbf{p}} - \mathbf{p}_0)'\mathbf{P}_0^{-1}(\hat{\mathbf{p}} - \mathbf{p}_0), \tag{7.12}$$

where $\mathbf{P}_0 = \text{diag}(\mathbf{p}_0) - \mathbf{p}_0\mathbf{p}_0'$ and $\mathbf{P}_0/n$ is the $(u-1) \times (u-1)$ multinomial covariance matrix of $\hat{\mathbf{p}}$ under the null hypothesis, and the operator $\text{diag}(\mathbf{p}_0)$ generates a diagonal matrix with diagonal elements p_{0j}. The covariance matrix $\mathbf{P}_0/n$ is a generalization of the case of $u = 2$ cells to the case of more than two cells. Note that the matrix formula of X_P^2 mimics that of the Wald statistic (7.9), the only difference being that $\mathbf{P}_0/n$ is used instead of $\hat{\mathbf{V}}_{des}$. In case of two cells, X_P^2 reduces to the simple formula $X_P^2 = (\hat{p}_1 - p_{01})^2/(p_{01}(1-p_{01})/n)$ previously considered, where the denominator corresponds to a binomial variance derived under the null hypothesis.

To examine the asymptotic distribution of the Pearson test statistic X_P^2, we generalize the previous results from the case of two cells to the case of $u > 2$ cells. In this case, X_P^2 is asymptotically distributed as a weighted sum $\delta_1 W_1 + \delta_2 W_2 + \cdots + \delta_{u-1} W_{u-1}$ of $u-1$ independent chi-squared random variables W_j each with one degree of freedom. The weights δ_j are eigenvalues of a *generalized design-effects matrix* $\mathbf{D} = \mathbf{P}_0^{-1}\mathbf{V}$, where $\mathbf{V}/n$ is the true covariance matrix of the proportion estimator vector $\hat{\mathbf{p}}$ based on the actual sampling design. These eigenvalues are also called *generalized design effects*. Note that, in general, they do not coincide with the design-effects d_j.

If the actual sampling design is simple random sampling, then the generalized design-effects δ_j are all equal to one because the true and assumed covariance matrices $\mathbf{V}/n$ and $\mathbf{P}_0/n$ coincide and, therefore, the generalized design-effects matrix is an identity matrix. The weighted sum $\sum_{j=1}^{u-1}\delta_j W_j$ then reduces to $\sum_{j=1}^{u-1} W_j$, i.e. a sum of $u-1$ independent chi-squared random variates χ_1^2 whose distribution obviously is χ^2 with $u-1$ degrees of freedom. Thus, under simple random sampling, the Pearson statistic X_P^2 is asymptotically chi-squared with $u-1$ degrees of freedom.

If the actual sampling design is more complex by involving clustering, then the true $\mathbf{V}/n$ and the assumed $\mathbf{P}_0/n$ do not necessarily coincide, and in this case, the generalized design-effects δ_j are not equal to one. The δ_j tend to be greater than one on average because of the clustering effect and, thus, the asymptotic distribution of the random variate $\sum_{j=1}^{u-1}\delta_j W_j$ is not assumed to be a chi-squared distribution with $u-1$ degrees of freedom. Therefore, the Pearson test statistic X_P^2 requires corrections similar to those used in the case of two cells. However, there are now more possibilities for an adjusted Pearson statistic, namely the so-called *first-order* and *second-order* Rao–Scott adjustments developed by Rao and Scott (1981). The aim of the first-order adjustment is to correct the asymptotic expectation of the Pearson statistic, and the second-order adjustment also involves

an asymptotically correct variance. Technically, both adjustments are based on eigenvalues of an estimated generalized design-effects matrix $\hat{\mathbf{D}}$.

We first consider a simple *mean deff adjustment* to X^2_{P}, due to Fellegi (1980) and Holt *et al.* (1980), and the first-order Rao–Scott adjustment. These adjustments are aimed at situations where the full design-based estimate $\hat{\mathbf{V}}_{des}$ is not available. If this estimate is provided, a more exact second-order adjustment is preferable.

The mean deff adjustment is based on the estimated design-effects $\hat{d}_j$ of the proportions $\hat{p}_j$. An adjusted statistic to (7.12) is calculated by dividing the observed value of the Pearson statistic by the average design effect:

$$X^2_{\mathrm{P}}(\hat{d}_\cdot) = X^2_{\mathrm{P}}/\hat{d}_\cdot, \tag{7.13}$$

where $\hat{d}_\cdot = \sum_{j=1}^{u} \hat{d}_j/u$ is an estimator of the mean $\overline{d}$ of the unknown design-effects d_j. We estimate the design effects by $\hat{d}_j = \hat{v}_{des}(\hat{p}_j)/(\hat{p}_j(1-\hat{p}_j)/n)$, where $\hat{v}_{des}(\hat{p}_j)$ are design-based variance estimators of the proportion estimators $\hat{p}_j$. This adjustment thus requires that the design-effect estimates of the u cell proportion estimates are available. Positive intra-cluster correlation gives a mean $\hat{d}_\cdot$ greater than one, and so the mean deff adjustment tends to remove the liberality in X^2_{P}. The mean deff adjustment can also be executed by calculating the effective sample size $\overline{n} = n/\hat{d}_\cdot$ and then inserting $\overline{n}$ into equation (7.12) of X^2_{P} in place of n.

The mean deff adjustment is approximate so that it does not involve exact correction to the asymptotic expectation of X^2_{P}, because the mean of the design effects is generally not equal to the mean of the generalized design effects. Under the null hypothesis, the asymptotic expectation of X^2_{P} is $E(X^2_{\mathrm{P}}) = \sum_{j=1}^{u-1} \delta_j$, so $E(X^2_{\mathrm{P}}/\overline{\delta}) = E(\chi^2_{u-1}) = u-1$, where the mean of the eigenvalues is $\overline{\delta} = \sum_{j=1}^{u-1} \delta_j/(u-1)$. This argument leads to a *first-order Rao–Scott adjustment* to X^2_{P} given by

$$X^2_{\mathrm{P}}(\hat{\delta}_\cdot) = X^2_{\mathrm{P}}/\hat{\delta}_\cdot, \tag{7.14}$$

where $\hat{\delta}_\cdot$ is an estimate of the mean $\overline{\delta}$ of the unknown eigenvalues. This mean can be estimated using the design-effect estimates by the equation

$$(u-1)\hat{\delta}_\cdot = \sum_{j=1}^{u} \frac{\hat{p}_j}{p_{0j}}(1-\hat{p}_j)\hat{d}_j$$

without estimating the eigenvalues themselves. Alternatively, $\hat{\delta}_\cdot$ can be obtained from the generalized design-effects matrix estimate $\hat{\mathbf{D}} = n\mathbf{P}_0^{-1}\hat{\mathbf{V}}_{des}$ by the equation $\hat{\delta}_\cdot = \mathrm{tr}(\hat{\mathbf{D}})/(u-1)$, i.e. by dividing the trace of $\hat{\mathbf{D}}$ by the degrees of freedom. The adjusted statistic $X^2_{\mathrm{P}}(\hat{\delta}_\cdot)$ is asymptotically chi-squared with $u-1$ degrees of freedom only if the eigenvalues δ_j are all equal, but the statistic is noted to work reasonably in practice if the variation in the estimated eigenvalues $\hat{\delta}_j$ is small. Because only design-effect estimates of $\hat{p}_j$ are needed, the statistic is also suitable

for secondary analyses from published tables if the design-effect estimates are supplied. The first-order Rao–Scott adjustment $X^2_{\mathrm{P}}(\hat{\delta}_\bullet)$ is more exact than the corresponding mean deff adjustment $X^2_{\mathrm{P}}(\hat{d}_\bullet)$, which can be taken as a conservative alternative to $X^2_{\mathrm{P}}(\hat{\delta}_\bullet)$.

The first-order Rao–Scott adjustment (7.14) is aimed at successfully correcting the Pearson test statistic X^2_{P} so that the asymptotic expectation would be equal to the degrees of freedom. If the variation in the estimated eigenvalues $\hat{\delta}_j$ is noted to be large, then a correction to the variance of X^2_{P} is also required. This is achieved by a *second-order Rao–Scott adjustment* based on the Satterthwaite (1946) method. The second-order adjusted Pearson statistic is given by

$$X^2_{\mathrm{P}}(\hat{\delta}_\bullet, \hat{a}^2) = X^2_{\mathrm{P}}(\hat{\delta}_\bullet)/(1 + \hat{a}^2), \tag{7.15}$$

where an estimator of the squared coefficient of variation a^2 of the unknown eigenvalues δ_j is

$$\hat{a}^2 = \sum_{j=1}^{u-1} \hat{\delta}_j^2/((u-1)\hat{\delta}_\bullet^2) - 1.$$

An estimator of the sum of the squared eigenvalues is given by

$$\sum_{j=1}^{u-1} \hat{\delta}_j^2 = \mathrm{tr}(\hat{\mathbf{D}}^2) = n^2 \sum_{j=1}^{u} \sum_{k=1}^{u} \hat{v}^2_{des}(\hat{p}_j, \hat{p}_k)/p_{0j}p_{0k},$$

where $\hat{v}_{des}(\hat{p}_j, \hat{p}_k)$ are variance and covariance estimators of $\hat{p}_j$ and $\hat{p}_k$. The degrees of freedom must also be adjusted for this statistic; $\mathbf{X}^2_{\mathrm{P}}(\hat{\delta}_\bullet, \hat{a}^2)$ is asymptotically chi-squared with Satterthwaite adjusted degrees of freedom $\mathrm{df}_S = (u-1)/(1+\hat{a}^2)$. Note that the full covariance-matrix estimate $\hat{\mathbf{V}}_{des}$ is required in the second-order adjustment, whereas in the first-order adjustment only the variance estimates $\hat{v}_{des}$ were needed.

In unstable situations, an F-correction to the first-order Rao–Scott adjustment (7.14) may be beneficial. It is given by

$$FX^2_{\mathrm{P}}(\hat{\delta}_\bullet) = X^2_{\mathrm{P}}/((u-1)\hat{\delta}_\bullet). \tag{7.16}$$

The statistic is referred to the F-distribution with $u-1$ and f degrees of freedom. Thomas and Rao (1987) noted this statistic as being better than the uncorrected first-order adjustment in unstable situations.

Neyman (Multinomial Wald) Statistic

The Neyman test statistic X^2_{N} was previously used as an alternative to the Pearson statistic. The Neyman statistic corresponds to a Wald statistic derived using an

assumption of a multinomial distribution on $\hat{\mathbf{p}}$. The Neyman statistic is

$$X^2_{\mathrm{N}} = n\sum_{j=1}^{u}(\hat{p}_j - p_{0j})^2/\hat{p}_j = n(\hat{\mathbf{p}} - \mathbf{p}_0)'\hat{\mathbf{P}}^{-1}(\hat{\mathbf{p}} - \mathbf{p}_0), \tag{7.17}$$

where $\hat{\mathbf{P}} = \mathrm{diag}(\hat{\mathbf{p}}) - \hat{\mathbf{p}}\hat{\mathbf{p}}'$ and $\hat{\mathbf{P}}/n$ is the estimated (empirical) multinomial covariance matrix. Note that this equation mimics equations (7.9) and (7.12) of the design-based Wald statistic and the Pearson statistic; the only difference is that $\hat{\mathbf{P}}/n$ is used instead of $\hat{\mathbf{V}}_{des}$ or $\mathbf{P}_0/n$. Under simple random sampling, X^2_{N} is asymptotically chi-squared with $u-1$ degrees of freedom, but for more complex designs the statistic requires adjustments similar to those used for the Pearson statistic. We thus have a mean deff adjustment for X^2_{N} given by $X^2_{\mathrm{N}}(\hat{d}_\bullet) = X^2_{\mathrm{N}}/\hat{d}_\bullet$, a first-order Rao–Scott adjustment $X^2_{\mathrm{N}}(\hat{\delta}_\bullet) = X^2_{\mathrm{N}}/\hat{\delta}_\bullet$, a second-order Rao–Scott adjustment $X^2_{\mathrm{N}}(\hat{\delta}_\bullet, \hat{a}^2) = X^2_{\mathrm{N}}(\hat{\delta}_\bullet)/(1+\hat{a}^2)$ and an F-corrected first-order Rao–Scott adjustment $FX^2_{\mathrm{N}}(\hat{\delta}_\bullet) = X^2_{\mathrm{N}}(\hat{\delta}_\bullet)/(u-1)$.

Test Statistic and Distributional Properties

Our discussion so far indicates that the asymptotic properties of a test statistic depend on the sampling design assumptions specific to the statistic and on the actual sampling design. More specifically, let $\mathbf{D} = \mathbf{P}^{-1}\mathbf{V}$ be a design-effects matrix, where $\mathbf{P}/n$ is the covariance matrix corresponding to the assumed sampling design and $\mathbf{V}/n$ is the true covariance matrix based on the actual design. Asymptotic distribution of a test statistic depends on the eigenvalues of such a design-effects matrix. If all the eigenvalues are equal to one, a test statistic of goodness of fit is asymptotically chi-squared with $u-1$ degrees of freedom.

For the Pearson test statistic, the assumed covariance matrix $\mathbf{P}/n$ was a multinomial $\mathbf{P}_0/n$. If the actual design was also simple random sampling, then the true $\mathbf{V}/n$ and assumed $\mathbf{P}/n$ would coincide and all the eigenvalues would be equal to one. But if the actual design is more complex, the covariance matrices do not coincide and the eigenvalues differ from the nominal value of one. Thus, an adjustment to X^2_{P} is required.

For the design-based Wald statistic, the situation is different because the assumed and actual sampling designs coincide. Thus, the covariance matrices $\mathbf{P}/n$ and $\mathbf{V}/n$ in $\mathbf{D}$ are equal by definition. So, if the actual design is simple random sampling, we put $\mathbf{P}/n = \mathbf{V}/n = \mathbf{P}_0/n$, and if the actual design is more complex, involving clustering and stratification, we put $\mathbf{P}/n = \mathbf{V}/n$. In both cases, the eigenvalues of the corresponding design-effects matrix are equal to one and no adjustment to X^2_{des} is required.

Residual Analysis

If a goodness-of-fit test does not support the null hypothesis, a residual analysis can be performed to study the deviations from H_0. For a simple random sample,

the standardized residuals are of the form

$$\hat{e}_j = (\hat{p}_j - p_{0j})/\text{s.e}_{srs}(\hat{p}_j), \quad j = 1, \ldots, u, \tag{7.18}$$

where $\text{s.e}_{srs}(\hat{p}_j)$ is the square root of the corresponding diagonal element of the multinomial covariance-matrix estimate $\hat{\mathbf{P}}/n$. A large absolute value of $\hat{e}_j$ indicates deviation from H_0. But in complex surveys, these standardized residuals can be too large because the multinomial standard errors tend to underestimate the true standard errors. We therefore derive the design-based standardized residuals by using the corresponding design-based standard errors $\text{s.e}_{des}(\hat{p}_j)$. Hence, we have

$$\hat{e}_j = (\hat{p}_j - p_{0j})/\text{s.e}_{des}(\hat{p}_j), \quad j = 1, \ldots, u. \tag{7.19}$$

Clearly, if design-effect estimates are noticeably larger than one, smaller standardized residuals are obtained by (7.19) relative to the multinomial counterparts. The design-based standardized residuals can be taken as approximate standard normal variates under the null hypothesis, so they can be referred to critical values from the $N(0, 1)$ distribution.

Example 7.1

Goodness-of-fit test of the age distribution for the MFH Survey. We consider a goodness-of-fit test for the age distribution of the MFH Survey subgroup of males aged 30–64 years, relative to the respective population age distribution. We have chosen the MFH design to demonstrate also the effects of a small number of sample clusters ($m = 48$) on test results. Sample and population age distributions with the estimated cell design effects of the proportion estimates are displayed in Table 7.1. The standardized design-based residuals are also included in the table.

Because the cell proportions are constrained to sum up to one, there are $u - 1 = 2$ degrees of freedom for the tests. The null hypothesis is stated as

Table 7.1 Estimated and hypothesized age distributions, design-effect estimates of the age proportions, and standardized residuals in the MFH Survey subgroup of 30–64-year-old males.

Age	n_j	Estimated $\hat{p}_j$	Hypothesized p_{0j}	Deff $\hat{d}_j$	Residuals $\hat{e}_j$
30–44	1329	0.492	0.521	1.51	−2.45
45–54	774	0.287	0.277	1.70	0.88
55–64	596	0.221	0.202	0.43	3.64
Total sample	2699	1.000	1.000		

$H_0 : p_j = p_{0j}$ with $j = 1, 2, 3$. The values of the unadjusted Pearson and Neyman test statistics, and the values of the mean deff adjustment and the first-order Rao–Scott adjustment to the Pearson statistic, can be calculated from Table 7.1 using the sample and population proportions $\hat{p}_j$ and p_{0j} and the design-effect estimates $\hat{d}_j$. But the second-order Rao–Scott adjustment and the Wald statistic require a full estimate $\hat{\mathbf{V}}_{des}$ of the proportion estimates. This estimate was obtained using the linearization method. For complete information, we supply the full 3×3 covariance-matrix estimate

$$\hat{\mathbf{V}}_{des} = 10^{-5} \times \begin{bmatrix} 13.9481 & -12.0731 & -1.8750 \\ -12.0731 & 12.9158 & -0.8427 \\ -1.8750 & -0.8427 & 2.7177 \end{bmatrix}.$$

For comparison, we display also the multinomial counterparts $\mathbf{P_0}/n = (\text{diag}(\mathbf{p}_0) - \mathbf{p}_0\mathbf{p}_0')/2699$ and $\hat{\mathbf{P}}/n = (\text{diag}(\hat{\mathbf{p}}) - \hat{\mathbf{p}}\hat{\mathbf{p}}')/2699$. These are

$$\mathbf{P}_0/n = 10^{-5} \times \begin{bmatrix} 9.2464 & -5.3471 & -3.8993 \\ -5.3471 & 7.4202 & -2.0731 \\ -3.8993 & -2.0731 & 5.9724 \end{bmatrix},$$

and

$$\hat{\mathbf{P}}/n = 10^{-5} \times \begin{bmatrix} 9.2603 & -5.2317 & -4.0286 \\ -5.2317 & 7.5817 & -2.3500 \\ -4.0286 & -2.3500 & 6.3786 \end{bmatrix}.$$

The covariance-matrix estimates $\mathbf{P}_0/n$ and $\hat{\mathbf{P}}/n$ can be used in the calculation of the design-effects matrix estimate $\hat{\mathbf{D}}$ and the Pearson and Neyman test statistics (7.12) and (7.17). Note that in the calculation of X^2_{des} in (7.9), and X^2_{P} and X^2_{N}, we need not use the full matrices but take the 2×2 submatrices from the estimates $\hat{\mathbf{V}}_{des}$, $\mathbf{P}_0/n$ and $\hat{\mathbf{P}}/n$ corresponding to the two elements of the vectors $\hat{\mathbf{p}}$ and $\mathbf{p}_0$. Of course, the Pearson and Neyman statistics can be calculated as well by using the standard formulae, which were also given in equations (7.12) and (7.17).

For the adjusted Pearson and Neyman test statistics, we obtain

$$\hat{d}_\bullet = \sum_{j=1}^{3} \hat{d}_j/3 = 1.21$$

$$\hat{\delta}_\bullet = \sum_{j=1}^{3} \hat{p}_j p_{0j}^{-1}(1 - \hat{p}_j)\hat{d}_j/2 = 1.17$$

$$1 + \hat{a}^2 = 2699^2 \sum_{j=1}^{3}\sum_{k=1}^{3} (\hat{v}^2_{des}(\hat{p}_j, \hat{p}_k)/p_{0j}p_{0k})/(2 \times 1.17^2) = 1.37$$

$$\text{df}_S = (u - 1)/(1 + \hat{a}^2) = 1.46.$$

Using these estimates, we obtain the following:

Neyman (multinomial Wald) statistic:

$$X^2_{\mathrm{N}} = 9.96 \text{ with 2 df (degrees of freedom) and a } p\text{-value } 0.007.$$

Pearson statistic:

$$X^2_{\mathrm{P}} = 10.15 \text{ with 2 df and a } p\text{-value } 0.006.$$

Mean deff adjustment to the Pearson statistic:

$$X^2_{\mathrm{P}}(\hat{d}_{\bullet}) = 10.15/1.21 = 8.38 \text{ with 2 df and a } p\text{-value } 0.015.$$

First-order Rao–Scott adjustment to the Pearson statistic:

$$X^2_{\mathrm{P}}(\hat{\delta}_{\bullet}) = 10.15/1.17 = 8.66 \text{ with 2 df and a } p\text{-value } 0.013.$$

F-corrected first-order Rao–Scott adjustment:

$$FX^2_{\mathrm{P}}(\hat{\delta}_{\bullet}) = 8.66/2 = 4.33 \text{ with 2 and 24 df and a } p\text{-value } 0.025.$$

Second-order Rao–Scott adjustment to the Pearson statistic:

$$X^2_{\mathrm{P}}(\hat{\delta}_{\bullet}, \hat{a}^2) = 8.66/1.37 = 6.30 \text{ with } 2/1.37 = 1.46 \text{ df and a } p\text{-value } 0.023.$$

Design-based Wald statistic:

$$X^2_{des} = 15.28 \text{ with 2 df and a } p\text{-value } 0.001.$$

F-corrected Wald statistics:

$$F_{1.des} = (24 - 3 + 2)/(24 \times 2) \times 15.28 = 7.32 \text{ with 2 and 23 df and a } p\text{-value } 0.003, \text{ and}$$

$$F_{2.des} = 15.28/2 = 7.64 \text{ with 2 and 24 df and a } p\text{-value } 0.003.$$

Of the test statistics introduced, the second-order Rao–Scott adjustment and the Wald statistic with an F-correction could be expected to provide the most adequate test results. The mean deff adjustment and the first-order Rao–Scott adjustment are aimed to be used only if the design-effect estimates in Table 7.1 are available but not the covariance-matrix estimate $\hat{\mathbf{V}}_{des}$.

The test results indicate that the uncorrected Pearson and Neyman statistics give liberal results relative to the adjusted Pearson tests, as expected. Of the adjusted tests, the second-order Rao–Scott adjustment and the F-corrected

first-order Rao–Scott adjustment are most conservative. The design-based Wald test, however, is unexpectedly liberal, and the F-corrections involve no apparent improvement in this case. The liberality may be due to the relatively few degrees of freedom ($f = 24$) for the estimate $\hat{\mathbf{V}}_{des}$, which might be unstable. Actually, the eigenvalues of the relevant 2×2 submatrix of $\hat{\mathbf{V}}_{des}$ are 0.0002552 and 0.0000135, and thus the condition number is 18.9, though this does not indicate serious instability.

Which one of the seven test statistics aimed at accounting for the clustering effects should be chosen in the MFH Survey where the degrees of freedom for $\hat{\mathbf{V}}_{des}$ are small? Assuming first that an estimate $\hat{\mathbf{V}}_{des}$ is provided, the second-order Rao–Scott adjustment would be chosen because of the apparent nondiagonality of $\hat{\mathbf{V}}_{des}$, and because the second-order correction is not expected to be seriously sensitive to instability problems. Although also asymptotically valid, the design-based Wald statistic, and its F-corrections, would be excluded in this case because of obvious liberality. It should be noticed that in other testing situations where the number of sample clusters is larger, the design-based Wald statistic will be a reasonable alternative. If an estimate $\hat{\mathbf{V}}_{des}$ is not available but the appropriate design-effect estimates are provided, the F-corrected first-order Rao–Scott adjustment would be chosen and this also seems to successfully reduce the effect of instability.

The test results do not support the conclusion that the sample and population age distributions were equal. A residual analysis for the design-based standardized residuals $\hat{e}_j$ indicates that the largest deviance is in the third age group, and the standardized residual exceeds the 1% critical value 2.33 from the $N(0, 1)$ distribution. The residuals are smaller than the multinomial counterparts, except in the last age group, which has a design-effect estimate noticeably smaller than one.

Rejection of H_0 suggests that it might be reasonable to weight the MFH Survey data set to better match the sample age distribution with the population age distribution. In Section 5.1, we demonstrated this by developing the appropriate poststratification weights. It was noted that this weighting caused some, small, differences in the weighted estimates, relative to the unweighted ones, in response variables that were apparently age-dependent.

7.3 PRELIMINARIES FOR TESTS FOR TWO-WAY TABLES

In a two-way table, a *test of homogeneity* is appropriate to study whether the class proportions of a categorical response variable are equal over a set of classes of a categorical predictor variable. A *test of independence* is stated when studying whether there is nonzero association between two categorical response variables. The two tests thus conceptually differ in the formulation of the hypotheses and in the interpretation of test results. Under a simple random sample, a multinomial-based test such as the Pearson test can be used with an identical formula of a test statistic for both hypotheses. For more complex designs involving clustering,

we also separate the tests technically, and derive different adjustments for the corresponding test statistics. We first introduce the preliminaries of the tests with a simple example from the MFH Survey.

Test of Independence

Let us first consider the test of independence in the simplest case of a two-way table. From the MFH Survey demonstration data set of size $n = 2699$ persons, we have the following frequency table with two categorical variables, PHYS (physical health hazards of work, 0: none, 1: some) and SYSBP (systolic blood pressure, $\leq$159 or >159):

	SYSBP		
PHYS	$\leq$159	>159	All
0	1857	362	2219
1	390	90	480
All	2247	452	2699

For an *independence hypothesis*, our question is whether the two variables are associated or not. This leads to the null hypothesis

$$H_0 : p_{jk} = p_{j+}p_{+k}, \quad j, k = 1, 2,$$

where p_{jk} are unknown population cell proportions and p_{j+} and p_{+k} are the corresponding row and column marginal proportions in an N element population with cell frequencies N_{jk}. We thus have

$$p_{jk} = N_{jk}/N \quad \text{and} \quad p_{11} + p_{12} + p_{21} + p_{22} = 1,$$
$$p_{j+} = p_{j1} + p_{j2} \quad \text{and} \quad p_{+k} = p_{1k} + p_{2k}.$$

Because of the constraints on the cell and marginal proportions, the null hypothesis reduces to $H_0 : p_{11} = p_{1+}p_{+1}$ with one degree of freedom for the test.

For the independence hypothesis, the table of observed cell and marginal proportions $\hat{p}_{jk} = \hat{n}_{jk}/n$, and $\hat{p}_{j+} = \hat{p}_{j1} + \hat{p}_{j2}$ and $\hat{p}_{+k} = \hat{p}_{1k} + \hat{p}_{2k}$, can now be derived using the observed cell frequencies $\hat{n}_{jk}$:

	SYSBP		
PHYS	$\leq$159	>159	All
0	0.6880	0.1342	0.8222
1	0.1445	0.0333	0.1778
All	0.8325	0.1675	1

Note that the cell proportions sum up to one over the table. A Pearson test statistic for the hypothesis of independence is

$$X_{\mathrm{P}}^2(I) = n \sum_{j=1}^{2} \sum_{k=1}^{2} \frac{(\hat{p}_{jk} - \hat{p}_{j+}\hat{p}_{+k})^2}{\hat{p}_{j+}\hat{p}_{+k}} = \frac{n(\hat{p}_{11} - \hat{p}_{1+}\hat{p}_{+1})^2}{\hat{p}_{1+}(1-\hat{p}_{1+})\hat{p}_{+1}(1-\hat{p}_{+1})},$$

which is a scaled measure of the squared differences of the observed proportions from their expected values under the null hypothesis of independence. For a standard inference on the null hypothesis, the Pearson statistic is referred to the chi-squared distribution with one degree of freedom. Calculated from the table above, the observed value of $X_{\mathrm{P}}^2(I)$ is 1.68 with a p-value of 0.195, clearly suggesting acceptance of the null hypothesis of independence.

Test of Homogeneity

For the independence hypothesis, both the classification variables SYSBP and PHYS were actually taken as response variables. It is also possible to look at the frequency table from another point of view. If we consider SYSBP as a response variable and PHYS as a predictor variable, for a *homogeneity hypothesis* our question is then whether the distributions of SYSBP in the two classes of PHYS are equal. This leads to a null hypothesis

$$H_0 : p_{1k} = p_{2k}$$

for both values of $k = 1, 2$. When compared to the independence hypothesis, we now have different population proportions for which it holds

$$p_{11} + p_{12} = 1 \quad \text{and} \quad p_{21} + p_{22} = 1.$$

Because of these constraints, the null hypothesis reduces to $H_0 : p_{11} = p_{21}$, and again there is one degree of freedom for the test.

For the homogeneity hypothesis, the table of observed cell proportions $\hat{p}_{1k} = \hat{n}_{1k}/\hat{n}_1$ and $\hat{p}_{2k} = \hat{n}_{2k}/\hat{n}_2$, where $\hat{n}_1 = \hat{n}_{11} + \hat{n}_{12}$ and $\hat{n}_2 = \hat{n}_{21} + \hat{n}_{22}$ are row marginal frequencies and observed marginal proportions are $\hat{p}_{j+} = 1$ and $\hat{p}_{+k} = (\hat{n}_{1k} + \hat{n}_{2k})/n$, is the following:

	SYSBP		
PHYS	≤159	>159	All
0	0.8369	0.1631	1
1	0.8125	0.1875	1
All	0.8325	0.1675	1

Note that both the row margins $\hat{p}_{1+}$ and $\hat{p}_{2+}$ are equal to one. A Pearson test statistic for the hypothesis of homogeneity is now given as

$$X^2_{\mathrm{P}}(H) = \sum_{j=1}^{2}\sum_{k=1}^{2} \frac{\hat{n}_j(\hat{p}_{jk} - \hat{p}_{+k})^2}{\hat{p}_{+k}} = \frac{(\hat{p}_{11} - \hat{p}_{21})^2}{\hat{p}_{+1}(1-\hat{p}_{+1})/\hat{n}_1 + \hat{p}_{+2}(1-\hat{p}_{+2})/\hat{n}_2},$$

which is again a measure of the squared differences of the observed proportions from their expected values, under the null hypothesis of homogeneity. For inference on the null hypothesis, this Pearson statistic is also referred to the chi-squared distribution with one degree of freedom. Although the formulae of $X^2_{\mathrm{P}}(H)$ and $X^2_{\mathrm{P}}(I)$ were written differently, the observed value, 1.68, for $X^2_{\mathrm{P}}(H)$ is the same as that in the test of independence, and the conclusion—accept the null hypothesis—also remains true.

Cell Design Effects

The Pearson tests of independence and homogeneity were executed assuming a simple random sample. But would the conclusions remain if we account for the clustering effect? This can be examined by calculating the design-effect estimates of the estimated cell and marginal proportions of the observed tables for the independence and homogeneity hypotheses. Table 7.2 would then be helpful. Cell design effects for the independence hypothesis are in the first DEFF column, and those for the homogeneity hypothesis are in the second DEFF column.

It is obvious that if the design-effect estimates are greater than one on average, then more conservative adjusted test statistics would be obtained, relative to the unadjusted ones, and, therefore, the conclusion of accepting the null hypotheses would remain. The mean of the cell design-effect estimates for the

Table 7.2 Cell and marginal percentages and design effects for the independence and homogeneity hypotheses in the MFH Survey.

		Test of independence		Test of homogeneity	
Physical health hazards	Systolic blood pressure	Cell percent	Deff of cell percent	Row percent	Deff of row percent
No	≤159	68.8	1.50	83.7	0.88
	>159	13.4	0.81	16.3	0.88
Yes	≤159	14.5	1.43	81.3	1.15
	>159	3.3	1.34	18.7	1.15

independence hypothesis is $\hat{d}_{\bullet}(I) = 1.27$, giving the mean deff adjusted Pearson statistic $X_{\mathrm{P}}^2(I, \hat{d}_{\bullet}) = 1.32$ with a p-value of 0.251. And the mean of the cell design-effect estimates for the homogeneity hypothesis is $\hat{d}_{\bullet}(H) = 1.01$, giving the mean deff adjusted Pearson statistic $X_{\mathrm{P}}^2(H, \hat{d}_{\bullet}) = 1.66$ with a p-value of 0.198. These design-based tests involve no new inferential conclusions, but, more importantly, they demonstrate that, because of different adjustments, the adjusted Pearson test statistics accounting for the clustering effect do not give numerically equal results, although the unadjusted ones do. Difference between the adjustments to $X_{\mathrm{P}}^2(I)$ and $X_{\mathrm{P}}^2(H)$ also holds for the Rao–Scott corrections, and the design-based Wald test statistics of independence and homogeneity hypotheses would not coincide either.

The test results also indicate that in the case of the MFH Survey, intra-cluster correlation has a greater effect on the test of independence than on the test of homogeneity. This might be so because we are working with cross-classes-type subgroups, and in part might be due to the few degrees of freedom available for the variance estimates. It should be noticed that the situation can also reverse: it has been noted in some surveys that inflation due to clustering is often less for tests of independence than for tests of homogeneity (Rao and Thomas 1988). This holds especially in cases in which the classes of the predictor variable are of segregated-type regions.

For the analysis of more general $r \times c$ tables from complex surveys, a design-based Wald statistic with an F-correction, and a second-order Rao–Scott adjustment to the standard Pearson and Neyman test statistics, can be constructed for tests of homogeneity and independence as in the case of the simple goodness-of-fit test. In secondary analyses from published tables, the mean deff and first-order Rao–Scott adjustments are possible if cell and marginal design-effect estimates are provided, but not the design-based covariance-matrix estimate of proportion estimators.

7.4 TEST OF HOMOGENEITY

In survey analysis literature, a test of homogeneity is usually used to study the homogeneity of the distribution of a response variable over a set of non-overlapping regions where independent samples are drawn using multi-stage sampling designs (e.g. Rao and Thomas 1988). It is thus assumed that the regions are segregated classes so that all elements in a sample cluster fall into the same region (class of the predictor variable). The classes of the response variable are typically cross-classes that cut across the regions. More generally, the test of homogeneity can be taken as the simplest example of a logit model with a binary or polytomous response variable and one categorical predictor variable whose type in practice is not restricted to a segregated class.

For a homogeneity hypothesis, assuming that columns of the table are formed by the classes of the response variable and rows constitute the regions, it is assumed

that each row-wise sum of cell proportions is equal to one. The population table is thus as follows:

	Response variable						
Region	1	2	...	k	...	c	All
1	p_{11}	p_{12}	$\cdots$	p_{1k}	$\cdots$	p_{1c}	1
2	p_{21}	p_{22}	$\cdots$	p_{2k}	$\cdots$	p_{2c}	1
$\vdots$	$\vdots$	$\vdots$	...	$\vdots$	...	$\vdots$	$\vdots$
j	p_{j1}	p_{j2}	$\cdots$	p_{jk}	$\cdots$	p_{jc}	1
$\vdots$	$\vdots$	$\vdots$	...	$\vdots$	...	$\vdots$	$\vdots$
r	p_{r1}	p_{r2}	$\cdots$	p_{rk}	$\cdots$	p_{rc}	1

For simplicity, we consider the case of only two regions and assume that the regions are of segregated classes type. A hypothesis of homogeneity of a c category response variable for $r = 2$ regions was given in Section 7.3 as $H_0 : p_{1k} = p_{2k}$, where $p_{1k} = N_{1k}/N_1$ and $p_{2k} = N_{2k}/N_2$ are unknown population proportions in the first and second regions respectively and $k = 1, \ldots, c$. The hypothesis can be written, using vectors, as $H_0 : \mathbf{p}_1 = \mathbf{p}_2$, where $\mathbf{p}_j = (p_{j1}, \ldots, p_{j,c-1})'$ denotes the population vector of row proportions p_{jk} in region j. There are thus $c - 1$ elements in each regional proportion vector, because the proportions must sum up to one independently for each region. Further, we denote by $\mathbf{p} = (p_{+1}, \ldots, p_{+,c-1})'$ the unknown common proportion vector under H_0, where $p_{+k} = N_{+k}/N$ and $N_{+k} = N_{1k} + N_{2k}$.

The estimated regional proportion vectors, based on independent samples from the regions, are denoted by $\hat{\mathbf{p}}_j = (\hat{p}_{j1}, \ldots, \hat{p}_{j,c-1})'$, where $\hat{p}_{jk} = \hat{n}_{jk}/\hat{n}_j$ is a consistent estimator of the corresponding population proportion p_{jk}, and $\hat{n}_{jk}$ and $\hat{n}_j$ are scaled weighted-up cell and marginal frequencies accounting for unequal element inclusion probabilities and adjustment for nonresponse, so that $\sum_{k=1}^{c} \hat{n}_{jk} = \hat{n}_j$. The $\hat{p}_{jk}$ are ratio estimators when we work with subgroups of the regional samples whose sizes are not fixed in advance, as we assume here as in the goodness-of-fit case. This also holds, for example, for the demonstration data sets from the MFH and OHC Surveys.

Design-based Wald Statistic

Let us denote by $\hat{\mathbf{V}}_{des}(\hat{\mathbf{p}}_1)$ the consistent covariance-matrix estimator of the proportion estimator vector $\hat{\mathbf{p}}_1$ in the first region, and have $\hat{\mathbf{V}}_{des}(\hat{\mathbf{p}}_2)$ correspondingly for $\hat{\mathbf{p}}_2$ in the second region. The covariance-matrix estimators can be calculated for each region in a similar manner as for the goodness-of-fit case. Using $\hat{\mathbf{V}}_{des}(\hat{\mathbf{p}}_1)$

and $\hat{\mathbf{V}}_{des}(\hat{\mathbf{p}}_2)$, a design-based Wald statistic X^2_{des} of a homogeneity hypothesis for two regions is given by

$$X^2_{des} = (\hat{\mathbf{p}}_1 - \hat{\mathbf{p}}_2)'(\hat{\mathbf{V}}_{des}(\hat{\mathbf{p}}_1) + \hat{\mathbf{V}}_{des}(\hat{\mathbf{p}}_2))^{-1}(\hat{\mathbf{p}}_1 - \hat{\mathbf{p}}_2), \tag{7.20}$$

because of segregated classes and $r = 2$. The Wald statistic is asymptotically chi-squared with $(2-1) \times (c-1) = (c-1)$ degrees of freedom. And also, if $c = 2$, then X^2_{des} reduces to $X^2_{des} = (\hat{p}_{11} - \hat{p}_{21})^2/(\hat{v}_{des}(\hat{p}_{11}) + \hat{v}_{des}(\hat{p}_{21}))$. X^2_{des} in (7.20) does not directly generalize to the case with more than two regions but is more complicated (see e.g. Rao and Thomas 1988).

The statistic X^2_{des} can be expected to work reasonably if a large number of sample clusters are available in each region. But if this is not the case, an instability problem can be encountered. F-corrected Wald statistics may then be used instead. By using $f = m - H$ as the overall degrees of freedom for the estimate $(\hat{\mathbf{V}}_{des}(\hat{\mathbf{p}}_1) + \hat{\mathbf{V}}_{des}(\hat{\mathbf{p}}_2))$, where m and H are the total number of sample clusters and strata in the two regions, the corrections are given by

$$F_{1.des} = \frac{f - (c-1) + 1}{f(c-1)} X^2_{des}, \tag{7.21}$$

which is referred to the F-distribution with $(c-1)$ and $(f - (c-1) + 1)$ degrees of freedom, and further,

$$F_{2.des} = X^2_{des}/(c-1), \tag{7.22}$$

which is referred to the F-distribution with $(c-1)$ and f degrees of freedom. These test statistics can be effective in reducing the effect of instability if f is not large relative to the number of classes c in the response variable.

Adjustments to Pearson and Neyman Test Statistics

A Pearson test statistic for the homogeneity hypothesis in the case of $r = 2$ regions is

$$X^2_{\mathrm{P}} = \sum_{j=1}^{2} \sum_{k=1}^{c} \frac{\hat{n}_j(\hat{p}_{jk} - \hat{p}_{+k})^2}{\hat{p}_{+k}} = (\hat{\mathbf{p}}_1 - \hat{\mathbf{p}}_2)'(\hat{\mathbf{P}}/\hat{n}_1 + \hat{\mathbf{P}}/\hat{n}_2)^{-1}(\hat{\mathbf{p}}_1 - \hat{\mathbf{p}}_2), \tag{7.23}$$

where $\hat{p}_{+k} = (\hat{n}_1\hat{p}_{1k} + \hat{n}_2\hat{p}_{2k})/(\hat{n}_1 + \hat{n}_2)$ are marginal proportion estimators over the rows of the table, i.e. estimators of the elements p_{+k} of the hypothesized common proportion vector $\mathbf{p}$ under H_0, and $\hat{\mathbf{P}} = \mathrm{diag}(\hat{\mathbf{p}}) - \hat{\mathbf{p}}\hat{\mathbf{p}}'$ such that $\hat{\mathbf{P}}/\hat{n}_1$ is the multinomial covariance-matrix estimator of the estimator vector $\hat{\mathbf{p}}$ for the first region and $\hat{\mathbf{P}}/\hat{n}_2$ correspondingly for the second region. Also, if $c = 2$, then X^2_{P} reduces to $\hat{n}_1\hat{n}_2(\hat{p}_{11} - \hat{p}_{21})^2/((\hat{n}_1 + \hat{n}_2)\hat{p}_{+1}(1 - \hat{p}_{+1}))$.

As an alternative, a Neyman test statistic can be used, which can be derived from the design-based Wald statistic (7.20) by assuming independent multinomial sampling in both regions:

$$X^2_{\mathrm{N}} = \sum_{j=1}^{2}\sum_{k=1}^{c} \frac{\hat{n}_j(\hat{p}_{jk} - \hat{p}_{+k})^2}{\hat{p}_{jk}} = (\hat{\mathbf{p}}_1 - \hat{\mathbf{p}}_2)'(\hat{\mathbf{P}}_1/\hat{n}_1 + \hat{\mathbf{P}}_2/\hat{n}_2)^{-1}(\hat{\mathbf{p}}_1 - \hat{\mathbf{p}}_2), \quad (7.24)$$

where $\hat{\mathbf{P}}_1 = \mathrm{diag}(\hat{\mathbf{p}}_1) - \hat{\mathbf{p}}_1\hat{\mathbf{p}}_1'$ and $\hat{\mathbf{P}}_1/\hat{n}_1$ is the multinomial covariance-matrix estimator for the first region and $\hat{\mathbf{P}}_2/\hat{n}_2$ correspondingly for the second region. Also, if $c = 2$, then X^2_{N} reduces to $(\hat{p}_{11} - \hat{p}_{21})^2/(\hat{p}_{11}(1 - \hat{p}_{11})/\hat{n}_1 + \hat{p}_{21}(1 - \hat{p}_{21})/\hat{n}_2)$. Note that the matrix formulae of X^2_{P} and X^2_{N} resemble that of the design-based Wald statistic, the only difference being which covariance-matrix estimator is used.

The Pearson and Neyman test statistics are valid for a simple random sample, i.e. they are chi-squared with $(c - 1)$ degrees of freedom for two regions. But under more complex designs, the statistics require adjustments that account for clustering effects. The adjustments are basically similar to those for the goodness-of-fit test, but, technically, they are obtained by different formulae.

For a mean deff adjustment and for a first-order Rao–Scott adjustment to X^2_{P} and X^2_{N}, the cell design-effect estimates in both regions are needed, and for a second-order Rao–Scott adjustment, a generalized design-effects matrix estimate is required. The design-effect estimators in region j are of the form

$$\hat{d}_{jk} = \hat{d}(\hat{p}_{jk}) = \hat{n}_j\hat{v}_{jk}/(\hat{p}_{+k}(1 - \hat{p}_{+k})), \quad j = 1, 2 \quad \text{and} \quad k = 1, \ldots, c,$$

where $\hat{v}_{1k}$ is the kth diagonal element of the covariance-matrix estimate $\hat{\mathbf{V}}_{des}(\hat{\mathbf{p}}_1)$ in the first region and $\hat{v}_{2k}$ is the corresponding element of $\hat{\mathbf{V}}_{des}(\hat{\mathbf{p}}_2)$. The generalized design-effects matrix estimate is

$$\hat{\mathbf{D}} = \frac{\hat{n}_1\hat{n}_2}{\hat{n}_1 + \hat{n}_2}\hat{\mathbf{P}}^{-1}(\hat{\mathbf{V}}_{des}(\hat{\mathbf{p}}_1) + \hat{\mathbf{V}}_{des}(\hat{\mathbf{p}}_2)). \quad (7.25)$$

Mean deff adjustments to the Pearson and Neyman test statistics are

$$X^2_{\mathrm{P}}(\hat{d}_\bullet) = X^2_{\mathrm{P}}/\hat{d}_\bullet \quad \text{and} \quad X^2_{\mathrm{N}}(\hat{d}_\bullet) = X^2_{\mathrm{N}}/\hat{d}_\bullet, \quad (7.26)$$

where

$$\hat{d}_\bullet = \sum_{j=1}^{2}\sum_{k=1}^{c} \hat{d}_{jk}/(2c)$$

is the mean of the design-effect estimates. By using the eigenvalues $\hat{\delta}_k$ of $\hat{\mathbf{D}}$, the first-order Rao–Scott adjustments to Pearson and Neyman test statistics (7.23)

and (7.24) are given by

$$X^2_{\mathrm{P}}(\hat{\delta}_\bullet) = X^2_{\mathrm{P}}/\hat{\delta}_\bullet \quad \text{and} \quad X^2_{\mathrm{N}}(\hat{\delta}_\bullet) = X^2_{\mathrm{N}}/\hat{\delta}_\bullet, \tag{7.27}$$

where

$$\hat{\delta}_\bullet = \mathrm{tr}(\hat{\mathbf{D}})/(c-1) = \frac{1}{c-1}\sum_{j=1}^{2}\left(1 - \frac{\hat{n}_j}{\hat{n}_1+\hat{n}_2}\right)\sum_{k=1}^{c}\frac{\hat{p}_{jk}}{\hat{p}_{+k}}(1-\hat{p}_{jk})\hat{d}_{jk}$$

is an estimator of the mean $\overline{\delta}$ of the eigenvalues δ_k of the unknown generalized design-effects matrix $\mathbf{D}$. Note that an estimate $\hat{\delta}_\bullet$ can also be computed directly from $\hat{\mathbf{D}}$ by first calculating the sum of its diagonal elements, i.e. the trace. Both adjustments are referred to the chi-squared distribution with $(c-1)$ degrees of freedom. The adjustments are approximative in the sense that they can be expected to work reasonably if the design-effect estimates, or the eigenvalues, do not vary considerably.

A second-order adjustment to X^2_{P} and X^2_{N} is more appropriate if the variation in the eigenvalue estimates $\hat{\delta}_k$ is noticeable. For the Pearson statistic, this adjustment is given by

$$X^2_{\mathrm{P}}(\hat{\delta}_\bullet, \hat{a}^2) = X^2_{\mathrm{P}}(\hat{\delta}_\bullet)/(1+\hat{a}^2), \tag{7.28}$$

where $\hat{a}^2$ is the squared coefficient of variation of the eigenvalue estimates $\hat{\delta}_k$. It is obtained by the formula

$$\hat{a}^2 = \sum_{k=1}^{c-1}\hat{\delta}_k^2/((c-1)\hat{\delta}_\bullet^2) - 1,$$

where the sum of squared eigenvalues can be obtained as the trace of the generalized design-effects matrix estimate raised to the second power:

$$\sum_{k=1}^{c-1}\hat{\delta}_k^2 = \mathrm{tr}(\hat{\mathbf{D}}^2).$$

The second-order Rao–Scott corrected Pearson test statistic is asymptotically chi-squared with Satterthwaite adjusted degrees of freedom $\mathrm{df}_S = (c-1)/(1+\hat{a}^2)$. A similar adjustment can be carried out to the first-order corrected Neyman statistic $X^2_{\mathrm{N}}(\hat{\delta}_\bullet)$ in (7.27).

If the regional covariance-matrix estimates $\hat{\mathbf{V}}_{des}(\hat{\mathbf{p}}_1)$ and $\hat{\mathbf{V}}_{des}(\hat{\mathbf{p}}_2)$ are based on a relatively small number of sample clusters, they might be unstable and, therefore, F-corrected first-order test statistics can be used instead. The Pearson statistic in (7.27) with an F-correction for two regions is given by

$$FX^2_{\mathrm{P}}(\hat{\delta}_\bullet) = X^2_{\mathrm{P}}(\hat{\delta}_\bullet)/(c-1) \tag{7.29}$$

referred to the F-distribution with $(c-1)$ and f degrees of freedom. This correction is analogous for the Neyman statistic.

Residual Analysis

Under rejection of the null hypothesis H_0 of homogeneity, the standardized residuals can be computed to detect cell deviations from the hypothesized proportions. Using the cell design-effect estimates $\hat{d}_{jk}$, we calculate the design-based standardized residuals

$$\hat{e}_{jk} = (\hat{p}_{jk} - \hat{p}_{+k})/\text{s.e}_{des}(\hat{p}_{jk} - \hat{p}_{+k}), \quad j = 1, 2 \quad \text{and} \quad k = 1, \ldots, c, \tag{7.30}$$

where a standard-error estimator $\text{s.e}_{des}(\hat{p}_{jk} - \hat{p}_{+k})$ of a raw residual is obtained from the design-based variance estimator, given by

$$\hat{v}_{des}(\hat{p}_{1k} - \hat{p}_{+k}) = \frac{\hat{n}_2(\hat{n}_2\hat{d}_{1k} + \hat{n}_1\hat{d}_{2k})}{(\hat{n}_1 + \hat{n}_2)^2}\hat{p}_{+k}(1 - \hat{p}_{+k})/\hat{n}_1, \quad k = 1, \ldots, c,$$

for the first region, and

$$\hat{v}_{des}(\hat{p}_{2k} - \hat{p}_{+k}) = \frac{\hat{n}_1(\hat{n}_2\hat{d}_{1k} + \hat{n}_1\hat{d}_{2k})}{(\hat{n}_1 + \hat{n}_2)^2}\hat{p}_{+k}(1 - \hat{p}_{+k})/\hat{n}_2, \quad k = 1, \ldots, c,$$

for the second region. Note that under simple random sampling, when $\hat{d}_{1k} = \hat{d}_{2k} = 1$, these variance estimators reduce to

$$\hat{v}_{srs}(\hat{p}_{1k} - \hat{p}_{+k}) = \frac{\hat{n}_2}{\hat{n}_1 + \hat{n}_2}\hat{p}_{+k}(1 - \hat{p}_{+k})/\hat{n}_1, \quad k = 1, \ldots, c,$$

for the first region, and

$$\hat{v}_{srs}(\hat{p}_{2k} - \hat{p}_{+k}) = \frac{\hat{n}_1}{\hat{n}_1 + \hat{n}_2}\hat{p}_{+k}(1 - \hat{p}_{+k})/\hat{n}_2, \quad k = 1, \ldots, c,$$

for the second region. It can be inferred from these formulae that under positive intra-cluster correlation, smaller design-based standardized residuals are obtained than those obtained from the equations based on an assumption of simple random sampling. The design-based standardized residuals can be referred to the critical values from the standard normal $N(0,1)$ distribution.

Example 7.2

The test of homogeneity for two populations in the OHC Survey. We consider the test of homogeneity of class proportions of the variable PSYCH, which is the

Table 7.3 Class proportions of PSYCH (psychic symptoms) in public services and other industries in the OHC Survey (design-effect estimates in parentheses).

Type of industry	PSYCH 1	PSYCH 2	PSYCH 3	All	Sample size
Public services	0.2939	0.3345	0.3716	1.00	1184
	(2.02)	(1.24)	(1.74)		
Other industries	0.3526	0.3216	0.3258	1.00	6657
	(1.73)	(1.23)	(1.57)		
All industries	0.3437	0.3236	0.3327	1.00	7841

first principal component of nine psychic symptoms measuring overall psychic strain, categorized into three nearly equally sized classes. The two populations are formed by the type of industry of establishment, constructed so that public services constitute the first subgroup and all the other industries are put into the second subgroup (Table 7.3). Note that the grouping follows industrial stratification and thus is of segregated type, and independent samples can be assumed to be drawn from each population. Of the 250 sample clusters available, 49 are in the first subgroup and 201 in the second, and the element data sets in both subgroups are taken to be self-weighting.

In public services, a larger proportion of serious psychic symptoms (class 3) is obtained than that obtained in other industries. A homogeneity hypothesis $H_0 : p_{1k} = p_{2k}$, $k = 1, 2, 3$, of the class proportions over the two populations is stated to examine the variation. Cell design-effect estimates, with an average 1.59, indicate a moderate clustering effect, which should be accounted for in a testing procedure. For the calculation of valid test statistics, we first obtain the two full covariance-matrix estimates $\hat{\mathbf{V}}_{des}(\hat{\mathbf{p}}_1)$ and $\hat{\mathbf{V}}_{des}(\hat{\mathbf{p}}_2)$. These are

$$\hat{\mathbf{V}}_{des}(\hat{\mathbf{p}}_1) = 10^{-5} \times \begin{bmatrix} 35.3394 & -12.1408 & -23.1986 \\ -12.1408 & 23.3570 & -11.2161 \\ 23.1986 & -11.2161 & 34.4148 \end{bmatrix},$$

and

$$\hat{\mathbf{V}}_{des}(\hat{\mathbf{p}}_2) = 10^{-5} \times \begin{bmatrix} 5.9177 & -2.3978 & -3.5200 \\ -2.3978 & 4.0417 & -1.6439 \\ -3.5200 & -1.6439 & 5.1639 \end{bmatrix}.$$

Because $c - 1 = 2$, we use the first two classes of PSYCH and the first 2×2 submatrices from the estimates $\hat{\mathbf{V}}_{des}(\hat{\mathbf{p}}_1)$ and $\hat{\mathbf{V}}_{des}(\hat{\mathbf{p}}_2)$ in the calculation of Wald statistics and Rao–Scott adjustments. For a design-based Wald test (7.20) of homogeneity, we get $X^2_{des} = 8.62$ with 2 degrees of freedom and a p-value 0.0134, thus indicating

non-homogeneity of the proportions over the populations. F-corrections (7.21) and (7.22) to X^2_{des} give $F_{1.des} = 4.29$, which referred to the F-distribution with 2 and 244 degrees of freedom attains a p-value 0.0147, and $F_{2.des} = 4.31$, which referred to the F-distribution with 2 and 245 degrees of freedom attains a p-value 0.0144. These corrections do not have a large impact on X^2_{des} because of the relatively large total number of sample clusters, in which case $\hat{\mathbf{V}}_{des}(\hat{\mathbf{p}}_1)$ and $\hat{\mathbf{V}}_{des}(\hat{\mathbf{p}}_2)$ can be assumed to be stable.

As another valid testing procedure, we calculate the second-order Rao–Scott adjustments (7.28) to the Pearson and Neyman test statistics (7.23) and (7.24). The unadjusted statistics give observed values $X^2_{\mathrm{P}} = 16.93$, with a p-value 0.0002, and $X^2_{\mathrm{N}} = 17.77$, with a p-value 0.0001, both significant at the 0.001 level, so they are very liberal relative to X^2_{des} as expected. For the Rao–Scott adjustments, a generalized design-effects matrix estimate (7.25) is first obtained:

$$\hat{\mathbf{D}} = \begin{bmatrix} 2.01374 & -0.03663 \\ 0.35554 & 1.23977 \end{bmatrix}.$$

The mean of the diagonal elements of $\hat{\mathbf{D}}$ is $\hat{\delta}_{\bullet} = \mathrm{tr}(\hat{\mathbf{D}})/2 = 1.627$, and the sum of the squared eigenvalues is $\sum_{k=1}^{2} \hat{\delta}_k^2 = \mathrm{tr}(\hat{\mathbf{D}}^2) = 5.566$. The second-order correction factor is thus $(1 + \hat{a}^2) = 1.052$, and this with Satterthwaite adjusted degrees of freedom $\mathrm{df}_S = 1.902$ gives $X^2_{\mathrm{P}}(\hat{\delta}_{\bullet}, \hat{a}^2) = 9.89$, with a p-value 0.0063, and $X^2_{\mathrm{N}}(\hat{\delta}_{\bullet}, \hat{a}^2) = 10.38$, with a p-value 0.0049, both significant at the 0.01 level. The results are somewhat liberal relative to those from the Wald test. These test results indicate that the design-based Wald statistic works adequately in the OHC case, unlike the MFH case (see Example 7.1).

We finally calculate the first-order adjustments (7.26) and (7.27) to X^2_{P} and X^2_{N} under the assumption that the only information provided for a homogeneity test is that given in Table 7.3. The estimated mean design effect is $\hat{d}_{\bullet} = 1.59$, and the corresponding adjustments to X^2_{P} and X^2_{N} are $X^2_{\mathrm{P}}(\hat{d}_{\bullet}) = 10.66$, with a p-value 0.0048, and $X^2_{\mathrm{N}}(\hat{d}_{\bullet}) = 11.19$, with a p-value 0.0037, both significant at the 0.01 level. By using cell design-effect estimates and cell proportions, we obtain $\hat{\delta}_{\bullet} = 1.627$, giving the first-order Rao–Scott adjustments $X^2_{\mathrm{P}}(\hat{\delta}_{\bullet}) = 10.41$, with a p-value 0.0055, and $X^2_{\mathrm{N}}(\hat{\delta}_{\bullet}) = 10.92$, with a p-value 0.0043, which are also significant at the 0.01 level. The F-corrections (7.29) to X^2_{P} and X^2_{N} give $FX^2_{\mathrm{P}} = 5.20$, with a p-value 0.0061, and $FX^2_{\mathrm{N}} = 5.46$, with a p-value 0.0048, indicating no obvious change in the results from the first-order corrected counterparts, again demonstrating stability of the testing situation.

Because all the tests suggest rejection of H_0 at least at the 0.05 level, we calculate the design-based standardized residuals, $\hat{e}_{jk}$ for both classes. Using (7.30), these are as follows:

PSYCH	Public services $\hat{e}_{1k}$	Other industries $\hat{e}_{2k}$
1	−2.79	2.79
2	0.78	−0.78
3	2.35	−2.35

The residuals sum up to zero across public services and other industries. Note that from absolute values of the standardized residuals the largest are in the first and third PSYCH classes. In the third PSYCH class, the direction of the difference favours those from public services, whereas in the first class the situation is the opposite. The design-based standardized residuals also exceed the 1% critical value 2.33 from the standard normal $N(0,1)$ distribution in these classes.

In the case where all relevant information is available, we conclude that the design-based Wald statistic provides an adequate and usable testing procedure for the homogeneity hypothesis. And if only cell design effects are provided, but not the two regional covariance-matrix estimates, we would choose the Rao–Scott adjustment to a Pearson or Neyman test statistic. But inferential conclusions remain unchanged independently of the test statistic chosen in the case considered; the strength of the conclusion to reject the null hypothesis of homogeneity of PSYCH proportions over the two populations, however, varies somewhat.

Logit modelling provides a convenient general framework for the test of a homogeneity hypothesis. A test of homogeneity of PSYCH proportions in the INDU (type of industry) classes in a 2×3 table can be taken as a simple example of a logit model for a polytomous response variable. A test of homogeneity is obtained by fitting the saturated logit model INTERCEPT + INDU, say, for PSYCH logits and then by testing by the Wald test the significance of the INDU term. The observed value of the Wald test statistic is $X^2_{des} = 8.13$ with a p-value 0.0171. The result, although slightly more conservative, is compatible with the previous results from the Wald test statistic X^2_{des}.

The Case of More than Two Regions

We have considered a test of homogeneity for two regions, where the regions constitute segregated classes. Derivation of a design-based Wald statistic, and the Rao–Scott adjustments to the Pearson and Neyman test statistics, for the case of more than two segregated regions is straightforward, but involves more matrix algebra. We omit the derivations and refer the reader to Rao and Thomas (1988).

The test of homogeneity for segregated classes is a special case of a more general testing situation with any type of categorical predictor variable. This case, with a binary response variable, is considered in Chapter 8 for logit modelling.

There, the assumption of segregated-type regions is relaxed, and we work with cross-classes also for the predictor variable. Then, the design-based covariance matrices of the response variable proportions cannot be estimated separately in the predictor variable subgroups, as was done in the segregated regions case, but the between-region covariance must be estimated as well. This covariance was assumed zero for segregated regions.

7.5 TEST OF INDEPENDENCE

A test of independence is applied to study whether there is nonzero association between two categorical variables within a population. Organized in an $r \times c$ contingency table, the data are thus assumed to be drawn from a single population with no fixed margins. Therefore, it is assumed that the sum of all population proportions p_{jk} in the population table equals one. The population table is thus:

First variable	Second variable						
	1	2	...	k	...	c	All
1	p_{11}	p_{12}	$\cdots$	p_{1k}	$\cdots$	p_{1c}	p_{1+}
2	p_{21}	p_{22}	$\cdots$	p_{2k}	$\cdots$	p_{2c}	p_{2+}
$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$
j	p_{j1}	p_{j2}	$\cdots$	p_{jk}	$\cdots$	p_{jc}	p_{j+}
$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$
r	p_{r1}	p_{r2}	$\cdots$	p_{rk}	$\cdots$	p_{rc}	p_{r+}
All	p_{+1}	p_{+2}	$\cdots$	p_{+k}	$\cdots$	p_{+c}	1

For the formulation of the null hypothesis, and for the interpretation of test results, it is important to note that we are now working in a symmetrical case where neither of the classification variables is assumed to be a predictor. The two response variables with r and c categories are typically of cross-classes or mixed-classes type so that they cut across the strata and clusters. A hypothesis of independence of the response variables was formulated in Section 7.3 as $H_0 : p_{jk} = p_{j+}p_{+k}$, where $p_{jk} = N_{jk}/N$, and $p_{j+} = \sum_{k=1}^{c} p_{jk}$ and $p_{+k} = \sum_{j=1}^{r} p_{jk}$ are marginal proportions with $j = 1, \ldots, r$ and $k = 1, \ldots, c$. It is obvious that if the actual unknown cell proportions p_{jk} were close to the expected cell proportions $p_{j+}p_{+k}$ under the null hypothesis, then the two variables can be assumed independent. This fact is utilized in the construction of appropriate test statistics for the independence hypothesis.

For the derivation of the test statistics of independence, we write the null hypothesis in an equivalent form, $H_0 : F_{jk} = p_{jk} - p_{j+}p_{+k} = 0$, where $j = 1, \ldots, r-1$

and $k = 1, \ldots, c-1$ because of the constraint $\sum_{j=1}^{r} \sum_{k=1}^{c} p_{jk} = 1$. The F_{jk} are thus the residual differences between the unknown cell proportions and their expected values under the null hypothesis, which states that the residual differences are all zero. The residuals can then be collected in a column vector $\mathbf{F} = (F_{11}, \ldots, F_{1,c-1}, \ldots, F_{r-1,1}, \ldots, F_{r-1,c-1})'$ with a total of $(r-1)(c-1)$ rows.

The estimated cell proportions $\hat{p}_{jk} = \hat{n}_{jk}/n$, obtained from a sample of n elements, provide consistent estimators of the corresponding unknown proportions p_{jk}, where $\hat{n}_{jk}$ are scaled weighted-up cell frequencies accounting for unequal element inclusion probabilities and nonresponse, such that $\sum_{j=1}^{r} \sum_{k=1}^{c} \hat{n}_{jk} = n$. The $\hat{p}_{jk}$ are ratio estimators when working with a subgroup of the total sample whose size is not fixed in advance, such as the demonstration data sets from the MFH and OHC Surveys. As for the goodness-of-fit and homogeneity hypotheses, we also make this assumption here.

Covariance-matrix Estimators

Let us first derive the covariance-matrix estimators of the estimated vector $\hat{\mathbf{F}}$ of the residual differences under various assumptions on the sampling design, to be used for a design-based Wald statistic and for Pearson and Neyman test statistics. The estimated vector of residual differences is

$$\hat{\mathbf{F}} = (\hat{F}_{11}, \ldots, \hat{F}_{1,c-1}, \ldots, \hat{F}_{r-1,1}, \ldots, \hat{F}_{r-1,c-1})', \tag{7.31}$$

where $\hat{F}_{jk} = \hat{p}_{jk} - \hat{p}_{j+}\hat{p}_{+k}$, and $\hat{p}_{j+}$ and $\hat{p}_{+k}$ are estimators of the corresponding marginal proportions. For the design-based Wald statistic, we derive the consistent covariance-matrix estimator $\hat{\mathbf{V}}_F$ of $\hat{\mathbf{F}}$, accounting for complexities of the sampling design, given by

$$\hat{\mathbf{V}}_F = \hat{\mathbf{H}}'\hat{\mathbf{V}}_{des}\hat{\mathbf{H}}, \tag{7.32}$$

where the $(r-1)(c-1) \times (r-1)(c-1)$ matrix $\hat{\mathbf{H}}$ is the matrix of partial derivatives of $\mathbf{F}$ with respect to p_{jk}, evaluated at $\hat{p}_{jk}$. The matrix $\hat{\mathbf{V}}_{des}$ is a consistent estimator of the asymptotic covariance matrix $\mathbf{V}/n$ of the vector of cell proportion estimators $\hat{\mathbf{p}} = (\hat{p}_{11}, \ldots, \hat{p}_{1,c-1}, \ldots, \hat{p}_{r-1,1}, \ldots, \hat{p}_{r-1,c-1})'$. An estimate $\hat{\mathbf{V}}_{des}$ is obtained by the linearization method as used previously for the goodness-of-fit and homogeneity hypotheses. In practice, $\hat{\mathbf{V}}_{des}$ can be calculated from the element-level data set by fitting a full-interaction linear model without an intercept, with the categorical variables as the model terms. The estimated model coefficients then coincide with the observed proportions, and the covariance-matrix estimate of the coefficients provides an estimate $\hat{\mathbf{V}}_{des}$.

The two multinomial covariance-matrix estimators of $\hat{\mathbf{F}}$ are as follows. For the Pearson test statistic, we derive an expected multinomial covariance-matrix estimator $\hat{\mathbf{P}}_{0F}/n$ of $\hat{\mathbf{F}}$ under the null hypothesis such that

$$\hat{\mathbf{P}}_{0F} = \hat{\mathbf{H}}'\hat{\mathbf{P}}_0\hat{\mathbf{H}}, \tag{7.33}$$

where $\hat{\mathbf{P}}_0 = \text{diag}(\hat{\mathbf{p}}_0) - \hat{\mathbf{p}}_0\hat{\mathbf{p}}_0'$ with $\hat{\mathbf{p}}_0$ being the vector of expected proportions under the null hypothesis, i.e. a vector with elements $\hat{p}_{j+}\hat{p}_{+k}$. And for the Neyman test statistic, we derive an observed multinomial covariance-matrix estimator $\hat{\mathbf{P}}_F/n$ of $\hat{\mathbf{F}}$ given by

$$\hat{\mathbf{P}}_F = \hat{\mathbf{H}}'\hat{\mathbf{P}}\hat{\mathbf{H}}, \tag{7.34}$$

where $\hat{\mathbf{P}} = \text{diag}(\hat{\mathbf{p}}) - \hat{\mathbf{p}}\hat{\mathbf{p}}'$. Note that all the covariance-matrix estimators of $\hat{\mathbf{F}}$ are of a similar form and use the same matrix $\hat{\mathbf{H}}$ of partial derivatives.

Design-based Wald Statistic

By using the estimated vector $\hat{\mathbf{F}}$ of residual differences with its consistent covariance-matrix estimate $\hat{\mathbf{V}}_F$ from (7.32), we obtain for the independence hypothesis a design-based Wald statistic

$$X^2_{des} = \hat{\mathbf{F}}'\hat{\mathbf{V}}_F^{-1}\hat{\mathbf{F}}, \tag{7.35}$$

which is asymptotically chi-squared with $(r-1)(c-1)$ degrees of freedom. As in the Wald tests for goodness of fit and homogeneity, this test statistic can suffer from instability problems in cases in which only few degrees of freedom f are available for an estimate $\hat{\mathbf{V}}_F$. F-corrections to X^2_{des} can then be used, where

$$F_{1.des} = \frac{f-(r-1)(c-1)-1}{f(r-1)(c-1)}X^2_{des}, \tag{7.36}$$

which is referred to the F-distribution with $(r-1)(c-1)$ and $(f-(r-1)(c-1)-1)$ degrees of freedom, and

$$F_{2.des} = \frac{X^2_{des}}{(r-1)(c-1)}, \tag{7.37}$$

which in turn is referred to the F-distribution with $(r-1)(c-1)$ and f degrees of freedom.

Adjustments to Pearson and Neyman Test Statistics

A Pearson test statistic for an independence hypothesis in Section 7.3 was given as

$$X^2_{\text{P}} = n\sum_{j=1}^{r}\sum_{k=1}^{c}\frac{(\hat{p}_{jk}-\hat{p}_{j+}\hat{p}_{+k})^2}{\hat{p}_{j+}\hat{p}_{+k}}. \tag{7.38}$$

A Neyman test statistic can be used as an alternative and is given by

$$X^2_{\mathrm{N}} = n \sum_{j=1}^{r} \sum_{k=1}^{c} \frac{(\hat{p}_{jk} - \hat{p}_{j+}\hat{p}_{+k})^2}{\hat{p}_{jk}}. \tag{7.39}$$

Observed values of these statistics can be obtained from the estimated cell and marginal proportions. And under simple random sampling, both test statistics are asymptotically chi-squared with $(r-1)(c-1)$ degrees of freedom.

For a convenient common framework, we write the Pearson and Neyman test statistics (7.38) and (7.39) using the corresponding matrix formulae,

$$X^2_{\mathrm{P}} = n\hat{\mathbf{F}}'\hat{\mathbf{P}}^{-1}_{0F}\hat{\mathbf{F}} \tag{7.40}$$

for the Pearson statistic, where the null multinomial covariance-matrix estimator $\hat{\mathbf{P}}_{0F}/n$ from (7.33) is used, and

$$X^2_{\mathrm{N}} = n\hat{\mathbf{F}}'\hat{\mathbf{P}}^{-1}_{F}\hat{\mathbf{F}} \tag{7.41}$$

for the Neyman statistic, where the empirical multinomial covariance-matrix estimator $\hat{\mathbf{P}}_F/n$ from (7.34) is used. Note that both statistics mimic the design-based Wald statistic X^2_{des} in (7.35), the only difference being which covariance-matrix estimator of the residual differences is used. It should also be noted that in the calculation of X^2_{des}, X^2_{P} and X^2_{N}, the vector $\hat{\mathbf{F}}$ is an $(r-1)(c-1)$ column vector, and the covariance-matrix estimates are $(r-1)(c-1) \times (r-1)(c-1)$ matrices. Thus, for example, in a 2×2 table, $\hat{\mathbf{F}}$ and the covariance-matrix estimates $\hat{\mathbf{P}}_{0F}$ and $\hat{\mathbf{P}}_F$ reduce to scalars.

In complex surveys, there is a similar motivation to adjusting the statistics X^2_{P} and X^2_{N} for the clustering effect as in the corresponding tests of goodness of fit and homogeneity. Asymptotically valid adjusted test statistics are obtained using second-order Rao–Scott corrections given by

$$X^2_{\mathrm{P}}(\hat{\delta}_\bullet, \hat{a}^2) = X^2_{\mathrm{P}}/(\hat{\delta}_\bullet(1 + \hat{a}^2)) \tag{7.42}$$

for the Pearson statistic (7.40), where

$$\hat{\delta}_\bullet = \mathrm{tr}(\hat{\mathbf{D}})/((r-1)(c-1))$$

is the mean of the eigenvalues $\hat{\delta}_l$ of the generalized design-effects matrix estimate

$$\hat{\mathbf{D}} = n\hat{\mathbf{P}}^{-1}_{0F}\hat{\mathbf{V}}_F, \tag{7.43}$$

and

$$\hat{a}^2 = \sum_{l=1}^{(r-1)(c-1)} \hat{\delta}^2_l/((r-1)(c-1)\hat{\delta}^2_\bullet) - 1$$

is again the squared coefficient of variation of the eigenvalue estimates $\hat{\delta}_l$, with the sum of squared eigenvalues given by

$$\sum_{l=1}^{(r-1)(c-1)} \hat{\delta}_l^2 = \mathrm{tr}(\hat{\mathbf{D}}^2).$$

The second-order adjusted statistic (7.42) is asymptotically chi-squared with Satterthwaite adjusted degrees of freedom

$$\mathrm{df}_S = \frac{(r-1)(c-1)}{(1+\hat{a}^2)}.$$

A similar second-order correction can also be made to X^2_{N}. There, a design-effects matrix estimate $\hat{\mathbf{D}} = n\hat{\mathbf{P}}_{0F}^{-1}\hat{\mathbf{V}}_F$ can alternatively be used.

Both the design-based Wald statistic X^2_{des} and the second-order Rao–Scott adjustments to X^2_{P} and X^2_{N} require availability of the full covariance-matrix estimate $\hat{\mathbf{V}}_{des}$ of the cell proportion estimators $\hat{p}_{jk}$. In secondary analysis situations, this estimate is not necessarily provided, but cell design-effect estimates $\hat{d}_{jk}$, possibly with marginal design-effect estimates $\hat{d}_{j+}$ and $\hat{d}_{+k}$, might be reported. By using these design-effect estimates, approximative first-order corrections can then be obtained. The simplest mean deff adjustment to the Pearson statistic X^2_{P} is calculated using the mean of the estimated cell design effects given by

$$X^2_{\mathrm{P}}(\hat{d}_{\cdot}) = X^2_{\mathrm{P}}/\hat{d}_{\cdot}, \tag{7.44}$$

where $\hat{d}_{\cdot} = \sum_{j=1}^{r}\sum_{k=1}^{c}\hat{d}_{jk}/(rc)$ is the average cell design effect. And the first-order Rao–Scott adjustment to X^2_{P} is given by

$$X^2_{\mathrm{P}}(\hat{\delta}_{\cdot}) = X^2_{\mathrm{P}}/\hat{\delta}_{\cdot}, \tag{7.45}$$

where $\hat{\delta}_{\cdot}$ can be calculated from the cell and marginal design effects by

$$\hat{\delta}_{\cdot} = \frac{1}{(r-1)(c-1)}\sum_{j=1}^{r}\sum_{k=1}^{c}\frac{\hat{p}_{jk}(1-\hat{p}_{jk})}{\hat{p}_{j+}\hat{p}_{+k}}\hat{d}_{jk} - \sum_{j=1}^{r}(1-\hat{p}_{j+})\hat{d}_{j+} - \sum_{k=1}^{c}(1-\hat{p}_{+k})\hat{d}_{+k}$$

without calculating the generalized design-effects matrix itself. Similar corrections can again be made to X^2_{N}. The statistics $X^2_{\mathrm{P}}(\hat{d}_{\cdot})$ and $X^2_{\mathrm{P}}(\hat{\delta}_{\cdot})$ are referred to the chi-squared distribution with $(r-1)(c-1)$ degrees of freedom. $X^2_{\mathrm{P}}(\hat{\delta}_{\cdot})$ is usually superior to $X^2_{\mathrm{P}}(\hat{d}_{\cdot})$, and the statistic $X^2_{\mathrm{P}}(\hat{\delta}_{\cdot})$ can be expected to work adequately if the variation in the eigenvalue estimates $\hat{\delta}_l$ is small.

If instability problems due to a relatively small f are expected, an F-correction to $X^2_{\mathrm{P}}(\hat{\delta}_{\bullet})$ can be obtained by

$$FX^2_{\mathrm{P}}(\hat{\delta}_{\bullet}) = X^2_{\mathrm{P}}(\hat{\delta}_{\bullet})/((r-1)(c-1)), \tag{7.46}$$

which is referred to the F-distribution with $(r-1)(c-1)$ and f degrees of freedom. A similar correction is also available for the first-order adjusted Neyman statistic $X^2_{\mathrm{N}}(\hat{\delta}_{\bullet})$.

Residual Analysis

If the null hypothesis of independence is rejected, then the standardized design-based cell residuals can be obtained for a closer examination of deviations from H_0. These residuals are given by

$$\hat{e}_{jk} = \frac{\hat{F}_{jk}}{\text{s.e}\,(\hat{F}_{jk})}, \tag{7.47}$$

where $\text{s.e}(\hat{F}_{jk})$ is the design-based standard-error estimate of $\hat{F}_{jk}$, i.e. square root of the corresponding variance estimate from (7.32). Under positive intra-cluster correlation, these design-based residuals tend to be smaller than the corresponding residuals calculated assuming simple random sampling. These would be obtained by inserting $\text{s.e}_0(\hat{F}_{jk})$ in place of $\text{s.e}(\hat{F}_{jk})$, where $\text{s.e}_0(\hat{F}_{jk})$ is the multinomial standard-error estimate of $\hat{F}_{jk}$, i.e. the square root of the corresponding variance estimate from (7.33).

Example 7.3

The test of independence of health hazards of work and psychic strain in the OHC Survey. Let us study whether the variables PHYS (physical health hazards of work: none or some) and PSYCH (overall psychic strain classified into three nearly equally sized classes) are associated or not. Note that both classification variables constitute cross-classes. The appropriate cross-tabulation is displayed in Table 7.4.

A hypothesis of independence is stated as $H_0 : p_{jk} = p_{j+}p_{+k}$ with $j = 1, 2$ and $k = 1, 2, 3$, or, analogously, $H_0 : p_{11} - p_{1+}p_{+1} = 0$ and $p_{12} - p_{1+}p_{+2} = 0$. The design-effect estimates of the cell proportions indicate a noticeable clustering effect, which is due to strong intra-cluster correlation for the variable PHYS, as can be seen from the corresponding marginal design-effect estimate, which is deff = 7.17. There is a natural interpretation for this unusually large design-effect estimate: separate establishments tend to be internally homogeneous with respect

Table 7.4 Cell and marginal proportions of variables PHYS (physical health hazards) and PSYCH (overall psychic strain) in the OHC Survey (design-effect estimates in parentheses).

	PSYCH				
PHYS	1	2	3	All	*n*
None	0.2276 (2.09)	0.2188 (2.26)	0.2078 (2.63)	0.6543 (7.17)	5130
Some	0.1161 (2.82)	0.1047 (2.37)	0.1250 (2.87)	0.3457 (7.17)	2711
All	0.3437 (1.77)	0.3236 (1.23)	0.3327 (1.61)	1.00	
n	2695	2537	2609		7841

to physical working conditions, but sites from different industries can differ noticeably from each other in their working conditions. For the variable PSYCH, on the other hand, marginal design effects are only moderate, which is also understandable because experiencing psychic symptoms cannot be expected to be a strongly workplace-specific phenomenon. The mean of cell design-effect estimates is also quite large, 2.51. It is therefore important that a valid testing procedure should account for the clustering effect.

For the test statistics (7.35), (7.38) and (7.39), the corresponding covariance-matrix estimates $\hat{\mathbf{V}}_F$, $\hat{\mathbf{P}}_{0F}$ and $\hat{\mathbf{P}}_F$ of residual differences $\hat{F}_{jk}$ are required.

Technically, in the calculation of these estimates, the full $(rc) \times (rc)$ estimate $\hat{\mathbf{H}}$ of the partial derivatives and the corresponding full covariance-matrix estimates $\hat{\mathbf{V}}_{des}$, $\hat{\mathbf{P}}_0$ and $\hat{\mathbf{P}}$ are used, but in the construction of the test statistics, only the $(r-1)(c-1) \times (r-1)(c-1)$ submatrices of these matrices are used. For the 2×3 table, we thus calculate the 6×6 full matrices but use only the 2×2 submatrices of these. A full 6×6 covariance-matrix estimate $\hat{\mathbf{V}}_{des}$ is first obtained using the linearization method. It is

$$\hat{\mathbf{V}}_{des} = 10^{-5} \begin{bmatrix} 4.6922 & 0.3207 & 0.6599 & -1.6442 & -1.6965 & -2.3321 \\ 0.3207 & 4.9264 & 1.7922 & -2.5751 & -2.1611 & -2.3030 \\ 0.6599 & 1.7922 & 5.5279 & -2.8972 & -2.5938 & -2.4890 \\ -1.6442 & -2.5751 & -2.8972 & 3.6938 & 1.9619 & 1.4608 \\ -1.6965 & -2.1611 & -2.5938 & 1.9619 & 2.8332 & 1.6562 \\ -2.3321 & -2.3030 & -2.4890 & 1.4608 & 1.6562 & 4.0072 \end{bmatrix}.$$

In addition to $\hat{\mathbf{V}}_{des}$, the matrix $\hat{\mathbf{H}}$ of partial derivatives is calculated to obtain the covariance-matrix estimate $\hat{\mathbf{V}}_F = \hat{\mathbf{H}}'\hat{\mathbf{V}}_{des}\hat{\mathbf{H}}$ of the vector of the residual differences, $\hat{\mathbf{F}}$. In the construction of the Wald statistic, we use the 2×1 vector of residual

differences,

$$\hat{\mathbf{F}} = \begin{bmatrix} \hat{F}_{11} \\ \hat{F}_{12} \end{bmatrix} = \begin{bmatrix} \hat{p}_{11} - \hat{p}_{1+}\hat{p}_{+1} \\ \hat{p}_{12} - \hat{p}_{1+}\hat{p}_{+2} \end{bmatrix} = 10^{-3} \begin{bmatrix} 2.778 \\ 7.162 \end{bmatrix},$$

and the corresponding 2×2 submatrix from the full $\hat{\mathbf{V}}_F$, calculated as

$$\hat{\mathbf{V}}_F = 10^{-6} \begin{bmatrix} 7.8147 & -2.8281 \\ -2.8281 & 6.3930 \end{bmatrix}.$$

For the design-based Wald statistic $X^2_{des} = \hat{\mathbf{F}}' \hat{\mathbf{V}}_F^{-1} \hat{\mathbf{F}}$, we obtain an observed value $X^2_{des} = 13.41$, which, referred to the chi-squared distribution with 2 degrees of freedom, attains a p-value 0.0012, significant at the 0.01 level. The F-corrections (7.36) and (7.37) to X^2_{des} give observed values $F_{1.des} = 6.68$, which, referred to the F-distribution with 2 and 244 degrees of freedom, attains a p-value 0.0015, and $F_{2.des} = 6.71$, which with 2 and 245 degrees of freedom attains the same p-value. The F-corrections do not contribute noticeably to the uncorrected X^2_{des}.

For the alternative asymptotically valid tests based on the second-order adjustment to the Pearson test statistic X^2_{P}, or the Neyman statistic X^2_{N}, we first calculate the estimated generalized design-effects matrix (7.43) as follows:

$$\hat{\mathbf{D}} = n\hat{\mathbf{P}}_{0F}^{-1} \hat{\mathbf{V}}_F = \begin{bmatrix} 1.30761 & 0.21651 \\ 0.08616 & 1.05628 \end{bmatrix}.$$

The first-order adjustment factor is $\hat{\delta}_{\bullet} = \mathrm{tr}(\hat{\mathbf{D}})/2 = 1.182$, and the sum of squared eigenvalues is $\sum_{l=1}^{2} \hat{\delta}_l = \mathrm{tr}(\hat{\mathbf{D}}^2) = 2.863$, giving a second-order correction factor $(1 + \hat{a}^2) = 1.025$. These figures indicate that the eigenvalues are close to one on average, and their variation is negligible.

For the unadjusted test statistics (7.38) and (7.39), the observed values $X^2_{\mathrm{P}} = 16.40$ and $X^2_{\mathrm{N}} = 16.59$ are obtained, both of which, referred to the chi-squared distribution with 2 degrees of freedom, attain a p-value 0.0003, which is significant at the 0.001 level. Note that X^2_{P} and X^2_{N} are considerably liberal relative to X^2_{des}. For the second-order Rao–Scott adjusted Pearson statistic (7.42), an observed value $X^2_{\mathrm{P}}(\hat{\delta}_{\bullet}, \hat{a}^2) = 13.68$ is obtained, which, referred to the chi-squared distribution with Satterthwaite adjusted degrees of freedom $\mathrm{df}_S = 1.952$, attains a p-value 0.0010. This test appears somewhat liberal relative to the design-based Wald statistic, which also seems to work reasonably in this OHC Survey testing situation (see Example 7.2).

With the availability of only limited information, we calculate the first-order adjustments (7.44), (7.45) and (7.46) to the Pearson statistic by using the design-effect estimates in Table 7.4. The mean deff adjustment, with an observed value $X^2_{\mathrm{P}}(\hat{d}_{\bullet}) = 6.60$ and a p-value 0.0369, is overly conservative relative to the first-order Rao–Scott adjustment $X^2_{\mathrm{P}}(\hat{\delta}_{\bullet}) = 14.02$, with a p-value 0.0009, and its

F-correction $FX^2_P(\hat{\delta}_\cdot) = 7.01$, which attains a p-value 0.0011. Conservativity of the mean deff adjustment arises because $\hat{d}_\cdot = 2.51$ considerably overestimates the mean $\overline{\delta}$ of the true eigenvalues, and the estimate $\hat{\delta}_\cdot = 1.182$, calculated using cell and marginal design-effect estimates, provides a much better estimate. This suggests a warning against the use of the mean deff adjustment if either of the classification variables is strongly intra-cluster correlated. The F-corrected first-order Rao–Scott adjustment works very reasonably when compared to the design-based Wald statistic and the second-order Rao–Scott adjustment.

The tests suggest rejection of the null hypothesis of independence of PHYS and PSYCH. We finally calculate the design-based standardized cell residuals by using (7.47):

	PHYS	
PSYCH	None $\hat{e}_{1k}$	Some $\hat{e}_{2k}$
1	0.99	−0.99
2	2.83	−2.83
3	−3.40	3.40

The residual analysis shows that the largest deviations are in the last PSYCH class so that the direction of the difference favours those suffering from physical health hazards of work. Standardized residuals in these classes exceed the 0.1% critical value 2.58 from the $N(0, 1)$ distribution. Note that the sum of residuals is zero across the two PHYS classes.

Also, in this testing situation, as in Example 7.2, the design-based Wald statistic behaves adequately because of the relatively large number of sample clusters (250), and we may conclude that the Wald test provides a reasonable testing procedure for the independence hypothesis of PHYS and PSYCH. And if only the cell and marginal design effects are provided, we would choose the F-corrected first-order Rao–Scott adjustment to the Pearson (or Neyman) statistic. But if only the cell design effects are provided and not the marginal design effects, difficulties would arise in obtaining an approximately valid testing procedure because of the apparent over-conservativity of the mean deff adjustment in such a case.

The test of independence in a two-way table can also be executed as a test of no interaction for an appropriate log-linear model with two categorical variables. The independence test is obtained by fitting the saturated log-linear model INTERCEPT + PHYS + PSYCH + PHYS*PSYCH, say, and then by testing with the Wald test the significance of the interaction of PHYS and PSYCH, i.e. the item PHYS*PSYCH. The design-based Wald statistic gives an observed value $X^2_{des} = 13.83$, with a p-value 0.0012, and is compatible with the previous results.

7.6 CHAPTER SUMMARY AND FURTHER READING

Summary

For a goodness-of-fit test and tests of homogeneity and independence on tables from complex surveys, testing procedures are available that properly account for the complexities of the sampling design. These complexities include the weighting of observations for obtaining consistently estimated proportions, and intra-cluster correlations, which arise due to the clustering and are usually positive. Generally, valid testing procedures include the design-based Wald test and the second-order adjustment to the Pearson and Neyman test statistics.

The design-based Wald test can be expected to work adequately when working with large samples in which a large number of sample clusters are also available. This was the case in the OHC Survey. A drawback to the Wald test is its sensitivity to such small-sample situations where only a small number of sample clusters are present, leading to unexpectedly liberal test results. The MFH Survey appeared to be an example of such a design. The degrees-of-freedom corrections to the Wald statistic, leading to F-type test statistics, can be used to account for possible instability. The second-order Rao–Scott adjustment to the Pearson and Neyman test statistics can be expected not to be seriously sensitive to instability problems. This adjustment appeared to work reasonably in both the OHC and MFH Surveys.

A full design-based covariance-matrix estimate is required for the design-based Wald test and for the second-order Rao–Scott adjustments. In secondary analyses on published tables, where such a covariance-matrix estimate is not supplied, only approximately valid first-order testing procedures are available. The mean deff adjustment to the standard test statistics can be used if only the cell design-effect estimates are provided. But this adjustment can be overly conservative, as seen in the example from the OHC Survey. The first-order Rao–Scott adjustment is superior to the mean deff adjustment, and, using an F-correction, the first-order adjustment can in some cases account for possible instability problems, as seen in the MFH Survey example.

Because a test of homogeneity can also be taken as a simple application of logit modelling, and a test of independence in turn as an application of log-linear modelling, these modelling approaches, implemented in software for the analysis of complex surveys, can also be used. For further training of these testing methodologies the reader is advised to visit the web extension of this book.

In hypothesis testing, a vector of finite-population cell proportions was considered. But, if the finite population is large, these proportions are close to the corresponding cell probabilities of the infinite superpopulation from which the finite population can be regarded as a single realization. Thus, the design-based inferences considered here also constitute an inference on the parameters of the appropriate infinite superpopulation.

Further Reading

The analysis of one-way and two-way frequency tables has received attention in the survey analysis literature. Articles by Holt *et al.* (1980) and Rao and Scott (1981, 1984, 1987) cover important theoretical developments of the 1980s. More applied sources include Hidiroglou and Rao (1987a, 1987b), and Rao and Thomas (1988, 1989). Thomas *et al.* (1996) evaluate various tests on independence on two-way tables under complex sampling.

There are also overviews and more specialize material available on this topic, such as the articles by Freeman and Nathan in the *Handbook of Statistics* (vol. 6, 1988) and a section in Särndal *et al.* (1992) and Lohr (1999). The duality between design-based and model-based inference is discussed, e.g. in Rao and Thomas (1988) and in Skinner *et al.* (1989). Rao and Thomas (2003) summarize many recent findings on the analysis of categorical response data from complex surveys.

8

Multivariate Survey Analysis

Multivariate methods provide powerful tools for the analysis of complex survey data. Multivariate analysis is discussed in this chapter in the case of one response variable and a set of predictor or explanatory variables. For this kind of analysis situation, logit models and linear models are widely used. Proper methods are available for fitting these models for intra-cluster correlated response variables from complex sampling designs. These methods have also been implemented in software products for survey analysis. With logit and linear modelling in complex surveys, as with the analysis of two-way tables, it is important to eliminate the effects of clustering from the estimation and test results. Examination of recent methodology for this task, supplemented with numerical examples, is the main focus in this chapter. The range of multivariate methods considered, and the basic logit and linear models, are introduced in Sections 8.1 and 8.2. The design-based and other analysis options used in multivariate analysis are also presented in Section 8.2. In Section 8.3, design-based analysis of categorical data is discussed and illustrated. Methods for logistic and linear regression analysis are treated in Section 8.4, and a summary is given in Section 8.5. The Occupational Health Care (OHC) Survey data, providing an example of a complex survey, is used in empirical applications. Materials presented in the examples are worked out further in the interactive web extension of the book.

8.1 RANGE OF METHODS

The aim in fitting multivariate models is to find a scientifically interesting but parsimonious explanation of the systematic variation of the response variable. This is achieved by modelling the variation with a reasonable set of predictor variables using the available survey data. For example, in a health survey based on

Practical Methods for Design and Analysis of Complex Surveys Risto Lehtonen and Erkki Pahkinen
 ISBN: 0-470-84769-7

a cluster sample of households, variation of health status and use of health services is to be studied in order to find possible high-risk population subgroups to target in developing a health promotion programme. Certain socioeconomic determinants of the sample households and demographic and behavioural characteristics of household members are used as predictor variables. In an educational survey based on cluster sampling of teaching groups, one may wish to study the effect of the teacher, and that of the students, on the differences in learning. Further, in a survey on health-related working conditions, the association of perceived psychic (psychological or mental) strain with certain physical and other working conditions can be studied, again on the basis of data from cluster sampling with industrial establishments as the clusters. In all these surveys, the data would be collected with cluster sampling, but inferences concern mainly a person-level population or, more generally, relationships of the person-level variables under a superpopulation framework.

Response variables in the example surveys were binary (chronic sickness is present or not present; psychic strain is low or high), polytomous (learning outcomes are poor, medium or good), or quantitative or continuous (the number of physician visits; principal component score of psychic strain). Logit modelling on a binary or polytomous response and linear modelling on continuous measurements provide two popular approaches to these cases. If cluster sampling is used, as in the example surveys, the response variables are exposed to intra-cluster correlations. The consequences of intra-cluster correlation are discussed briefly in the following introductory example.

Introductory Example

Let us consider more closely the cases of a binary and a continuous response variable. With categorical predictors, the data for a binary response can be arranged in a table of proportions, and for a continuous response, in a table of means. From the OHC Survey, we have the following table of perceived overall psychic strain (PSYCH), which is originally a continuous variable of scores of the first principal component from a set of psychic symptoms. For a binary response, the variable PSYCH is recoded so that the value zero indicates strain below the mean (low-strain group), and the value one indicates strain above the mean (high-strain group). In the table, we have three categorical predictors, each with two classes: sex and age of respondent, and the variable PHYS (physical health hazards), which measures physical working conditions coded so that the value one indicates more hazardous work. The domains are formed by cross-classifying the predictors, and they cut across the sample clusters. The main interests are in the relation of psychic strain to physical working conditions. In Example 7.3, statistically significant dependence was noted for these variables, although in a slightly different setting where PSYCH was recoded as a three-class variable.

The percentage of persons experiencing above-average psychic strain in the whole sample is of course 50%, and in the risk group (PHYS = 1) this percentage was noted to be 52.2%, i.e. only slightly higher than in the other group. But, when inspecting the variation of percentage estimates in Table 8.1, it appears that there are certain subgroups with a high proportion of persons suffering from psychic strain. For both sexes, the proportions tend to increase with increasing age and, in a given age group, the proportions are higher for those involved in physically more hazardous work. There might also exist an interaction between age and physical working conditions.

Thus, the variation in the proportions of the binary response is quite logical. Obviously, the variation in the means of the corresponding continuously measured psychic strain follows a similar pattern. A logit analysis would be chosen for the analysis of the domain proportions, and linear modelling is appropriate for the domain means. Because the predictors are categorical, an analysis-of-variance-type model would be selected in both cases. If the data were obtained with simple random sampling (SRS), the analysis would technically be a standard one: take a procedure for binomial logit modelling and for linear analysis of variance (ANOVA) from any commercial program package, search for well-fitting and parsimonious logit and linear models and draw conclusions.

But in the OHC Survey, cluster sampling was used with establishments as the clusters. Positive intra-cluster correlation can thus be expected for the response variable PSYCH, as in Example 7.3. This correlation can disturb the analysis in such a way that if it is ignored, erroneous conclusions might be drawn. From Table 8.1 it can be seen that design-effect estimates of proportions are larger than

Table 8.1 Proportion (%) of persons in the upper psychic strain group, and mean of the continuously measured psychic strain, in domains formed by sex, age and physical working conditions of respondent, and design-effect estimates of the proportions and means (the OHC Survey; $n = 7841$ employees).

				PSYCH (Binary)		PSYCH (Continuous)	
Domain	SEX	AGE	PHYS	%	deff	Mean	deff
1	Males	−44	0	41.9	1.16	−0.193	1.14
2			1	47.2	1.33	−0.084	1.36
3		45−	0	46.1	0.87	−0.075	1.05
4			1	52.0	1.18	0.139	1.25
5	Females	−44	0	54.1	1.23	0.065	1.61
6			1	62.0	1.38	0.264	1.46
7		45−	0	53.2	1.65	0.098	1.74
8			1	70.0	1.47	0.656	1.44
All				50.0	1.69	0.000	1.97

one on average, with an overall design-effect estimate deff = 1.7, indicating a noticeable clustering effect. For a proper analysis, this clustering effect should be taken into account, and a simpler model for the variation of PSYCH proportions can be obtained than by ignoring the clustering effects, as will be seen in Example 8.1.

Two Main Approaches

There are two main approaches available for proper multivariate analysis of an intra-cluster correlated response variable such as PSYCH. If intra-cluster correlation is taken to be a nuisance, one may make efforts to eliminate this disturbance effect from the estimation and test results, as was done in Chapter 7. The *nuisance approach*, covering a variety of methods for logit and linear modelling, has been developed over a long period, mainly within the context of survey sampling. This approach is sometimes referred to as the *aggregated* approach. In Chapter 8, we will discuss methods commonly used in fitting logit and linear models for complex survey data under the nuisance approach, based on variants of least squares (LS) estimation and maximum likelihood (ML) estimation.

If, on the other hand, clustering is interesting as a structural property of the population, it can be examined with appropriate models. This approach has been developed under a general framework of *multi-level modelling* for hierarchically structured data sets. Multi-level modelling can also be applied to multivariate analysis of correlated responses from clustered designs. However, for complex surveys, the nuisance approach has had a dominant role, and it is the main approach used here. The alternative approach, which can also be called *disaggregated*, will be briefly discussed in this chapter and demonstrated in Chapter 9.

Estimation Methods

There are alternative asymptotically valid estimation methods for modelling intra-cluster correlated response variables. For a binary or polytomous response variable, we apply a variant of the *generalized least squares* (*GLS*) estimation in cases where the data are arranged in a multidimensional table such as Table 8.1. In using GLS for complex survey data, element weights are incorporated in the estimation equations. We call it henceforth the *generalized weighted least squares* (*GWLS*) method. This simple noniterative method will be discussed in Section 8.3 for logit and linear modelling of categorical data. The GWLS method, introduced in Grizzle *et al.* (1969) and Koch *et al.* (1975), is applicable to a combination of linear, logarithmic and exponential functions on proportions. Thus, in addition to logit and linear models, log-linear models are also covered.

A widely used method for fitting models for binary, polytomous and count response variables in complex surveys is based on a modification of ML estimation such that the element weights are incorporated in the estimating equations. The

method, called *pseudolikelihood* (PML) estimation, will be considered in Section 8.4 for logit analysis on a binary response. In linear modelling on a continuous response, LS estimation will be used where element weights are also incorporated in the estimation; the method will be called the *WLS method*. In all these methods, proper design-based methods using approximation techniques, introduced in Chapter 5, are applied in covariance-matrix estimation of estimated regression coefficients. Linear and nonlinear models considered are special cases of a broad methodology for fitting *generalized linear models* following Nelder and Wedderburn (1972) and McCullagh and Nelder (1989) covering, for example linear, logit and log-linear models.

The third method is based on the methodology of *generalized estimating equations (GEE)* (Liang and Zeger 1986). The model parameters are estimated using the so-called *multivariate quasilikelihood* method. We will briefly discuss and apply this method in Section 8.4, because the method, like the PML method, has its roots in generalized linear models methodology.

In testing procedures, design-based Wald test statistics and second-order Rao–Scott adjusted test statistics can be used, providing asymptotically valid testing procedures. However, the test statistics may suffer from instability problems, especially when the number of sample clusters is small. Instability can disturb the behaviour of a design-based Wald statistic, resulting in overly liberal test results relative to the nominal levels and leading to unnecessarily complex models. This property is similar to that noted for Wald tests on two-way tables. To protect against the effects of instability, certain degrees-of-freedom corrections such as F-corrections are available.

Although there are many similarities in the estimation methods, their applicability and properties differ in certain respects. For further discussion, we next define the main types of linear and logit models, and more formally introduce the corresponding models.

8.2 TYPES OF MODELS AND OPTIONS FOR ANALYSIS

Three Types of Models

In linear models, the expectation of a continuous response variable is related to a linear expression on the predictors. In logit models, a nonlinear function of the expectation of a binary response variable, called a *logit* or *logistic* function, is related to a linear expression on the predictors. Note that both models share the property that the expression on the predictors is a linear one. But the essential difference is that in a linear model this predictor part is linearly related to the response variable and in a logit model a nonlinear relationship is postulated.

For introducing the types of linear and logit models, it is instructive to consider separately the case of multidimensional tables with categorical predictors and the case where the predictors are purely continuous (or at least one of them is). In

both instances, the response variable can be binary, polytomous, quantitative or continuous.

In multidimensional tables, such as Table 8.1, the predictors are categorical qualitative or categorized quantitative variables, and depending on additional assumptions on their types, special cases of linear and logit models are obtained. In models of ANOVA type, the classes of each predictor are taken to be qualitative. Sex, occupation, social class and type of industry are examples of commonly used predictors. For categorized quantitative predictors, monotonic ordering can be assumed on the classes of each predictor, and desired scores can be assigned to the classes. The predictors can then be taken to be continuous, leading to regression-type models. Age, systolic blood pressure, monthly income of a household and first principal component of psychic symptoms are examples of such predictors, each categorized into a small number of classes. Note that the classes of an originally quantitative variable can also be taken to be qualitative, as in Example 7.3. If both qualitative and quantitative categorical predictors are present, we may call the model an analysis of covariance or ANCOVA-type model. For ANOVA and ANCOVA models, it is common to include interaction terms in the model and test their significance, which often constitutes an essential part of model building.

Sometimes it is desirable to work with quantitative predictors without categorizing them and arranging the data in a multidimensional table. Thus, we have at least one continuous predictor, and depending on the types of the other predictors, the corresponding models are obtained. If all the predictors are continuous measurements, we have a regression-type model, and additional qualitative predictors result in an ANCOVA-type model. It should be noted that in this case we actually model individual-level differences, whereas in the former case we are modelling differences between subgroups of the population.

In the analysis of a continuous response variable, the traditional ANOVA, regression analysis and ANCOVA models constitute the commonly used special cases of a linear model. We use analogous terminology for logit models with a binary or polytomous response variable. For these, we therefore have the corresponding logit ANOVA, logit (or logistic) regression and logit (or logistic) ANCOVA types of models.

Logit and Linear Models for Proportions

The following examples often deal with logit and linear modelling on domain proportions of a binary response variable because of the simplicity and popularity of this analysis situation in practice. Let us thus introduce the logit and linear models in the case where the data are organized in a multidimensional table such that there are u domains that are formed by cross-classifying the categorical predictors, and the response variable is binary. A logit or a linear model can then be postulated for examining the systematic variation of the estimated domain

proportions of the response variable across the domains. The situation is thus essentially similar to that of Table 8.1.

Under a logit model, we deal with logarithms of *ratios of proportions* p_{j1} and p_{j2}, where the former is the proportion of 'success'. We denote this proportion by p_j; thus the other is $p_{j2} = 1 - p_j$. The variation is modelled by relating the functions of the form $\log(p_j/(1-p_j))$ of the unknown proportions p_j to linear functions of the form $b_1x_{j1} + b_2x_{j2} + \cdots + b_sx_{js}$, where 'log' refers to natural logarithm. A function $\log(p_j/(1-p_j))$ is called the *logit* or *log odds* of success. In the linear functions, b_k are the model coefficients to be estimated, of which the first coefficient b_1 is an intercept term. The values x_{jk} are for the predictor or explanatory variables x_k, with a constant value of one assigned to the first variable x_1. Other variables depend on the model type. In logit ANOVA, x_k are indicator variables for the classes of the predictors. In logistic regression, they are continuous-valued scores assigned to the classes or the original continuous measurements. And in logit ANCOVA, the x-variables constitute a mixture of indicator variables and continuous variables. Interpretation of the coefficients b_k depends on the model type and on the parametrization used under a specific model. An advantage of the logit model is that *odds-ratio*-type statistics are readily available, and in special cases, interpretations with the concepts of independence and conditional independence are also possible.

Under linear modelling on proportions, on the other hand, we deal directly with *differences of proportions*. Thus, the population proportions p_j are related linearly to the linear functions $b_1x_{j1} + b_2x_{j2} + \cdots + b_sx_{js}$. This model formulation can be equally appropriate as a logit formulation, and it involves certain convenient interpretations. But interpretations by independence or related natural terminology are excluded.

The logit and linear models can be compactly written in a matrix form. Let $\mathbf{p} = (p_1, \ldots, p_u)'$ be the vector of unknown domain proportions, $\mathbf{b} = (b_1, \ldots, b_s)'$ be the vector of model coefficients and let $\mathbf{X}$ be the $u \times s$ matrix of x_{jk} such that the columns of the matrix represent the values of the variables x_k. Usually, $\mathbf{X}$ is called the *model matrix*. A hypothesized model can be written in the form

$$F(\mathbf{p}) = \mathbf{Xb}, \tag{8.1}$$

where, in the case of a logit model, the function vector $F(\mathbf{p})$ of the unknown proportion vector $\mathbf{p}$ is formulated as

$$F(\mathbf{p}) = F(\mathbf{f}(\mathbf{b})) = \log\left(\frac{\mathbf{f}(\mathbf{b})}{1-\mathbf{f}(\mathbf{b})}\right), \tag{8.2}$$

and, in the case of a linear model, the function vector $F(\mathbf{p})$ equals $\mathbf{p}$ because F is simply an identity function. Further, for a logit model, the function vector $\mathbf{f}(\mathbf{b})$ is

derived using the inverse of the logit function:

$$\mathbf{f}(\mathbf{b}) = F^{-1}(\mathbf{Xb}) = \frac{\exp(\mathbf{Xb})}{1 + \exp(\mathbf{Xb})}, \tag{8.3}$$

where 'exp' refers to the exponential function. For a linear model, this function vector is obviously $\mathbf{f}(\mathbf{b}) = \mathbf{Xb}$. An important motivation for the logit function is that the values of the function vary between zero and one, i.e. in the same range as the proportions p_j themselves. Therefore, predicted proportions from a fitted logit model always fall in the range (0,1). This property does not necessarily hold for the linear model formulation.

As an illustration of the matrix expressions (8.1)–(8.3), let us consider the case with two dichotomous predictors A and B for logit and linear ANOVA models for proportions p_j of a binary response variable. There are thus four domains ($u = 4$) and the table of the unknown proportions p_j is as follows:

Domain	A	B	p_j
1	1	1	p_1
2	1	2	p_2
3	2	1	p_3
4	2	2	p_4

We have three sources of variation in the table: that due to the effect of A, that due to the effect of B and that due to the effect of the interaction of A and B. In order to cover all these sources of variation, a total of four coefficients b_k are included in the model $F(\mathbf{p}) = \mathbf{Xb}$. The coefficient b_1 is the intercept, b_2 is assigned to A, b_3 is assigned to B and b_4 is assigned to the interaction of A and B. This model is called a *saturated* model, and by choosing a specific model matrix $\mathbf{X}$, it can be expressed as

$$\begin{bmatrix} F(p_1) \\ F(p_2) \\ F(p_3) \\ F(p_4) \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \end{bmatrix}, \tag{8.4}$$

where for a logit model the functions $F(p_j)$ are the logits

$$F(p_j) = \text{logit}(p_j) = \log\left(\frac{p_j}{1 - p_j}\right), \quad j = 1, 2, 3, 4,$$

and for a linear model the functions are $F(p_j) = p_j$. In the model matrix $\mathbf{X}$ of (8.4), we first have a column of ones for the indicator variable x_1. Then, there are three columns of contrasts with values 1 or −1, of which the first is for the predictor A, i.e. for the indicator variable x_2, the second is for the predictor B, i.e. for the indicator

variable x_3, and the last one is for the interaction of A and B, i.e. for the indicator variable x_4. Note that each indicator variable sums to zero in this parametrization, and there is one indicator variable for each predictor and its interaction, because the predictors are two-class variables. Generally, there are $t-1$ columns in the model matrix for a t-class variable, and $(t-1) \times (v-1)$ columns for an interaction of a t-class variable and a v-class variable, corresponding to the degrees of freedom for a model term. The sum of these degrees of freedom is the number s of model coefficients.

The parametrization just applied is sometimes called a *marginal* or full-rank centre-point parametrization. Under this parametrization, for categorical predictors with more than two classes, each indicator variable is used with the others to contrast a given class with the average of all classes. For example, in a logit ANOVA model, the coefficients b_k indicate differential effects on a logit scale, i.e. with respect to the average of all the fitted logits, and in a linear ANOVA model, they indicate differential effects on the untransformed scale, i.e. with respect to the average of all the fitted proportions.

It is important for proper inferences that we are fully aware of the specific parametrization applied, because there are also other commonly used parametrizations. For example, a parametrization called *partial* or reference-cell can be used. There, a specific reference class is assumed, and each indicator variable is used with the others to compare a given class with the reference class. Under this parametrization, we put zeros in place of -1 in the previous model matrix $\mathbf{X}$. This parametrization is especially useful when a definite reference group can be stated. In a logit model, the coefficients now indicate differential effects with respect to the fitted logit in the reference class, and in a linear model, differential effects with respect to the fitted proportion in the reference class. An odds ratio $\mathrm{OR}(b_k) = \exp(b_k)$ interpretation is readily available for logit models under partial parametrization.

Under these parametrizations, we have for the functions $F(p_j)$:

Marginal	Partial
$F(p_1) = b_1 + b_2 + b_3 + b_4$	$F(p_1) = b_1 + b_2 + b_3 + b_4$
$F(p_2) = b_1 + b_2 - b_3 - b_4$	$F(p_2) = b_1 + b_2$
$F(p_3) = b_1 - b_2 + b_3 - b_4$	$F(p_3) = b_1 + b_3$
$F(p_4) = b_1 - b_2 - b_3 + b_4$	$F(p_4) = b_1 + b_4$

Note that, because the functions $F(p_j)$ must be equal for both parametrizations, the corresponding coefficients b_k from these parametrizations cannot coincide. So, for example, the coefficient b_1 in the marginal parametrization is not equal to the b_1 in the partial parametrization.

Our discussion so far has been on logit and linear ANOVA models on domain proportions. A similar discussion applies for linear ANOVA models on domain means of a continuous response. For logit and linear regression and ANCOVA

models on binary responses, and for the corresponding linear models on continuous responses, the model matrices, however, are different, involving different interpretation of the model parameters.

Model Building in Practice

When fitting a specified logit or linear model, the primary task is to estimate the model coefficients b_k and the variances of the estimated coefficients. Using the resulting estimates, adequacy of the model is assessed by examining the goodness of fit of the model, and tests of linear hypotheses are executed on model coefficients. In practice, model building often involves repetition of this procedure several times for alternative models.

Let us consider further the logit and linear models on proportions. In a model-fitting procedure using standard notation, the previous ANOVA-type models can be written as $\mathrm{F(P)} = \log(\mathrm{P}/(1-\mathrm{P})) = \mathrm{A} + \mathrm{B} + \mathrm{A} * \mathrm{B}$ for a logit model, and $\mathrm{F(P)} = \mathrm{P} = \mathrm{A} + \mathrm{B} + \mathrm{A} * \mathrm{B}$ for a linear model with a binary response variable. There are three model terms corresponding to the predictors: two main effects and an interaction term. The model is saturated because it includes all the terms possible in this situation; the intercept term is included as a default in all the models. This kind of notation is commonly used for requesting a specified model structure, i.e. the terms desired in the linear part of the model, in many programs for linear and logit analysis.

A saturated model, including all possible main effects and interaction terms, is seldom interesting because the model includes as many parameters as there are degrees of freedom available. Also, the saturated model fits the data perfectly. In a model-building procedure, the aim is to reduce the saturated model in order to find a well-fitting model, which is parsimonious, so that as few model terms as possible are included.

Using the above notation, the possible models in these logit and linear ANOVA cases are as follows:

$$\begin{aligned}
&\mathrm{F(P)} = \mathrm{A} + \mathrm{B} + \mathrm{A} * \mathrm{B} && \text{(saturated model)},\\
&\mathrm{F(P)} = \mathrm{A} + \mathrm{B} && \text{(main effects model)},\\
&\mathrm{F(P)} = \mathrm{A} && \text{(model for the predictor A only)},\\
&\mathrm{F(P)} = \mathrm{B} && \text{(model for the predictor B only), and}\\
&\mathrm{F(P)} = \mathrm{INTERCEPT} && \text{(null model)}.
\end{aligned}$$

Reduced models are obtained by hierarchically removing statistically nonsignificant terms from a model. This procedure corresponds to removing columns (or sets of columns) from the model matrix. Usually, a well-fitting model for further use and for interpretation is found between the saturated and null models.

A model-building procedure in linear ANOVA on domain means resembles that of logit and linear ANOVA on domain proportions. In logistic and linear regression or ANCOVA-type models involving continuous predictors, an appropriate model is usually searched for by consecutively entering statistically significant or scientifically interesting terms, beginning from the null model. In these models, it should be noted that interactions are not allowed between the continuous predictors.

In complex surveys, estimation of the model coefficients of a logit ANOVA, ANCOVA or regression models on domain proportions can be executed by the GWLS, the PML or the GEE method. For logistic regression and ANCOVA models on a binary or polytomous response with strictly continuous predictors, the PML or the GEE method is used. In practice, all these models can be conveniently fitted with software for survey analysis.

Before entering into the details of modelling by GWLS, PML and GEE methods, we discuss in greater depth the special features of multivariate analysis when working with complex surveys. A number of options will be introduced for proper analysis under different sampling-design assumptions.

Options for Analysis

Here, we introduce a set of options for multivariate analysis of complex survey data involving clustering, stratification, multi-stage sampling and nonignorable nonresponse. In the presence of such complexities, consistent estimators of model coefficients and their variances, and valid test results, can be obtained by appropriately weighting the observations due to unequal inclusion probabilities and nonresponse, and by appropriately accounting for the intra-cluster correlations.

Three specific analysis options are presented: a design-based option and two options assuming simple random sampling (SRS), with or without replacement. Usually, a with-replacement assumption is used. We call the first option the *design-based* option, and it uses the actual, possibly complex, sampling design. In SRS-based options, an assumption of simple random sampling is made, irrespective of the possibly more complex sampling design actually used. The first SRS-based option incorporates the weighting due to adjustment for nonignorable unit nonresponse. We call it the *weighted SRS* option. The second SRS-based option is called the *unweighted SRS* option. It ignores the sampling complexities including the weighting.

An analysis under the design-based option accounts for all the sampling complexities, that is, weighting, stratification and clustering. The weighted SRS option ignores the stratification and clustering, and the unweighted SRS option ignores all the sampling complexities. The SRS options can be used as a reference for the design-based option when quantifying the effects of the design complexities on analysis results.

Under the design-based option, intra-cluster correlations, unequal element inclusion probabilities and adjustment for nonresponse can be properly accounted

for. This option is evidently the most appropriate for multivariate analysis in complex surveys. Therefore, design-based analysis is widely used in survey analysis, and it will be adopted in this chapter as the main analysis option.

The design-based option can in practice be applied in various ways, depending on special features of the sampling design and on software available for the analysis. Sampling designs involving weighting due to stratification or poststratification and several stages of sampling often require approximations to conveniently fit the design-based option. For data from two-stage stratified cluster sampling with a large population of clusters, a simple solution for this option is to reduce the design to one-stage stratified sampling where the primary sampling units are assumed to be drawn with replacement. This approximation is common in complex analytical surveys. Use of this approximation requires access to an element-level data set, which includes variables for stratum and cluster identification and for weighting. The approximation was used in the design-based analysis of frequency tables in Chapter 7 and will be used for multivariate analyses in this chapter.

In more advanced use of the design-based option, additional features of the sampling design can be accounted for, if necessary, for proper estimation. Examples are when the variation is due to several stages of sampling or sampling of clusters is with unequal probabilities without replacement. This presupposes the availability of population counts at each stage of sampling, and the calculation of single and joint selection probabilities of each primary sampling unit and each pair of PSUs in each first-stage stratum. Thus, more information must be supplied for an analysis program.

In addition to the above refinements, analysis under the design-based option can involve reorganization of the sample clusters into strata using the collapsed stratum technique, if only one primary sampling unit was originally drawn from each stratum, as was the case in the Mini-Finland Health (MFH) Survey. In some cases, additional weighting for poststratification is desirable. Many of these features have been implemented in software products for complex surveys.

In multivariate analysis of domain proportions of a binary response, it is assumed for the design-based option that an appropriate design-based covariance-matrix estimate of proportions can be calculated. In Chapter 5, we introduced a technique for obtaining a consistent covariance-matrix estimate based on the linearization method. Sample reuse methods, such as the jackknife, can also be used. This estimate is allowed to be nondiagonal because the correlations of the proportions from separate domains can be nonzero, which is the case when working with cross-classes or mixed classes. But when working with domains constituting segregated classes, it can be assumed that correlations of the proportions from separate domains are zero, because all elements in a given cluster fall in the same domain. In this case, the design-based covariance-matrix estimate simplifies to a diagonal matrix.

The SRS-based analysis options assume a binomial covariance matrix of the domain proportions, which is diagonal by definition. The validity of this assumption depends on the actual sampling design and the domain structure.

The SRS-based options assume simple random sampling with replacement. Under the weighted SRS option, it is assumed that the domain proportions are consistently estimated using the appropriate element weights, and a binomial covariance matrix is assumed for these proportions. Under the unweighted SRS option, simple random sampling with replacement is assumed, and the data set is assumed to be self-weighting. Thus, all the complexities of the sampling design are ignored.

Because the two versions of the SRS-based option are not valid for complex surveys involving clustering, they will be used as reference options for the design-based option and in the construction of appropriate generalized design-effect matrices. The weighted SRS option is used when assessing the magnitude of the clustering effects on results from multivariate analyses, and the unweighted SRS option can be used as a reference option for the design-based option when examining the effects of all the complexities of the sampling design on analysis results, including the effect of weighting procedures.

The analysis options with respect to sampling design are summarized below:

Option	Allowing weights	Allowing stratification	Allowing clustering
Design-based	Yes	Yes	Yes
Weighted SRS	Yes	No	No
Unweighted SRS	No	No	No

It should be noticed that in multivariate survey analysis, as in the analysis of two-way tables, the design-based approach to inference also constitutes inference on the parameters of the corresponding superpopulation model, provided that the finite population is large (see Rao and Thomas 1988).

8.3 ANALYSIS OF CATEGORICAL DATA

The GWLS method of generalized weighted least squares estimation provides a simple technique for the analysis of categorical data with ANOVA-type logit and linear models on domain proportions. Allowing all the complexities of a sampling design including stratification, clustering and weighting, the design-based option provides a generally valid GWLS analysis. Analysis under the weighted or unweighted SRS options assuming simple random sampling serves as a reference when studying the effects of clustering and weighting on results.

The GWLS method is computationally simple because it is noniterative for both logit and linear models on proportions. The alternative PML and GEE methods of pseudolikelihood and generalized estimating equations for logit models are, as iterative methods, computationally more demanding. For logit regression with

continuous predictors, which are not categorized, the PML and GEE methods can be used but the GWLS method is inappropriate. The application area of the GWLS method is thus more limited than that of PML and GEE methods.

In surveys with large samples, closely related results are usually attained by any of the methods. But in fitting ANOVA-type models there can be many multi-class predictors included in the model and, therefore, the number of domains can be large, and a large element-level sample size is required to obtain a reasonably large number of observations falling in each domain. This is especially important for the GWLS method, which is mainly used in large-scale surveys where the sample sizes can be in thousands of persons, as is the case in the OHC and MFH Surveys. For proper behaviour of GWLS, PML and GEE methods, a large number of sample clusters is beneficial. Recall that this property holds for the OHC Survey.

We consider the GWLS method for a binary response variable and a set of categorical predictors. The data can thus be arranged into a multidimensional table, such as Table 8.1, where the u domains are formed by cross-classifying the categorical predictors and the proportions p_j of the binary response are estimated in each domain. The consistent estimates $\hat{p}_j$, used under the design-based and weighted SRS options, are weighted ratio-type estimators of the form $\hat{p}_j = \hat{n}_{j1}/\hat{n}_j$, where $\hat{n}_{j1}$ is the weighted sample sum of the binary response in domain j, and $\hat{n}_j$ are weighted domain sample sizes. The unweighted proportion estimates $\hat{p}_j^U$, used under the unweighted SRS option, are obtained using the unweighted counterparts n_{j1} and n_j.

When applying the GWLS method for logit and linear modelling under an analysis option, the starting point is the calculation of the corresponding proportion estimate vector and its covariance-matrix estimate. By using these estimates, the model coefficients are estimated, together with a covariance matrix of the estimated coefficients, and using these, fitted proportions and their covariance-matrix estimates are obtained. Further, the Wald test of goodness of fit of the model, and desired Wald tests of linear hypotheses on the model coefficients, are executed. Finally, residual analysis is carried out to more closely examine the fit of the selected model.

Design-based GWLS Estimation

Under the design-based option, a consistent *GWLS estimator* $\hat{\mathbf{b}}_{des}$, denoted $\hat{\mathbf{b}}$ for short in this section, of the $s \times 1$ model coefficient vector $\mathbf{b}$ for a model $F(\mathbf{p}) = \mathbf{Xb}$ is given by

$$\hat{\mathbf{b}} = (\mathbf{X}'(\mathbf{H}\hat{\mathbf{V}}_{des}\mathbf{H})^{-1}\mathbf{X})^{-1}\mathbf{X}'(\mathbf{H}\hat{\mathbf{V}}_{des}\mathbf{H})^{-1}F(\hat{\mathbf{p}}), \tag{8.5}$$

where $\hat{\mathbf{V}}_{des}$ is a consistent estimator of the covariance matrix of the consistent domain proportion estimator vector $\hat{\mathbf{p}}$, and $\mathbf{H}\hat{\mathbf{V}}_{des}\mathbf{H}$ is a covariance-matrix estimator of the function vector $F(\hat{\mathbf{p}})$. An estimate $\hat{\mathbf{V}}_{des}$ is obtained using, for example, the linearization method as described in Chapter 5. The GWLS estimating

equations (8.5) are thus based on the consistently estimated functions $F(\hat{p}_j)$ and their design-based covariance-matrix estimate. The equations also indicate that no iterations are needed to obtain the estimates $\hat{b}_k$. A justification for the label 'GWLS' is that element weights are used in obtaining the proportion vector estimate and its covariance-matrix estimate, which are supplied to the GLS estimating equations.

The GWLS estimator $\hat{\mathbf{b}}$ from (8.5) applies for both logit and linear models on domain proportions. But the matrix $\mathbf{H}$ in the covariance-matrix estimator of the function vector differs. In the logit model, the diagonal $u \times u$ matrix $\mathbf{H}$ of partial derivatives of the functions $F(\hat{p}_j)$ has diagonal elements of the form $h_j = 1/(\hat{p}_j(1 - \hat{p}_j))$. And in the linear model, the matrix $\mathbf{H}$ is an identity matrix with ones on the main diagonal and zeros elsewhere.

Under a partial parametrization of a logit ANOVA model (see Section 8.2), where the columns of the model matrix $\mathbf{X}$ corresponding to the classes of the predictors are binary variables, a log odds ratio interpretation can be given to the estimates $\hat{b}_k$. Thus, an estimate $\exp(\hat{b}_k)$ is the *odds ratio* for the corresponding class with respect to the reference class adjusted for the effects of the other terms in the model. This interpretation of the estimated model coefficients is common in epidemiology and also in social sciences.

A covariance-matrix estimate $\hat{\mathbf{V}}_{des}(\hat{\mathbf{b}})$ of the estimated model coefficients $\hat{b}_k$ from (8.5) is used in obtaining Wald test statistics for the coefficients. This $s \times s$ covariance matrix is given by

$$\hat{\mathbf{V}}_{des}(\hat{\mathbf{b}}) = (\mathbf{X}'(\mathbf{H}\hat{\mathbf{V}}_{des}\mathbf{H})^{-1}\mathbf{X})^{-1}. \tag{8.6}$$

With proper choice of $\mathbf{H}$, this estimator applies again for both logit and linear models. Diagonal elements of $\hat{\mathbf{V}}_{des}(\hat{\mathbf{b}})$ provide the design-based variance estimates $\hat{v}_{des}(\hat{b}_k)$ of the estimated coefficients $\hat{b}_k$ to be used in obtaining the corresponding standard-error estimates $\text{s.e}_{des}(\hat{b}_k) = \hat{v}_{des}^{1/2}(\hat{b}_k)$. Under a logit model, using these standard-error estimates, for example, an approximative 95% confidence interval for an odds ratio $\exp(b_k)$ can be calculated as follows:

$$\exp(\hat{b}_k \pm 1.96 \times \text{s.e}_{des}(\hat{b}_k)). \tag{8.7}$$

Two additional covariance-matrix estimators are useful in practice. These are the $u \times u$ covariance-matrix estimator $\hat{\mathbf{V}}_{des}(\hat{\mathbf{F}})$ of the vector $\hat{\mathbf{F}} = \mathbf{X}\hat{\mathbf{b}}$ of the fitted logits and the covariance-matrix estimator $\hat{\mathbf{V}}_{des}(\hat{\mathbf{f}})$ of the vector $\hat{\mathbf{f}} = F^{-1}(\mathbf{X}\hat{\mathbf{b}})$ of the fitted proportions. These are

$$\hat{\mathbf{V}}_{des}(\hat{\mathbf{F}}) = \mathbf{X}\hat{\mathbf{V}}_{des}(\hat{\mathbf{b}})\mathbf{X}' \tag{8.8}$$

and

$$\hat{\mathbf{V}}_{des}(\hat{\mathbf{f}}) = \hat{\mathbf{H}}^{-1}\hat{\mathbf{V}}_{des}(\hat{\mathbf{F}})\hat{\mathbf{H}}^{-1}. \tag{8.9}$$

For a linear model, these covariance matrices obviously coincide, because the fitted functions are equal to the fitted proportions. For a logit model, the diagonal matrix $\hat{\mathbf{H}}$ has diagonal elements of the form $\hat{h}_j = 1/(\hat{f}_j(1-\hat{f}_j))$, and the terms $\hat{f}_j = f_j(\hat{\mathbf{b}})$ are elements of the vector $\hat{\mathbf{f}}$ of fitted proportions calculated using the equation

$$\hat{\mathbf{f}} = \mathbf{f}(\hat{\mathbf{b}}) = \exp(\mathbf{X}\hat{\mathbf{b}})/(1+\exp(\mathbf{X}\hat{\mathbf{b}})). \tag{8.10}$$

The diagonal elements of the covariance-matrix estimates (8.8) and (8.9) are needed to obtain the design-based standard errors of the fitted functions and of the fitted proportions.

Goodness of Fit and Related Tests

Examining goodness of fit of the model is an essential part of a logit and linear modelling procedure on domain proportions. Various goodness-of-fit statistics can be obtained by first partitioning the total variation *(total chi-square)* in the table into the variation due to the model *(model chi-square)* and into the residual variation *(residual chi-square)*. Hence, we have

$$\textit{total chi-square} = \textit{model chi-square} + \textit{residual chi-square}$$

similar to the partition of the total sum of squares for usual linear regression and ANOVA. A design-based Wald test statistic X^2_{des} measuring the residual variation is commonly used as an indicator of goodness of fit of the model. This statistic is given by

$$X^2_{des} = (F(\hat{\mathbf{p}}) - \mathbf{X}\hat{\mathbf{b}})'(\mathbf{H}\hat{\mathbf{V}}_{des}\mathbf{H})^{-1}(F(\hat{\mathbf{p}}) - \mathbf{X}\hat{\mathbf{b}}), \tag{8.11}$$

which is asymptotically chi-squared with $u-s$ degrees of freedom under the design-based option. A small value of this statistic, relative to the residual degrees of freedom, indicates good fit of the model, and obviously, the fit is perfect for a saturated model. A Wald statistic denoted by $X^2_{des}(overall)$, measuring the variation due to the overall model, is used to test the hypothesis that all the model coefficients are zero. It is given by

$$X^2_{des}(overall) = F(\hat{\mathbf{p}})'(\mathbf{H}\hat{\mathbf{V}}_{des}\mathbf{H})^{-1}F(\hat{\mathbf{p}}) - X^2_{des}, \tag{8.12}$$

where the first quadratic form measures the total variation and the second is the residual chi-square (8.11) for the model under consideration. This statistic is asymptotically chi-squared with s degrees of freedom. Also, a Wald statistic denoted by $X^2_{des}(gof)$ can be constructed for the hypothesis that all the model parameters, except the intercept, are zero. This statistic is defined as the difference of the observed values of the residual chi-square statistic (8.11) for the model where only the intercept is included and for the model including all the terms of the

current model, and therefore, it is asymptotically chi-squared with $s-1$ degrees of freedom. The statistic $X^2_{des}(overall)$ is sometimes called a test for the overall model, and $X^2_{des}(gof)$ a test of goodness of fit. Note that all these test statistics apply for both logit and linear models on domain proportions.

Linear hypotheses H_0: $\mathbf{Cb} = \mathbf{0}$ on the model coefficient vector $\mathbf{b}$ can be tested using the Wald statistic

$$X^2_{des}(\mathbf{b}) = (\mathbf{C}\hat{\mathbf{b}})'(\mathbf{C}\hat{\mathbf{V}}_{des}(\hat{\mathbf{b}})\mathbf{C}')^{-1}(\mathbf{C}\hat{\mathbf{b}}), \tag{8.13}$$

where $\mathbf{C}$ is the desired $c \times s$ ($c \leq s$) matrix of contrasts. The statistic is asymptotically chi-squared with c degrees of freedom under the design-based option. This statistic is used, for example, in the testing of hypotheses $H_0 : b_k = 0$ on single parameters of the model using the Wald statistics

$$X^2_{des}(b_k) = \hat{b}^2_k / \hat{v}_{des}(\hat{b}_k), \qquad k = 1, \ldots, s,$$

which are asymptotically chi-squared with one degree of freedom. Note that for the corresponding *t-test statistic* the equation $t^2_{des}(b_k) = X^2_{des}(b_k)$ holds.

Another asymptotically valid testing procedure for linear hypotheses on model parameters is based on a second-order Rao–Scott adjustment to a binomial-based Wald test statistic using the Satterthwaite method. This technique is similar to that used in Chapter 7 on the Pearson and Neyman test statistics. We first calculate the GWLS estimate $\hat{\mathbf{b}} = \hat{\mathbf{b}}_{bin}$ by using in (8.5) the binomial covariance-matrix estimate $\hat{\mathbf{V}}_{bin}$ of $\hat{\mathbf{p}}$ in place of $\hat{\mathbf{V}}_{des}$, and construct the corresponding Wald test statistic $X^2_{bin}(\mathbf{b})$:

$$X^2_{bin}(\mathbf{b}) = (\mathbf{C}\hat{\mathbf{b}})'(\mathbf{C}\hat{\mathbf{V}}_{bin}(\hat{\mathbf{b}})\mathbf{C}')^{-1}(\mathbf{C}\hat{\mathbf{b}}),$$

where $\hat{\mathbf{V}}_{bin}(\hat{\mathbf{b}})$ is the covariance-matrix estimate of the binomial GWLS estimates obtained by using the estimate $\hat{\mathbf{V}}_{bin}$ in place of $\hat{\mathbf{V}}_{des}$ in (8.6). The second-order corrected Wald statistic is given by

$$X^2_{bin}(\mathbf{b}; \hat{\delta}_\bullet, \hat{a}^2) = \frac{X^2_{bin}(\mathbf{b})}{\hat{\delta}_\bullet(1+\hat{a}^2)}, \tag{8.14}$$

where the first-order and second-order adjustment factors $\hat{\delta}_\bullet$ and $(1+\hat{a}^2)$ are calculated from the $c \times c$ generalized design-effects matrix estimate

$$\hat{\mathbf{D}} = (\mathbf{C}\hat{\mathbf{V}}_{bin}(\hat{\mathbf{b}})\mathbf{C}')^{-1}(\mathbf{C}\hat{\mathbf{V}}_{des}(\hat{\mathbf{b}})\mathbf{C}') \tag{8.15}$$

so that

$$\hat{\delta}_\bullet = \mathrm{tr}(\hat{\mathbf{D}})/c$$

is the mean of the eigenvalues $\hat{\delta}_k$ of the generalized design-effects matrix estimate, and

$$(1 + \hat{a}^2) = \sum_{k=1}^{c} \hat{\delta}_k^2 / (c\hat{\delta}_\bullet^2),$$

where the sum of squared eigenvalues is calculated by the formula

$$\sum_{k=1}^{c} \hat{\delta}_k^2 = \text{tr}(\hat{\mathbf{D}}^2).$$

The second-order adjusted statistic $X^2_{bin}(\mathbf{b}; \hat{\delta}_\bullet, \hat{a}^2)$ is asymptotically chi-squared under the design-based option with Satterthwaite adjusted degrees of freedom $\text{df}_S = c/(1 + \hat{a}^2)$. If $c = 1$, as in tests on separate parameters of a model, we have $(1 + \hat{a}^2) = 1$ because the generalized design-effects matrix reduces to a scalar and the adjustment reduces to a first-order adjustment. The test statistics are available in software products for the analysis of complex surveys.

Unstable Situations

Because the Wald statistics X^2_{des}, $X^2_{des}(overall)$ and $X^2_{des}(gof)$ of goodness of fit, and the statistic $X^2_{des}(\mathbf{b})$ of linear hypotheses on model parameters, are asymptotically chi-squared under the design-based option, they can be expected to work reasonably well if the number m of sample clusters is large relative to the number u of domains. But the test statistics can become overly liberal relative to the nominal significance levels if the covariance-matrix estimate $\hat{\mathbf{V}}_{des}$ appears unstable. This can happen if the degrees of freedom $f = m - H$ are small for an estimate $\hat{\mathbf{V}}_{des}$, relative to the residual or model degrees of freedom.

There are certain F-corrected Wald test statistics available to protect against the effects of instability similar to those used in Chapter 7 for hypotheses of homogeneity and independence. For the goodness-of-fit test statistic (8.11), these degrees-of-freedom corrections are

$$F_{1.des} = \frac{f - (u - s) + 1}{f(u - s)} X^2_{des}, \tag{8.16}$$

referred to the F-distribution with $(u - s)$ and $(f - (u - s) + 1)$ degrees of freedom, and

$$F_{2.des} = X^2_{des}/(u - s), \tag{8.17}$$

referred in turn to the F-distribution with $(u - s)$ and f degrees of freedom. These F-corrections can also be derived for the Wald statistics $X^2_{des}(overall)$ and $X^2_{des}(gof)$, using the corresponding degrees of freedom s or $(s - 1)$ in place of $(u - s)$.

Similar F-corrections can be derived for the Wald test statistics of linear hypotheses on model parameters. For the statistic (8.13), these are

$$F_{1.des}(\mathbf{b}) = \frac{f-c+1}{fc} X^2_{des}(\mathbf{b}) \tag{8.18}$$

and

$$F_{2.des}(\mathbf{b}) = X^2_{des}(\mathbf{b})/c, \tag{8.19}$$

referred to the F-distributions with c and $(f-c+1)$, and c and f degrees of freedom, respectively.

Second-order Rao–Scott adjustments can be expected to be robust to instability problems. However, for the second-order corrected statistic (8.14), an F-correction can be derived. It is given by

$$F_{bin}(\mathbf{b}; \hat{\delta}_{\bullet}, \hat{a}^2) = (1+\hat{a}^2) X^2_{bin}(\mathbf{b}; \hat{\delta}_{\bullet}, \hat{a}^2)/c = X^2_{bin}(\mathbf{b})/(c\hat{\delta}_{\bullet}), \tag{8.20}$$

which is referred to the F-distribution with df_S and f degrees of freedom.

The impact of these F-corrections on p-values of the tests is small if f is large. However, if f is relatively small, and especially if f and the residual degrees of freedom are close, the corrections can be effective. Under serious instability, the statistics $F_{1.des}$, and $F_{1.des}(\mathbf{b})$ or $F_{bin}(\mathbf{b}; \hat{\delta}_{\bullet}, \hat{a}^2)$, are preferable. These corrections have been implemented as testing options in software products for the analysis of complex surveys.

Residual Analysis

It is desirable to examine more closely the fit of the selected model by calculating the raw and standardized residuals. These can be used in detecting possible outlying domain proportions. The raw residuals are simple differences $(\hat{p}_j - \hat{f}_j)$ of the fitted proportions $\hat{f}_j$ from the corresponding observed proportions $\hat{p}_j$. Under the design-based option, the standardized residuals are calculated by first obtaining a covariance-matrix estimate $\hat{\mathbf{V}}_{res}$ of the raw residuals given by

$$\hat{\mathbf{V}}_{res} = \mathbf{H}^{-1}(\mathbf{H}\hat{\mathbf{V}}_{des}\mathbf{H} - \hat{\mathbf{V}}_{des}(\hat{\mathbf{F}}))\mathbf{H}^{-1}, \tag{8.21}$$

where $\mathbf{H}\hat{\mathbf{V}}_{des}\mathbf{H}$ and $\hat{\mathbf{V}}_{des}(\hat{\mathbf{F}})$ are the design-based covariance-matrix estimates of the vector $F(\hat{\mathbf{p}})$ of the observed functions and the vector $\hat{\mathbf{F}} = \mathbf{X}\hat{\mathbf{b}}$ of the fitted functions, respectively, and the matrix $\mathbf{H}$ depends on which model type, logit or linear, is fitted. Using (8.21), the standardized residuals are calculated as

$$\hat{e}_j = (\hat{p}_j - \hat{f}_j)/\sqrt{\hat{v}_j}, \quad j = 1, \ldots, u, \tag{8.22}$$

where $\hat{v}_j$ are the diagonal elements of the residual covariance matrix $\hat{\mathbf{V}}_{res}$. A large standardized residual indicates that the corresponding domain is poorly accounted for by the model. Because the standardized residuals are approximate standard normal variates, they can be referred to critical values from the $N(0,1)$ distribution.

Design Effect Estimation

A principal property of the GWLS method is its flexibility, not only for various model formulations but also for alternative sampling designs. The design-based GWLS method appeared valid under the design-based option involving a complex multi-stage design with clustering and stratification. But the GWLS method can also be used for simpler designs with the choice of an appropriate proportion estimator and its covariance-matrix estimator reflecting the complexities of the sampling design.

Under the weighted SRS option, the consistent proportion estimate $\hat{\mathbf{p}}$ and its binomial covariance-matrix estimate $\hat{\mathbf{V}}_{bin}(\hat{\mathbf{p}})$ are used in equations (8.5) and (8.6) to obtain the corresponding GWLS estimate $\hat{\mathbf{b}}$ of model coefficients and the covariance-matrix estimate $\hat{\mathbf{V}}_{bin}(\hat{\mathbf{b}})$. The same holds for the unweighted SRS option, where the unweighted counterparts $\hat{\mathbf{p}}^U$ and $\hat{\mathbf{V}}_{bin}(\hat{\mathbf{p}}^U)$ are used. The GWLS estimating equations indicate that the estimates $\hat{b}_k$ obtained under the SRS-based options would not numerically coincide with those from the design-based option.

The SRS-based options are restrictive in the sense that the effect of clustering on standard-error estimates of estimated model coefficients cannot be accounted for. This effect is indicated in design-effect estimates of model coefficient estimates. The design-effect estimates are calculated by using the diagonal elements of the covariance-matrix estimates $\hat{\mathbf{V}}_{des}(\hat{\mathbf{b}})$ and $\hat{\mathbf{V}}_{bin}(\hat{\mathbf{b}}^*)$ of the model coefficients. Hence, we have

$$\hat{d}(\hat{b}_k) = \hat{v}_{des}(\hat{b}_k)/\hat{v}_{bin}(\hat{b}_k^*), \quad k = 1, \ldots, s, \tag{8.23}$$

where $\hat{b}_k^*$ denotes the estimated model coefficients obtained under the weighted or unweighted SRS option. Under the unweighted SRS option, these design-effect estimates indicate the contribution of all the sampling complexities, and under the weighted SRS option, the contribution of clustering is indicated. It is often instructive to calculate the design-effect estimates under both SRS options, because then the contribution of the weighting to design effects can be examined.

Criteria for Choosing a Model Formulation

Which one of the model formulations for proportions, logit or linear, should be chosen? In certain sciences, one type is more standard than the other, but

taking an explicit position in favour of either of the types generally is not possible. It appears that there are gains with the logit formulation, such as possibilities for interpretation with odds ratios, and in certain cases with standard independence concepts. Moreover, being a member of the broad category of so-called exponential family models, a logit model for binomial proportions involves convenient statistical properties that are not shared with linear models for binomial proportions. Although these properties do not necessarily apply to logit models in complex surveys, attention has also been directed to the use of logit models for this kind of survey.

The linear model formulation on proportions, on the other hand, provides a simple modelling approach that is especially convenient for those familiar with linear ANOVA on continuous measurements. Being additive on a linear scale, the coefficients of a linear model describe differences of the proportions themselves, not their logits. In practice, however, logit and linear GWLS estimation results on model coefficients do not markedly differ if proportions are in the range 0.2–0.8, say. In the following example, we compare the logit and linear model formulations in a typical health sciences analysis.

Example 8.1

Logit and linear ANOVA with the GWLS method. Let us apply the GWLS method for logit and linear modelling on domain proportions in the simple OHC Survey setting displayed in Table 8.1. Our aim is to model the variation of domain proportions of the binary response variable PSYCH, measuring overall psychic strain, across the $u = 8$ domains formed by sex and age of respondent, and the variable PHYS describing the respondent's physical working conditions. Table 8.2 provides a more complete description of the analysis situation. The original domain sample sizes $\hat{n}_j$ and the number m_j of sample clusters covered by each domain are included in addition to the domain proportions $\hat{p}_j$, standard errors s.e_j and design effects $\hat{d}_j$. Note that the domain proportions vary around the value 0.5.

The design-based option provides valid GWLS logit and linear modelling in this analysis. The sampling design involves clustering effects, as indicated by design-effect estimates of proportions being on average greater than one. The average design-effect estimate is 1.28. Further, the domains constitute cross-classes, which is indicated by the fact that each domain covers a reasonably large number of sample clusters. More apparently, this property can be seen from the design-based covariance-matrix estimate $\hat{\mathbf{V}}_{des}$ of domain proportions displayed in Figure 8.1. It can be noted that there exist nonzero covariance terms in the off-diagonal part of the covariance-matrix estimate. The estimate also seems relatively stable, because covariance estimates are much smaller than the corresponding variance estimates. The condition number of $\hat{\mathbf{V}}_{des}$ is 12.1, which also indicates stability. The corresponding binomial covariance-matrix estimate $\hat{\mathbf{V}}_{bin}$ is displayed for comparison.

Table 8.2 Proportion $\hat{p}_j$ of persons in the upper psychic strain group, with standard error estimates s.e_j and design-effect estimates $\hat{d}_j$ of the proportions, and domain sample sizes $\hat{n}_j$ and the number of sample clusters m_j (the OHC Survey).

Domain j	SEX	AGE	PHYS	$\hat{p}_j$	s.e_j	$\hat{d}_j$	$\hat{n}_j$	m_j
1	Males	−44	0	0.419	0.0128	1.16	1734	230
2			1	0.472	0.0145	1.33	1578	198
3		45–	0	0.461	0.0178	0.88	690	186
4			1	0.520	0.0247	1.18	483	138
5	Females	−44	0	0.541	0.0125	1.23	1966	240
6			1	0.620	0.0270	1.38	447	152
7		45–	0	0.532	0.0236	1.65	740	185
8			1	0.700	0.0391	1.48	203	101
All				0.500	0.0073	1.69	7841	250

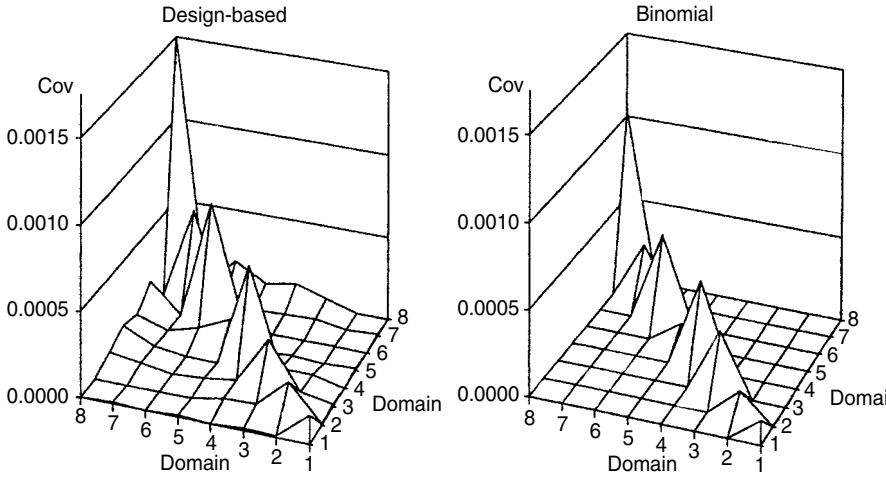


Figure 8.1 Design-based and binomial covariance-matrix estimates $\hat{\mathbf{V}}_{des}$ and $\hat{\mathbf{V}}_{bin}$ of domain proportion estimates $\hat{p}_j$.

We consider the model-building process under the design-based option, and use the unweighted SRS option as a reference. There are three predictors, and together with their main effects, an intercept, and four interaction terms, a total of eight model terms appear in the saturated logit and linear ANOVA models, which can be written in the form

$$\text{F(P)} = \text{INTERCEPT} + \text{SEX} + \text{AGE} + \text{PHYS} + \text{SEX} * \text{AGE}$$
$$+ \text{SEX} * \text{PHYS} + \text{AGE} * \text{PHYS} + \text{SEX} * \text{AGE} * \text{PHYS},$$

where the function is $F(P) = \log(P/(1-P))$ for the logit model and $F(P) = P$ for the linear model, and P stands for proportions of the upper PSYCH group.

In the model-building process, we first fit the saturated logit and linear models and test the significance of the interaction term of all the three predictors. If it appears nonsignificant, we remove the term, and study the two-variable interactions, in turn, for further reduction of the model. Model building is completed when a reasonably well-fitting reduced model is attained. This stepwise process is an example of the so-called *backward elimination* common in fitting of log-linear and logit ANOVA models.

Let us consider more closely the results on logit model fitting. Under the design-based option, the main effects model appeared reasonably well-fitting and could not be further reduced. Results for the model reduction are given in Table 8.3. There, the values of X^2_{des} for a difference Wald statistic are obtained, for example, in the comparison of the saturated model 5 and the model 4. The difference statistic is calculated as $X^2_{des}(overall; 5) - X^2_{des}(overall; 4) = 78.84 - 76.90 = 1.94$, and compared to the chi-squared distribution with one degree of freedom attains a nonsignificant p-value 0.1635, and thus, the interaction term can be removed from the model 5. The observed value of the Wald statistic of goodness of fit of the main effects model (Model 1) is $X^2_{des} = 78.84 - 72.39 = 6.45$, which with 4 degrees of freedom attains a p-value 0.1681, indicating reasonably good fit.

Substantial reduction of the saturated logit model was possible, and the model-building procedure produced quite a simple structure including the main effects terms only. So, the suspected interaction of SEX and PHYS appeared nonsignificant. We return to this conclusion later when fitting logit models under the SRS-based analysis options.

Table 8.3 Observed values of the Wald statistics X^2_{des} (*overall*) for overall models, and the differences statistics X^2_{des} when compared with reduced logit ANOVA models, under the design-based analysis option.

		Overall		Model		Difference	
Model	df	X^2_{des}	p-value	comparison	df	X^2_{des}	p-value
5	8	78.84	0.0000	—	1	—	—
4	7	76.90	0.0000	5–4	1	1.94	0.1635
3	6	76.09	0.0000	4–3	1	0.81	0.3693
2	5	74.78	0.0000	3–2	1	1.31	0.2533
1	4	72.39	0.0000	2–1	1	2.39	0.1218

Model 5: SEX + AGE + PHYS + SEX*AGE + SEX*PHYS + AGE*PHYS + SEX*AGE*PHYS
Model 4: SEX + AGE + PHYS + SEX*AGE + SEX*PHYS + AGE*PHYS
Model 3: SEX + AGE + PHYS + SEX*PHYS + AGE*PHYS
Model 2: SEX + AGE + PHYS + SEX*PHYS
Model 1: SEX + AGE + PHYS

In the partial parametrization used here, for each predictor the model coefficient for the first class is set to zero. The first class of the last domain is the reference domain—here domain 7 in Table 8.2. There are four coefficients b_k to be estimated in the main effects models. GWLS estimates $\hat{b}_k$ are actually obtained under the following model matrix:

$$\mathbf{X} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{bmatrix}.$$

The fitted models can be written with $\hat{b}_k$ and the model matrix as

$$F(\hat{f}_j) = \hat{b}_1 + \hat{b}_2(\text{SEX})_j + \hat{b}_3(\text{AGE})_j + \hat{b}_4(\text{PHYS})_j, \quad j = 1, \ldots, 8,$$

where $F(\hat{f}_j) = \log(\hat{f}_j/(1-\hat{f}_j))$ for the logit model, and $F(\hat{f}_j) = \hat{f}_j$ for the linear model, and the indicator variable values for SEX, AGE and PHYS are in the second, third and fourth columns of the model matrix **X**.

Let us consider more closely the estimation and test results for the main effects logit model. The estimation results for the model coefficients are displayed in Table 8.4.

Table 8.4 Estimates from design-based logit ANOVA on overall psychic strain (model fitting by the GWLS method).

Model term	Beta coefficient	Design effect	Standard error	*t*-test	*p*-value	Odds ratio	95% confidence interval for OR	
							Lower	Upper
Intercept	−0.3282	1.32	0.0635	−7.02	0.0000	0.72	0.66	0.79
Sex								
Males*	0	n.a.	0	n.a.	n.a.	1	1	1
Females	0.4663	1.44	0.0579	8.06	0.0000	1.59	1.42	1.79
Age								
−44*	0	n.a.	0	n.a.	n.a.	1	1	1
45–	0.1385	1.23	0.0570	2.43	0.0159	1.15	1.03	1.28
Physical health hazards								
No*	0	n.a.	0	n.a.	n.a.	1	1	1
Yes	0.2568	1.30	0.0574	4.48	0.0000	1.29	1.16	1.45

* Reference class; parameter value set to zero.
n.a. not available.

In the table, a positive value of the estimated coefficients $\hat{b}_2$ and $\hat{b}_3$ for females and for the older group is obtained as expected, and the corresponding t-tests attain significant p-values. The sex−age adjusted estimate $\hat{b}_4$ for the PHYS class of more hazardous work is positive, involving a clearly significant t-test. It should be noticed that the absolute value of the t-test statistic used here corresponds to the square root of the F-corrected Wald statistic (8.19). The design-effect estimates $\hat{d}(\hat{b}_k)$ of the estimated model coefficients are larger than one owing to the clustering effect. Thus, binomial standard-error estimates of the model coefficients would be smaller than the corresponding design-based estimates.

Using the estimate $\hat{b}_4 = 0.2568$ for the interesting parameter of the PHYS class of more hazardous work, the corresponding sex−age adjusted odds ratio estimate with its 95% confidence interval can be obtained by (8.7). The odds ratio (OR) estimate is $\exp(\hat{b}_4) = 1.29$, and its 95% confidence interval is calculated as

$$\exp(0.2568 \pm 1.96 \times 0.0574) = (1.16, 1.45).$$

The sex−age adjusted odds of experiencing a higher level of psychic strain is thus 1.3 times higher for persons under more hazardous working conditions than for those in the group of less hazardous work. This result is consistent with the t-test results, because the 95% confidence interval does not include the value one, which is the odds ratio for the reference group.

We next turn to the test results on the model terms in the final main effects ANOVA model (Table 8.5). There is a set of observed values from different Wald test statistics and their F-corrections. Let us consider more closely the tests for the model terms. The first test statistic corresponds to the original design-based Wald statistic (8.13), and the second statistic is the F-corrected statistic (8.18). The third statistic is the Satterthwaite corrected binomial statistic (8.14), and finally, the fourth statistic is the F-corrected statistic (8.20). The design-based Wald statistic $X^2_{des}(\mathbf{b})$ and the second-order corrected binomial statistic $X^2_{bin}(\mathbf{b}; \hat{\delta}_\bullet, \hat{a}^2)$ provide similar results. The design-based Wald statistic thus works adequately in this

Table 8.5 Observed values and p-values of test statistics for model terms in the final logit ANOVA model on overall psychic strain (model fitting by the GWLS method).

Contrast	Df	(1) Design-based Wald test	p-value	(2) F-correction to (1)	p-value	(3) Rao−Scott 2nd order adjustment to binomial Wald test	p-value	(4) F-correction to (3)	p-value
SEX	1	64.92	0.0000	64.92	0.0000	64.92	0.0000	64.92	0.0000
AGE	1	5.90	0.0151	5.90	0.0159	5.90	0.0153	5.90	0.0159
PHYS	1	20.04	0.0000	20.04	0.0000	20.04	0.0000	20.04	0.0000

(1) Equation (8.13), (2) Equation (8.18), (3) Equation (8.14), (4) Equation (8.20)

case, which is primarily due to the stability of the covariance-matrix estimate $\hat{\mathbf{V}}_{des}(\hat{\mathbf{b}})$. Because there is a large number of degrees of freedom $f = 245$ for an estimate $\hat{\mathbf{V}}_{des}(\hat{\mathbf{b}})$, the F-corrected tests do not contribute substantially to the p-values of the original tests.

Although there is no controversy about the results from the alternative test statistics in this analysis situation, there can be situations where the choice of an adequate statistic is crucial. This is especially so if the number m of sample clusters is small and the number of domains u is close to m. Then, some of the F-corrected statistics can be chosen to protect against the effects of instability.

For a more detailed examination of the model fit, let us now calculate the fitted proportions and the raw and standardized residuals for a residual analysis. These are displayed in Table 8.6.

The observed and fitted proportions are close, except in the last three domains where the largest raw residuals can be obtained. The standardized residuals in the last two groups exceed the 5% critical value 1.96 from the $N(0,1)$ distribution; so the model fit is somewhat questionable for these domains. It should be noticed that the fitted proportions and the residuals are independent of the parametrization of the model.

It would be useful to consider briefly the logit analysis under the other analysis options as a reference to the results from the design-based option. In this, we are especially interested in the importance of the term SEX*PHYS, describing the interaction of SEX and PHYS, which appeared nonsignificant under the design-based option. The results from the Wald tests are in Table 8.7.

The interaction of SEX and PHYS appears significant when ignoring the clustering effect by using the unweighted SRS option. A more complex model is thus obtained than under the design-based option. These results suggest further warnings on ignoring the clustering effect even if it is not very serious as indicated in the medium-sized domain design-effect estimates.

Table 8.6 Observed and fitted PSYCH proportions $\hat{p}_j$ and $\hat{f}_j$ with their standard errors, and raw and standardized residuals $(\hat{p}_j - \hat{f}_j)$ and $\hat{e}_j$ for the logit ANOVA Model 1 under the design-based option.

Domain	SEX	AGE	PHYS	$\hat{p}_j$	s.e $(\hat{p}_j)$	$\hat{f}_j$	s.e $(\hat{f}_j)$	$(\hat{p}_j - \hat{f}_j)$	$\hat{e}_j$
1	Males	−44	0	0.419	0.0128	0.419	0.0114	0.0000	0.0000
2			1	0.472	0.0145	0.482	0.0122	−0.0100	−1.270
3		45−	0	0.461	0.0178	0.453	0.0142	0.0082	0.771
4			1	0.520	0.0247	0.517	0.0167	0.0029	0.160
5	Females	−44	0	0.541	0.0125	0.534	0.0115	0.0062	1.306
6			1	0.620	0.0270	0.597	0.0160	0.0222	2.012
7		45−	0	0.532	0.0236	0.569	0.0156	−0.0363	−2.073
8			1	0.700	0.0391	0.630	0.0199	0.0692	1.993

Table 8.7 Wald tests $X^2(\mathbf{b})$ for the significance of the interaction term SEX*PHYS in Model 2 under the design-based and unweighted SRS analysis options.

		Design-based		Unweighted SRS	
Term	df	X^2_{des}	p-value	X^2_{bin}	p-value
SEX*PHYS	1	2.39	0.1218	3.97	0.0463

Let us turn to the corresponding design-based analysis with a linear model for the proportions of Table 8.2. In this situation, logit and linear formulations of an ANOVA model lead to similar results because proportions do not deviate much from the value 0.5. The main effects model (Model 1) is chosen, and results on model fit, residuals, and on significance of the model terms, are close to those for the logit model. But the estimates of the model coefficients differ and are subject to different interpretations. For the logit model with the partial parametrization, an estimated coefficient indicates differential effect on a logit scale of the corresponding class from the estimated intercept being the fitted logit for the reference domain. And for the linear model, an estimated coefficient indicates differential effect on a linear scale of the corresponding class from the estimated intercept, which is now the fitted proportion for the reference domain.

The linear model formulation thus involves a more straightforward interpretation of the estimates of the model coefficients. Under Model 1, these estimates are as follows:

$$\begin{aligned}
\hat{b}_1 &= 0.5705 \quad \text{(Intercept)} \\
\hat{b}_2 &= -0.1172 \quad \text{(Differential effect of SEX} = \text{Males)} \\
\hat{b}_3 &= -0.0355 \quad \text{(Differential effect of AGE} = -44) \\
\hat{b}_4 &= 0.0650 \quad \text{(Differential effect of PHYS} = 1).
\end{aligned}$$

The fitted proportion for falling into the upper psychic strain group is thus 0.57 for females in the older age group whose working conditions are less hazardous, and for males in the same age group, $0.57 - 0.12 = 0.45$. The highest fitted proportion, $0.57 + 0.07 = 0.64$, is for the older age group females doing more hazardous work. Also, the fitted proportions are close to those obtained with the corresponding logit ANOVA model.

8.4 LOGISTIC AND LINEAR REGRESSION

The PML method of pseudolikelihood is often used on complex survey data for logit analysis in analysis situations similar to the GWLS method. But the applicability of the PML method is wider, covering not only models on domain proportions of

a binary or polytomous response but also the usual regression-type settings with continuous measurements as the predictors. We consider in this section first a PML analysis on domain proportions and then a more general situation of logit modelling of a binary response with a mixture of continuous measurements and categorical variables as predictors. Finally, an example is given of linear modelling for a continuous response variable in an ANCOVA setting.

In PML estimation of model coefficients and their asymptotic covariance matrix, we use a modification of the maximum likelihood (ML) method. In the ML estimation for simple random samples, we work with unweighted observations and appropriate likelihood equations can be constructed, based on standard distributional assumptions, to obtain the ML estimates of the model coefficients and the corresponding covariance-matrix estimate. Using these estimates, standard likelihood ratio (LR) and binomial-based Wald test statistics can be used for testing the model adequacy and linear hypotheses on the model coefficients.

Under more complex designs involving element weighting and clustering, an ML estimator of the model coefficients and the corresponding covariance-matrix estimator are not consistent and, moreover, the standard test statistics are not asymptotically chi-squared with appropriate degrees of freedom. For consistent estimation of model coefficients, the standard likelihood equations are modified to cover the case of weighted observations. In addition to this, a consistent covariance-matrix estimator of the PML estimators is constructed such that the clustering effects are properly accounted for. Using these consistent estimators, appropriate asymptotically chi-squared test statistics are derived.

The PML method can be conveniently introduced in a setting similar to the GWLS method, assuming again a binary response variable and a set of categorical predictors. The data set is arranged in a multidimensional table, such as Table 8.1, with u domains, and our aim is to model the variation of the domain proportion estimates $\hat{p}_j$ across the domains. The variation is modelled by a logit model of the type given in (8.1) and (8.2). A PML logit analysis for domain proportions, covering logit ANOVA, ANCOVA and regression models with categorical predictors can be carried out under any of the analysis options previously introduced by using the corresponding domain proportion estimator vector and its covariance-matrix estimate, and the steps in model-building are equivalent to those in the GWLS method. The design-based analysis option provides a generally valid PML logit analysis for complex surveys. In practice, a PML logit analysis under the design-based option requires access to specialized software for survey analysis.

Design-based and Binomial PML Methods

Under both design-based and weighted SRS options, a consistent *PML estimator* $\hat{\mathbf{b}}_{pml}$ for the vector $\mathbf{b}$ of the s model coefficients b_k in a logit model $F(\mathbf{p}) = \mathbf{Xb}$ is obtained by iteratively solving the PML estimating equations

$$\mathbf{X}'\mathbf{W}\mathbf{f}(\hat{\mathbf{b}}_{pml}) = \mathbf{X}'\mathbf{W}\hat{\mathbf{p}}, \tag{8.24}$$

where $\mathbf{W}$ is a $u \times u$ diagonal weight matrix with weights $w_j = \hat{n}_j$ on the main diagonal, and $\mathbf{f} = \exp(\mathbf{Xb})/(1 + \exp(\mathbf{Xb}))$ is the inverse function of the logit function. It is essential in (8.24) that the weighted domain sample sizes $\hat{n}_j$ and the weighted proportion estimates $\hat{p}_j$ be used, not their unweighted counterparts n_j and $\hat{p}_j^U$ as in the ML method, i.e. under the unweighted SRS option. This is for consistency of the PML estimators. The corresponding vector (8.5) of the GWLS estimates can be used as an initial value for the PML iterations. Note that under the linear formulation of the ANOVA model, the function vector $\mathbf{f}(\hat{\mathbf{b}}_{pml})$ would be linear in $\hat{b}_k$ and, thus, no iterations are needed. Henceforth, in this section we denote the vector of PML estimates of logit model coefficients by $\hat{\mathbf{b}}$ for short.

Because the vector $\hat{\mathbf{b}}$ of PML estimates is equal under the design-based and weighted SRS options, so also are the vectors $\hat{\mathbf{F}} = \mathbf{X}\hat{\mathbf{b}}$ and $\hat{\mathbf{f}} = F^{-1}(\mathbf{X}\hat{\mathbf{b}})$ of fitted logits and fitted proportions. The equality also holds for estimated odds ratios, which can be obtained as $\exp(\hat{b}_k)$ under the partial parametrization of the model. Fitted proportions $\hat{f}_j = f_j(\hat{\mathbf{b}})$ are estimated under both options by the formula

$$\hat{\mathbf{f}} = \mathbf{f}(\hat{\mathbf{b}}) = \exp(\mathbf{X}\hat{\mathbf{b}})/(1 + \exp(\mathbf{X}\hat{\mathbf{b}})). \tag{8.25}$$

Let us derive under the weighted SRS and design-based options the $s \times s$ covariance-matrix estimators of the PML estimator vector $\hat{\mathbf{b}}$ calculated by (8.24). Assuming simple random sampling, the covariance-matrix estimator is given by

$$\hat{\mathbf{V}}_{bin}(\hat{\mathbf{b}}) = (\mathbf{X}'\mathbf{W}\hat{\Delta}\mathbf{W}\mathbf{X})^{-1}, \tag{8.26}$$

where the diagonal elements of the diagonal $u \times u$ matrix $\hat{\Delta}$ are binomial-type variances $\hat{f}_j(1 - \hat{f}_j)/\hat{n}_j$. The binomial covariance-matrix estimator (8.26) is not consistent for complex sampling designs involving clustering. For these designs, we derive a more complicated consistent covariance-matrix estimator that is valid under the design-based option:

$$\hat{\mathbf{V}}_{des}(\hat{\mathbf{b}}) = \hat{\mathbf{V}}_{bin}(\hat{\mathbf{b}})\mathbf{X}'\mathbf{W}\hat{\mathbf{V}}_{des}\mathbf{W}\mathbf{X}\hat{\mathbf{V}}_{bin}(\hat{\mathbf{b}}). \tag{8.27}$$

This estimator is of a 'sandwich' form such that the design-based covariance-matrix estimator $\hat{\mathbf{V}}_{des}$ of the proportion vector $\hat{\mathbf{p}}$ acts as the 'filling'.

Approximate confidence intervals for odds ratio estimates $\exp(b_k)$ under the design-based and weighted SRS options can be calculated by (8.7) using the corresponding variance estimates $\hat{v}_{des}(\hat{b}_k)$ and $\hat{v}_{bin}(\hat{b}_k)$ of the PML estimates $\hat{b}_k$, as in the GWLS method. Also, the design-effect estimates $\hat{d}(\hat{b}_k)$ of the model coefficients $\hat{b}_k$ can be obtained by (8.23), again analogously to the GWLS method.

Expressions for the consistent covariance-matrix estimators $\hat{\mathbf{V}}_{des}(\hat{\mathbf{F}})$ and $\hat{\mathbf{V}}_{des}(\hat{\mathbf{f}})$ of the vector $\hat{\mathbf{F}}$ of fitted logits and the vector $\hat{\mathbf{f}}$ of fitted proportions are similar under the design-based option to those of the GWLS method, as given in equations (8.8) and (8.9). The PML analogue $\hat{\mathbf{V}}_{des}(\hat{\mathbf{b}})$ from (8.27) and the corresponding

matrix $\hat{\mathbf{H}}$ must of course be used in the equations. And under the weighted SRS option, the covariance-matrix estimators $\hat{\mathbf{V}}_{bin}(\hat{\mathbf{F}})$ and $\hat{\mathbf{V}}_{bin}(\hat{\mathbf{f}})$ are derived similarly by using the binomial estimator (8.26) in the equations in place of its design-based counterpart.

A residual covariance-matrix estimator is needed for conducting a proper residual analysis under the design-based option. This $u \times u$ estimator is given by

$$\hat{\mathbf{V}}_{res} = \mathbf{A}\hat{\mathbf{V}}_{des}\mathbf{A}', \tag{8.28}$$

where the matrix $\mathbf{A}$ is obtained by the formula

$$\mathbf{A} = \mathbf{I} - \hat{\Delta}\mathbf{W}\mathbf{X}(\mathbf{X}'\mathbf{W}\hat{\Delta}\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}$$

with $\mathbf{I}$ being a $u \times u$ identity matrix. Using this estimate, design-based standardized residuals of the form (8.22) can then be calculated.

There are thus many similarities between the PML formulae and those derived for the GWLS method. The main differences lie in the way the estimates of model coefficients and their covariance-matrix estimate are calculated. More similarities are evident in the testing procedures. All the test statistics derived for the GWLS method are also applicable to the PML method.

Under the design-based option, goodness of fit of the model can be tested with the design-based Wald statistic X^2_{des} given by (8.11). When examining the model fit more closely, PML analogues to the Wald statistics $X^2_{des}(overall)$ and $X^2_{des}(gof)$ can be used. The Wald statistics (8.13) and (8.14) for linear hypotheses on model parameters are applicable as well. Finally, in unstable situations, the F-corrected Wald and Rao–Scott statistics (8.16)–(8.20) can be used. It should be noted that the PML estimates from (8.24) and the corresponding covariance-matrix estimate (8.27) must be used in the calculation of these test statistics under the design-based option. These test statistics are available in commonly used software products for logit analysis for complex survey data.

In testing procedures for the weighted and unweighted SRS options, the corresponding binomial covariance-matrix estimates are used in the test statistics in place of those from the design-based option. As an alternative to the Wald statistics, LR test statistics can be used, which for the design-based option should be adjusted using the Rao–Scott methodology. A second-order adjustment to LR test statistics similar to (8.14) for the binomial-based Wald statistic provides asymptotically chi-squared test statistics. The residual covariance-matrix estimate (8.28) can be used in deriving an appropriate generalized design-effects matrix estimate for the adjustments.

The main application area of the PML method for complex surveys is under the design-based option, and the weighted and unweighted SRS options are used as the reference when examining the effects of weighting and intra-cluster correlation on standard-error estimates of model coefficients and on p-values of Wald test statistics.

Logistic Regression

The PML method can also be used in strictly regression-type logit analyses on a binary response variable from a complex survey, where the predictors are continuous measurements. In logistic regression, we work with an element-level data set without aggregating these data into a multidimensional table. So, the measured values of the continuous predictor variables constitute the columns in an $n \times s$ model matrix $\mathbf{X}$ for a logistic regression model. But all the other elements of the PML estimation remain unchanged, and consistent PML estimates with their consistent covariance-matrix estimate are obtained in a way similar to that described for the design-based analysis option. Moreover, a logistic ANCOVA can be performed by incorporating categorical predictors into the logistic regression model. Then, interaction terms of the continuous and categorical predictors can also be included.

A logistic regression model is usually built by entering predictors into the model using subject-matter criteria or significance measures of potential predictors. In this, t-tests $\mathrm{t}_{des}(b_k)$, or the corresponding Wald tests $X^2_{des}(b_k)$, on model coefficients can be used as previously and, under the design-based option, asymptotic properties of these test statistics remain unchanged.

Instability of an estimate $\hat{\mathbf{V}}_{des}(\hat{\mathbf{b}})$ from (8.27) can destroy the distributional properties of the test statistics on model coefficients in such small-sample situations where the number of sample clusters is small. Usual degrees-of-freedom, F-corrections to the Wald and t-test statistics can then be used.

The GEE methodology of generalized estimating equations can also be used for logistic modelling on complex survey data. In this method, the model coefficients are estimated using the multivariate quasilikelihood technique, and intra-cluster correlations are taken as nuisances. Using an estimated intra-cluster correlation structure, a 'robust' estimator of the covariance matrix of the model coefficients can be obtained, basically similar to the 'sandwich' form in the PML method. Thus, the GEE method can be used to account for the clustering effects. We describe only briefly the method and give an example for logistic ANCOVA in the OHC Survey.

The GEE method was originally developed for accounting for the possible correlation of observations in fitting generalized linear models in the context of longitudinal surveys (Liang and Zeger 1986). The methodology has been further described and illustrated in Liang *et al.* (1992) and Diggle *et al.* (2002).

Two alternatives of the GEE method have been presented. A preliminary GEE method with an independent correlation assumption relates to the standard PML method where observations are assumed independent within clusters for the estimation of the regression coefficients, but are allowed to be correlated for the estimation of the covariance matrix of the estimated regression coefficients. In covariance-matrix estimation, a 'sandwich' form of estimator is used. In a more advanced GEE method, assuming an exchangeable correlation structure, observations are allowed to be correlated within clusters in the estimation of both

regression coefficients and the covariance matrix of estimated regression coefficients. There, a 'working' intra-cluster correlation is estimated and incorporated in the estimation procedure of regression coefficients and the covariance matrix of estimated coefficients.

A generalized linear model can be compactly written as

$$\mathrm{E_M}(\mathrm{g}(\mathbf{y})) = \mathbf{Xb}, \tag{8.29}$$

where $\mathrm{E_M}$ refers to the expectation under the model and the function g refers to the so-called link function postulating a relationship between the expectation of the response variable vector $\mathbf{y}$ and the linear part $\mathbf{Xb}$ of the model. Special cases of link functions are identity, logistic and logarithmic functions used in linear models for continuous responses, logistic models for binary responses and log-linear models for count data, respectively.

The covariance structure of observations within clusters is modelled by

$$\mathbf{V}_i = \phi \mathbf{A}_i^{1/2} \mathbf{R}(\alpha) \mathbf{A}_i^{1/2}, \qquad i = 1, \ldots, m, \tag{8.30}$$

where $\mathbf{A}_i$ is a diagonal matrix of variances $V(y_k)$ in cluster i and $\mathbf{R}(\alpha)$ is the 'working' correlation matrix specified by the (possibly vector-valued) correlation parameter α of observations in cluster i. The parameter ϕ denotes the dispersion parameter of the corresponding member of the exponential family of distributions. Under an independent correlation assumption, all off-diagonal elements α of the 'working' correlation matrix are set to zero. Under an exchangeable correlation of pairs of observations within a cluster, the parameter α is a scalar and requires estimation. In an estimation procedure to obtain an estimate $\hat{\mathbf{b}}$, Newton–Raphson-type algorithms are usually used. The covariance-matrix estimate $\hat{\mathbf{V}}_{des}(\hat{\mathbf{b}})$ is obtained using a 'sandwich' type estimator (see equation (8.27)). Element weights can be incorporated in a GEE estimation procedure. GEE and the weighted analogue can be applied using suitable software for the analysis of complex surveys.

The GEE method has been shown to produce consistent estimates of model parameters and their covariance matrices, independently of a correct specification of the 'working' correlation structure. In the next two examples, we apply logistic ANCOVA first with the PML method and then with the GEE method assuming an exchangeable intra-cluster correlation structure. For further training on the PML and GEE methods in logistic modelling on the OHC Survey data, the reader is advised to visit the web extension of the book.

Example 8.2

Logistic ANCOVA with the PML method. Let us consider in a slightly more general setting the analysis situation of Example 8.1, where a logit ANOVA model was fitted by the GWLS method to proportions in a multidimensional table. We now

fit a logistic ANCOVA model using the PML method, by entering some of the predictors as continuous measurements in the model. The design-based analysis option is applied, providing valid PML analysis.

The binary response variable PSYCH measures high psychic strain, and we take the variables AGE, PHYS (physical working conditions) and CHRON (chronic morbidity) as continuous predictors such that AGE is measured in years and PHYS and CHRON are binary. Thus there are four predictors, of which SEX is taken as a qualitative predictor. So, the interaction of SEX with AGE, PHYS and CHRON can also be examined.

A model with SEX, AGE, PHYS and CHRON as the main effects and an interaction term of SEX and AGE was taken as the final model, because the other interactions appeared nonsignificant at the 5% level. Results of the model coefficients are displayed in Table 8.8.

The fitted logit ANCOVA model can be written using the estimated coefficients $\hat{b}_k$ and the corresponding model matrix $\mathbf{X}$ similar to the ANOVA modelling in Example 8.1:

$$F(\hat{f}_l) = \hat{b}_1 + \hat{b}_2(\text{SEX})_l + \hat{b}_3(\text{AGE})_l + \hat{b}_4(\text{PHYS})_l + \hat{b}_5(\text{CHRON})_l + \hat{b}_6(\text{SEX} * \text{AGE})_l,$$

where $l = 1, \ldots, 7841$, and $F(\hat{f}_l) = \log(\hat{f}_l/(1-\hat{f}_l))$. The values for the model terms are obtained from the corresponding columns of the 7841×6 model matrix $\mathbf{X}$. There, SEX, PHYS and CHRON are binary, and AGE has its original values (age

Table 8.8 Design-based logistic ANCOVA on overall psychic strain with the PML method.

Model term	Beta coefficient	Design effect	Standard error	*t*-test	*p*-value	Odds ratio	95% confidence interval for OR	
							Lower	Upper
Intercept	0.1964	1.56	0.1572	1.25	0.2127	1.22	0.89	1.66
Sex								
Males	−0.9926	1.43	0.2033	−4.88	0.0000	0.37	0.25	0.55
Females*	0	n.a.	0	n.a.	n.a.	1	1	1
Age	−0.0046	1.55	0.0041	−1.12	0.2624	1.00	0.99	1.00
Physical health hazards	0.2765	1.39	0.0596	4.64	0.0000	1.32	1.17	1.48
Chronic morbidity	0.5641	1.17	0.0575	9.82	0.0000	1.76	1.57	1.97
Sex, Age								
Males	0.0131	1.41	0.0051	2.56	0.0111	1.01	1.00	1.02
Females*	0	n.a.	0	n.a.	n.a.	1	1	1

* Reference class; parameter value set to zero.
n.a. not available.

in years). Note the difference in the ANCOVA model matrix when compared with that for the ANOVA model.

The *t*-tests on model coefficients indicate that the coefficients for the interesting predictors, physical working conditions and chronic morbidity are strongly associated with experiencing psychic strain. Persons in hazardous work, and chronically ill persons are more likely to suffer from psychic strain than healthy persons and persons whose working conditions are less hazardous. Note that the sex−age adjusted coefficient $\hat{b}_5$ for CHRON is larger than $\hat{b}_4$ for PHYS. Thus, in the model, chronic morbidity is more important as a predictor of psychic strain. This can also be seen in the odds ratio (OR) estimates provided in Table 8.8.

Odds ratios with their approximative 95% confidence intervals (in parenthesis) thus are

$$\text{PHYS: Odds ratio} = \exp(0.2765) = 1.32 \quad (1.17,\ 1.48),$$

$$\text{CHRON: Odds ratio} = \exp(0.5641) = 1.76 \quad (1.57,\ 1.97).$$

We may thus conclude that odds for experiencing a higher level of psychic strain, adjusted for sex, age and chronic morbidity, is about 1.3 times higher for those in more hazardous work than for those in less hazardous work. This conclusion was similar in Example 8.1, where a closely related odds ratio and confidence interval were obtained. Furthermore, the odds of experiencing much psychic strain, adjusted for sex, age and working conditions, are about 1.8 times higher for chronically ill persons than for healthier persons. Because neither of the 95% confidence intervals covers the value one, the corresponding odds ratios differ significantly (at the 5% level) from one. It should be noted that the binomial-based confidence intervals would be narrower especially for the predictor PHYS, for which the design-effect estimate is larger than for CHRON.

An analysis under the SRS options yield the same final model as the design-based analysis, but the observed values of the test statistics are somewhat larger and thus more liberal test results are attained.

Finally, let us examine more closely the fitted proportions $\hat{f}_l$ for the upper psychic strain group under the present model. The results are summarized in Figure 8.2 by plotting the proportions against the predictors included in the model. Fitted proportions increase with increasing age for males, and decrease for females. At a given age, the proportions are larger for the chronically ill and for those in more hazardous work than in the reference groups. Also, in females the fitted proportions tend to be larger than in males in all the corresponding domains, although the differences decline with increasing age.

Example 8.3

Logistic ANCOVA with the GEE method. Let us consider further the analysis situation of Example 8.2, where a logistic ANCOVA model was fitted by the PML method. We now fit a logistic ANCOVA model using the GEE method with

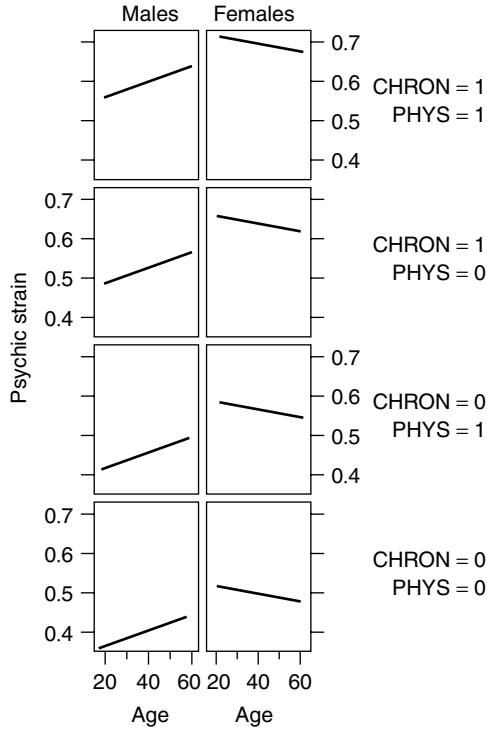


Figure 8.2 Fitted proportions of falling into the high psychic strain group for the final logistic ANCOVA model.

an assumed exchangeable correlation of pairs of observations within a cluster. Similarly as in Example 8.2, our response variable is the binary PSYCH measuring psychic strain. The variable SEX is included in the model as a categorical predictor and AGE, PHYS (physical working conditions) and CHRON (chronic morbidity) as continuous predictors such that AGE is measured in years and PHYS and CHRON are binary. We fit the same model as in Example 8.2.

Results are shown in Table 8.9. A comparison with logistic ANCOVA with the PML method in Example 8.2 indicates that the results are quite similar, and our inferential conclusions remain the same. There are, however, certain differences. First, the estimated beta coefficients have changed. Absolute values of estimates are larger than in the PML application, except for the CHRON effect. Standard-error estimates are somewhat smaller than the PML counterparts. Hence, the observed t-statistics tend to be larger involving slightly more liberal tests than in the PML case. These differences are due to the fact that in the GEE method with an exchangeable correlation structure, the correlation of observations also contributes to the estimation of the beta parameters. The 'working' intra-cluster

Table 8.9 Design-based logistic ANCOVA on overall psychic strain with the GEE method under exchangeable intra-cluster correlation structure.

Model Term	Beta coefficient	Design effect	Standard error	*t*-test	*p*-value
Intercept	0.2292	1.44	0.1524	1.50	0.1338
Sex					
Males	−1.0290	1.36	0.2000	−5.14	0.0000
Females*	0	n.a.	0	n.a.	n.a.
Age	−0.0057	1.43	0.0039	−1.45	0.1489
Physical health hazards	0.3011	1.31	0.0587	5.13	0.0000
Chronic morbidity	0.5569	1.14	0.0568	9.81	0.0000
Sex, Age					
Males	0.0144	1.33	0.0050	2.88	0.0044
Females*	0	n.a.	0	n.a.	n.a.

* Reference class; parameter value set to zero.
n.a. not available.

correlation is estimated as $\hat{\alpha} = 0.0189$. Using the expression $\text{deff} = 1 + (\overline{m} - 1)\hat{\alpha}$, where $\overline{m}$ is the average cluster size, this corresponds to an average design effect of 1.57.

Linear Modelling on Continuous Responses

We have extensively considered the modelling of binary response variables from complex surveys. The GWLS, PML and GEE methods were used, covering logit and linear modelling on categorical data and logit modelling with continuous predictors. These types of multivariate models are most frequently found in analytical surveys, for example, in social and health sciences. But in some instances it is appropriate to model a quantitative or continuous response variable, such as the number of physician visits or blood pressure. We discuss briefly the special features of multivariate analysis in such cases, and give an illustrative example of a special case of linear ANCOVA.

Linear modelling provides a convenient analysis methodology for analysis situations with a continuous response variable and a set of predictors. This situation was present in Examples 8.2 and 8.3, where the dichotomized PSYCH was analysed with a logistic ANCOVA model. There the original continuous variable on psychic strain could be taken as the response variable as well, leading to linear ANCOVA modelling. For a simple random sample, the analysis would be based on ordinary least squares (OLS) estimation with a standard program for linear modelling. For the OHC Survey data set, which is based on cluster sampling, the design-based approach with weighted least squares (WLS) estimation provides proper linear modelling.

Under the design-based option, similar complexities to those of the previous modelling techniques enter into linear modelling. In the estimation technique and testing procedures, however, no novel elements are involved compared to those already introduced for modelling with the GWLS, PML and GEE methods. So, we first aim at consistent estimation of the model coefficients and consistent estimation of the covariance matrix of the estimated coefficients. These require weighting with appropriate element weights, and the construction of a covariance-matrix estimator of the model coefficient estimates properly accounting for the clustering effects.

A linear regression model can be written compactly in matrix form as

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}, \tag{8.31}$$

where $\mathbf{y}$ is the vector of response variable values, $\mathbf{X}$ is the model matrix, $\mathbf{b}$ is the vector of regression coefficients to be estimated and $\mathbf{e}$ is the vector of random errors.

Under the design-based and weighted SRS options, the vector $\mathbf{b}$ is consistently estimated by solving the weighted normal equations

$$\mathbf{X}'\mathbf{W}\mathbf{X}\hat{\mathbf{b}} = \mathbf{X}'\mathbf{W}\mathbf{y}, \tag{8.32}$$

where the diagonal elements of $\mathbf{W}$ are the rescaled element weights w_l^{**}. Under the unweighted SRS option, the weights are all one, and the estimation reduces to usual OLS estimation. The WLS estimator $\hat{\mathbf{b}}$ is given by

$$\hat{\mathbf{b}} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{y}. \tag{8.33}$$

Under the design-based option, as for the design-based PML method for proportions, the covariance matrix of the estimator $\hat{\mathbf{b}}$ can be estimated consistently by a 'sandwich' type estimator. Also, desired tests of model adequacy and of linear hypotheses on model coefficients can be executed using test statistics similar to the Wald and F-statistics used in the GWLS, PML and GEE methods for logit and linear modelling on proportions.

Linear modelling under the design-based option can be carried out in practice most conveniently with appropriate software for survey analysis.

Example 8.4

Linear ANCOVA modelling with the WLS method on perceived psychic strain. In Examples 8.2 and 8.3, a logistic ANCOVA model was fitted on the dichotomized variable PSYCH of psychic strain. A linear ANCOVA model is now fitted on the original variable PSYCH, whose values are scores of the first standardized principal component of nine psychic symptoms. Thus, the average of PSYCH is zero and the variance is one. The distribution of PSYCH is, however, somewhat skewed; there are numerous persons in the data set not experiencing any of the psychic

symptoms in question. The range of values of PSYCH is $(-1, 4.7)$, and the median of the distribution is -0.4.

We include the same variables as in the previous two examples as the potential predictors in the linear ANCOVA model. The predictor SEX is taken to be qualitative, and AGE, PHYS and CHRON are taken to be continuous, and we also study the pairwise interactions of SEX and the continuous predictors. The model is fitted by the WLS method, and the model-building produces a similar ANCOVA model as in Examples 8.2 and 8.3. Thus, all the main effects and the interaction of SEX and AGE appear significant.

The fitted linear ANCOVA model on PSYCH can be written using the estimated coefficients $\hat{b}_k$ and the corresponding model matrix $\mathbf{X}$, as in the logistic model on a binary PSYCH:

$$\hat{f}_l = \hat{b}_1 + \hat{b}_2(\text{SEX})_l + \hat{b}_3(\text{AGE})_l + \hat{b}_4(\text{PHYS})_l + \hat{b}_5(\text{CHRON})_l + \hat{b}_6(\text{SEX} * \text{AGE})_l,$$

where $l = 1, \ldots, 7841$, and the values for the model terms are obtained from the model matrix $\mathbf{X}$ of Example 8.2. Results on the ANCOVA model coefficients with the continuously measured psychic strain as the response variable are displayed in Table 8.10.

The signs of model coefficients and the t-test results follow a similar pattern to those in the corresponding logit ANCOVA model in Example 8.2. The model coefficients, however, have different interpretations from those in the logit model. In a logit ANCOVA, we were working on a logit scale on the binary response, whereas we are now dealing with continuous measurements on a linear scale. Thus, the coefficients of the linear ANCOVA model can be interpreted in the usual linear regression context.

Under the weighted SRS analysis option, the same ANCOVA model would be obtained, and the results on model coefficients would be equal. But the standard errors of the model coefficients would be smaller because the design-effect estimates $\hat{d}(\hat{b}_k)$ are greater than one. However, this does not affect the inferences from the t-test results.

The continuous response variable PSYCH offered good possibilities for the demonstration of linear modelling due to the continuity of the response variable, although the distribution was somewhat skewed. Count variables, such as the number of physician visits in a given time interval or related variables whose distribution can be very skewed, are often met with in practice. Modelling of such quantitative response variables can involve such symmetrizing transformations as logarithmic, often used in econometrics, or Box–Cox transformations, prior to the fitting of a linear model. Moreover, a linear model formulation can even be inappropriate for such variables. Then, other regression modelling techniques should be used: for example, Poisson regression and a negative binomial model to account for extra-Poisson variation. These methods belong to a class of generalized linear models for correlated

Table 8.10 Design-based linear ANCOVA on overall psychic strain with the WLS method.

Model term	Beta coefficient	Design effect	Standard error	*t*-test	*p*-value
Intercept	−0.0121	1.70	0.0831	−0.15	0.8846
Sex					
Males	−0.4975	1.48	0.0997	−4.99	0.0000
Females*	0	n.a.	0	n.a.	n.a.
Age	−0.0001	1.60	0.0021	−1.02	0.9804
Physical health hazards	0.1772	1.37	0.0290	6.11	0.0000
Chronic morbidity	0.3922	1.17	0.0294	13.33	0.0000
Sex, Age					
Males	0.0057	1.39	0.0025	2.25	0.0252
Females*	0	n.a.	0	n.a.	n.a.

* Reference class; parameter value set to zero.
n.a. not available.

response variables. For these models, for example, the pseudolikelihood and generalized estimating equations methods can be successfully used under the nuisance approach.

Methods for the Disaggregated Approach

Methods for multivariate analysis considered so far fall under the nuisance or aggregated approach, where the aim is to clean out the possibly disturbing clustering effects from the analysis results in order to attain consistent estimation and asymptotically valid testing. Under the disaggregated approach, on the other hand, intra-cluster correlation structures are intrinsically interesting, and the estimation of these correlations constitutes an essential part of the analysis. This often occurs in social and educational surveys when working with hierarchically structured data sets. Clustering with villages, establishments or schools constitute common examples of sources of such a hierarchical structure.

There are advanced methods available for multivariate analysis of intra-cluster correlated response variables from hierarchically structured data sets. The methodology of *multi-level modelling* is based on *generalized linear mixed models*, where certain random effects are incorporated in the model. These constitute a new class of models not yet considered in this book; in all the previous models, the model parameters have been taken as fixed effects. Applications of multi-level modelling have been mainly in linear modelling of continuous response variables from educational surveys, where schools or teaching groups are used as the clusters (Goldstein 1987, 2002). Multi-level models have also been developed for binary and polytomous responses, and appropriate computing algorithms are available. We will use multi-level modelling in Section 9.4 for a continuous

response variable from clustered educational data. There, a brief introduction to the method will be given.

8.5 CHAPTER SUMMARY AND FURTHER READING

Summary

Linear and logit modelling of an intra-cluster correlated response variable were considered in this chapter mainly under the nuisance approach. The principal aim was to successfully remove the effects of intra-cluster correlations from the estimation and test results. The severity of these effects, however, varies under different sampling designs and therefore various analysis options were introduced for proper analysis in practice.

A design-based analysis option provides a generally valid analysis option for multivariate analysis in complex surveys. Under this option, the complexities of the sampling design can be properly accounted for, including clustering, stratification and weighting. Analysis under the design-based option requires access to the element-level data set, and availability of proper software for survey analysis. Also, under stratified element sampling and simple random sampling, the weighted and unweighted SRS options can be used for valid analysis. Under the weighted SRS option, only the weighting is covered, and the unweighted SRS option ignores all the sampling complexities. These options are thus inappropriate for clustered designs of complex surveys.

Under any of the analysis options, logit and linear ANOVA, ANCOVA and regression analysis on domain proportions of a binary or polytomous response variable can be carried out by the GWLS method of generalized weighted least squares estimation in a data set arranged into a multidimensional table. The GWLS method, applied under the design-based option, provides valid analysis for such tables from complex surveys. For reliable results, a large element sample and a large number of sample clusters are required; these conditions are usually met in large-scale analytical surveys such as the OHC Survey based on a stratified cluster-sampling design. With a small number of sample clusters, instability problems can arise, making the estimation and test results unreliable. This problem can be successfully handled using appropriate correction techniques for the test statistics.

The PML method of pseudolikelihood estimation can be used in analysis situations similar to the GWLS method, but its main applications are in logistic regression with continuous predictors where the GWLS method fails. Under the design-based option, the PML method provides valid logit analysis for complex surveys. It is also beneficial for the PML method that the number of sample clusters is large, and similar adjustments are available for unstable cases, as for the GWLS method. We applied the PML method for logistic ANCOVA modelling in an OHC Survey case study on a binary response variable.

The PML method covers not only logistic regression models but also other model types from the class of generalized linear models. So, linear models on continuous responses are also covered. We briefly introduced the GEE method of generalized estimating equations. The GEE version assuming an exchangeable correlation structure within the clusters was applied for logistic ANCOVA modelling on a binary response, and the results were similar to those from the PML application.

In the case studies on selected multivariate analysis situations from the OHC Survey, it appeared that accounting for sampling complexities, especially for the clustering effects, can be crucial for valid inferences. We shall demonstrate this important conclusion further in Chapter 9, where additional case studies from other complex survey data sets will be given.

The nuisance, or aggregated, approach provides a reasonable and manageable analysis strategy for different kinds of multivariate analysis situations on an intra-cluster correlated response variable. In the alternative disaggregated approach, the intra-cluster correlations are taken as intrinsically interesting parameters to be estimated as well as the model coefficients. We discussed briefly multi-level modelling, applicable for hierarchically structured data sets. The method of multi-level modelling will be demonstrated in the next chapter.

Further Reading

Multivariate analysis of complex surveys has received considerable attention in the literature. Advances in the methodology can be found in Binder (1983), Rao and Scott (1984, 1987), Roberts *et al.* (1987), Rao *et al.* (1989) and Scott *et al.* (1990), covering, for example, the weighted least squares, pseudolikelihood and quasi-likelihood methods for logit and related analysis of categorical data from complex surveys. The book edited by Skinner *et al.* (1989) covers many of the important advances in multivariate analysis under both the aggregated and disaggregated approaches. Rao and Thomas (1988) and Korn and Graubard (1999) provide more applied sources on the methodology. The book edited by Chambers and Skinner (2003) includes several articles on different views into the analysis methodology for complex survey data.

Rao *et al.* (1993) discuss regression analysis with two-stage cluster samples. Binder (1992) addresses the fitting of proportional hazards models to complex survey data. The analysis of categorical data with nonresponse is considered in Binder (1991), and Glynn *et al.* (1993) consider multiple imputation in linear models.

Multi-level modelling is introduced in Goldstein (1987, 1991), and is further developed in Goldstein and Rasbash (1992) and Goldstein (2002). Pfeffermann *et al.* (1998) consider weighting issues in multi-level modelling. Modelling by the generalized estimating equations is introduced in Liang and Zeger (1986), and is further developed in Liang *et al.* (1992) and Diggle *et al.* (2002). Horton and

Lipsitz (1999) discuss software and Ziegler *et al.* (1998) address literature on GEE methodology. Breslow and Clayton (1993) give general results on approximate inference in the framework of generalized linear mixed models. Analysis of complex longitudinal survey data is discussed in Clayton *et al.* (1998) and Feder *et al.* (2000).

9

More Detailed Case Studies

Four additional case studies are selected to provide a more subject-matter-oriented demonstration of the survey methodology discussed in this book. The first case study (Section 9.1) deals with monitoring the quality of data collection in a long-term survey. A number of statistics introduced earlier in this book are used as quality indicators. Empirical findings are from a *passenger transport survey*. The data-collection period covered a full calendar year with equal-sized monthly samples.

The second case study (Section 9.2) is from a *business survey* that is an example of resolving sampling frame problems often met in the production of business statistics. The estimation of the annual mean salary of certain occupational groups is discussed when two different frames are present. This results in a data-collection strategy of mixed type in which three-quarters of data are collected by the census-type and one-quarter by the survey-type. In addition, our analysis on the business survey proves that the clustering effect should be accounted for calculating employer-level statistics from a sample in which the sampling units are firms.

In the case study from a *socioeconomic survey* (Section 9.3), a logit model is fitted to categorical data from a cluster-sampling design with households as the clusters. The main emphasis is not only on pointing out the importance of accounting for the clustering effects but also on the importance of adequate selection of model type for analysis. Here, analysis of variance and regression-type logit models are used, which lead to different conclusions.

In the final case study (Section 9.4), we introduce and demonstrate an approach of modelling hierarchically structured data sets using multi-level regression models, applied to clustered survey data from a multinational *educational survey*. These models differ from the methods of the nuisance approach, as used in the preceding case study, in the sense that in multi-level modelling, the hierarchical structure of the population is reflected in the structure of the model. Some interesting comparisons between countries are also included.

Practical Methods for Design and Analysis of Complex Surveys Risto Lehtonen and Erkki Pahkinen
 ISBN: 0-470-84769-7

9.1 MONITORING QUALITY IN A LONG-TERM TRANSPORT SURVEY

Data-collection operations in many surveys can be of a long-term nature covering, for example, a whole calendar year. Good examples from this type of social surveys are consumer attitude surveys and travel or mobility surveys in which the total sample is divided into 12 equal-sized subsamples. This kind of survey strategy is targeted at two different goals: to collect monthly cross-sectional data and to compile yearly data to catch seasonal, cyclic or trend characteristics of the phenomena. In such surveys, a major issue is the maintenance of uniform data quality throughout the entire survey period. In this, monitoring the quality of the data-collection procedure becomes important.

In this case study, a set of 20 statistical quality indicators are presented to monitor possible deviations in quality for each data-collection wave. The indicators cover important aspects of sampling and nonsampling errors. Some indicators are defined earlier in this book, such as the coefficient of variation, coverage rate, response rate and intra-class correlation. More extensive consideration of different survey errors can be found in Groves (1989). Cox *et al.* (1995) deals with the subject in the context of business surveys. Biemer and Lyberg (2003) gives a non-technical introduction to survey quality.

Passenger Transport Survey

The use of quality indicators is demonstrated in a long-term survey, the Passenger Transport Survey, conducted by the Finnish Ministry of Transport and Communications in 1998–1999. The survey totalled 18 250 sampling units divided into equal-sized monthly slots of 1500 persons. Data were collected by computer-assisted telephone interview (CATI). The main results and survey processes are reported in Pastinen (1999).

For monitoring the homogeneity of quality, two report formats were developed for the monthly data-collection slots. The indicators were calculated for each successive data-collection wave and compiled into a report format presenting the values for the current sample and the cumulative sample. Monthly calculated quality reports served as a basis for monitoring the homogeneity of the data-collection process. Using these data, operations to correct the process could be made when necessary.

An aim of the survey was to describe the mobility of people registered in Finland, aged six years or over. The sample was selected by stratified simple random sampling with proportional allocation. Stratification was based on age/sex/area groupings. Data collection was timed in 12 monthly waves, each including 1500 sampling units selected from the Central Population Register. The data was collected between July 1998 and June 1999. The survey covered every day for a full period of 12 months so that temporal variation in mobility could be taken into account. Data were processed on a monthly basis thus resulting in 12 data files.

To ensure the quality of the fieldwork, the interviewers received advance training and the data collection was monitored on a monthly basis. The interviewers were also given regular feedback on their performance so that the material they were collecting would be of consistent quality. The prospective respondents were provided with advance information about the survey. For example, each respondent received a contact letter detailing the background and objectives of the survey.

We first present empirical findings on the four key quality indicators: coverage rate (%), response rate (%), interviewer effect and coefficient of variation (%). Then, one of the two report formats used for monitoring the quality of monthly collected data is briefly discussed.

Monitoring Coverage Rate

In this survey, coverage rate (%) is defined as follows. The frame population for sampling consists of a relevant population register. The frame for telephone numbers consists of a register of phone numbers and the names of persons. Coverage error is present if these two registers do not coincide. We estimate the coverage rate by

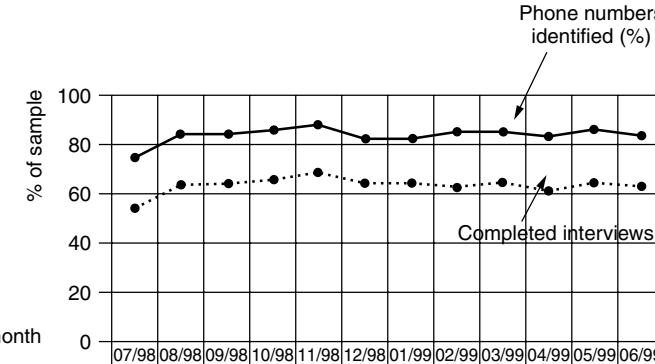
$$\text{COVERAGE RATE (\%)} = (n_F/n) \times 100,$$

where n_F is the number of sample persons whose phone number is identified in the frame and n is the sample size.

The phone penetration serves as an example. In a computer-assisted telephone interview, the target population might be all the adult persons living in private households in the country. The frame population, list of a database of phone numbers, includes only persons who can be contacted by phone. Usually, this frame population is noticeably smaller than the target population, thus causing an under-coverage error. This is a nonsampling error due to non-observation.

General telephone coverage in Finland is very high, as reported in Kuusela (2000). Over 96% of households owned either ordinary or mobile phones or both in 1996. The high density of phones does not ensure that telephone interviewing is a successful data-collection mode in the sense of good coverage. A considerable drawback is usually met during the identifying process of phone numbers. As seen in Figure 9.1, the proportion of identified phone numbers in the Passenger Transport Survey is about 85%. Under-coverage is thus 15%.

In phone interviews, the contact-making starts by locating up-to-date information on addresses and phone numbers. The addresses may be culled from a recent national census register, but finding phone numbers often causes problems. However, even if a household has a phone it is not guaranteed that the phone number will be found.



	07/98	08/98	09/98	10/98	11/98	12/98	01/99	02/99	03/99	04/99	05/99	06/99
Phone numbers identified (%)	75	84	84	86	88	83	83	86	85	85	87	85
Responses (%)	56	64	65	67	70	65	66	64	63	62	64	64

Figure 9.1 Percentage of sample persons for whom phone numbers were identified and interviews were completed in each survey month.

During the first survey month in July 1998, the percentage of phone numbers identified remained below average. After this defect was found and adjusted, the seeking out of phone numbers could be speeded up and the outcome could be improved over the following months.

Monitoring Response Rate

Response rate (%) indicates the proportion of participating sample persons. A measure for response rate is

$$\text{RESPONSE RATE (\%)} = \frac{I}{I + R + NC + O} \times 100,$$

where I = number of interviewed persons
R = number of refusals known to be eligible
NC = number of non-contacts known to be eligible
O = number of other eligible sample units non-interviewed.

The seriousness of nonresponse is twofold: firstly, it decreases the effective sample size thus inflating standard errors of estimates and secondly, possibly causes nonresponse bias if respondents give values for study variables that would systematically deviate those of the nonrespondents. Therefore, survey organization recorded by continuous basis reasons for nonresponse (see Table 9.1).

As seen again from Figure 9.1, the monthly calculated response rate was about 65% except in July 1998, which was the first survey month. Temporal variation of the response rate is insignificant during the total survey period. In the Passenger Transport Survey, finding phone numbers had a high correlation with the final

response rate, and if a low number of phone numbers was found there was little the interviewers could do to raise the response rate. This explains an exceptionally low response rate (%) in July 1998. The response rate was 10% units smaller than the average rate due to low identification (%) of phone numbers. Compared to similar national surveys there are no significant discrepancies in the response rate. Groves *et al.* (2001) reports several national surveys from the point of view of nonresponse.

Monitoring Interviewer Effect

Interviewer effect belongs to the class of nonsampling errors. A telephone or personal interview is a social interaction process between the interviewer and the respondent. Biemer *et al.* (1991) list four ways in which the interviewer effect might occur: (a) the survey interview is seen as a structured social interaction, (b) variations among interviewers when filling in questionnaires, (c) differing word emphasis or intonation and (d) individual reaction to respondent difficulties. All four factors might cause correlated answers within interviewers. A commonly used statistic to assess this source of survey error is the intra-class correlation coefficient (Kish 1962). Denoting by $\overline{m}$ the average size of the workload of an interviewer, intra-class correlation can be estimated from the formula

$$\hat{\rho}_{\text{int}} = \frac{\left(\frac{\hat{V}_b - \hat{V}_w}{\overline{m}}\right)}{\left(\frac{\hat{V}_b - \hat{V}_w}{\overline{m}}\right) + \hat{V}_w},$$

where the interviewer variance component is $\hat{V}_b$ measured as the between mean square in a one-way analysis of variance with interviewers as the factor, and $\hat{V}_w$ is the corresponding within mean square. Its value varies as $-\frac{1}{\overline{m}} \leq \hat{\rho}_{\text{int}} \leq 1$. Note that this formula deviates from that given earlier for systematic sampling and cluster sampling in Chapters 2 and 3. There, the intra-class correlation was defined in a design-based setting. The starting point here is a model for measurement error caused by interviewers, and thus the coefficient of intra-class correlation is calculated in a model-based setting allowing also for varying workload size. The contribution of the intra-class correlation caused by the interviewer effect should be included in the standard error estimate of an estimate. For example, had a nonzero $\hat{\rho}_{\text{int}}$ met, the estimated design variance of the sample mean should be multiplied by an inflating factor of deff $= 1 + (\overline{m} - 1)\hat{\rho}_{\text{int}}$.

Next, an empirical finding is presented from the Passenger Transport Survey. As a study variable, the number of trips per person per day was selected. In Figure 9.2, the estimated $\hat{\rho}_{\text{int}}$ is presented as a monthly figure and on a cumulative basis. Note

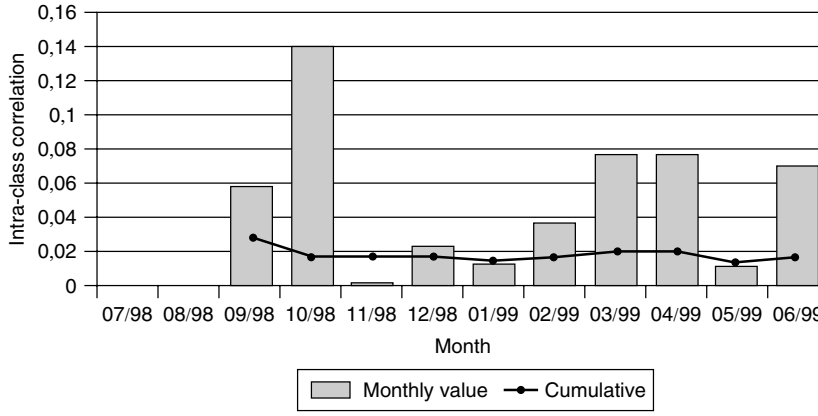


Figure 9.2 Intra-class correlation of the number of trips per person per day.

that the monthly figures are calculated separately for each month's workloads, and for cumulative figures, the workloads of each interviewer are combined over respective months. The first two months (June and July 1998) are lacking because the monitoring of this characteristic started in August 1998.

On the basis of the cumulative figures, an average $\hat{\rho}_{\text{int}}$ is about 0.02. Many research findings show that in large-scale telephone surveys the value of $\rho_{\text{int}} \approx 0.02$ is typical (Groves 1989). As an interviewer effect, if present as in this case, it broadens the confidence interval thereby absorbing the cumulative effect of the increasing sample size, which in turn decreases the sampling error. This finding recommends limiting the maximum workload per interviewer in large long-term surveys in order to prevent overly large samples from being interviewed by the same interviewer. Because of this, the sample persons should be assigned to each interviewer randomly, a practice that evens out the interviewer effect within sampled elements (Biemer *et al.* 1991).

One can calculate the average inflating effect of the intra-class correlation on the sample mean by taking the monthly value as the basis. In June 1999, $\hat{\rho}_{\text{int}}$ was 0.071 and the average workload $\overline{m}$ of the interviewers consisted of 96 respondents. Thus, the inflating factor is $\text{deff} = 1 + \hat{\rho}_{\text{int}}(\overline{m} - 1) = 1 + 0.071(96 - 1) = 6.75$. For example, to adjust for the interviewer effect, the estimated standard errors of sample means should be multiplied by the square root of this factor or

$$\text{s.e}(\overline{y}) = \sqrt{6.75} \times \text{s.e}(\overline{y})_{p(s)} = 2.60 \times \text{s.e}(\overline{y})_{p(s)}.$$

Monitoring Sampling Error Using the Coefficient of Variation

Coefficient of variation (%), denoted as $\text{C.V}(\hat{\theta})\%$ measures the relative sampling error. For a non-negative study variable y, the estimated coefficient of variation

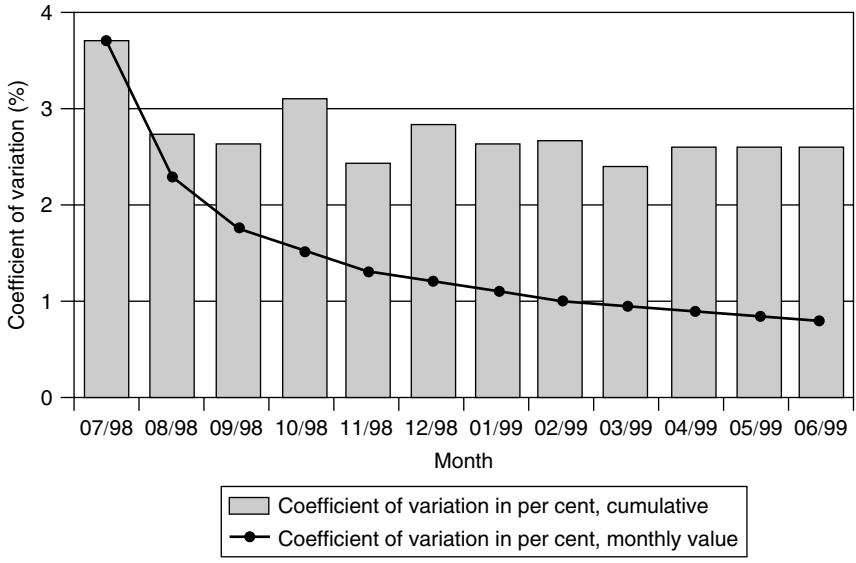


Figure 9.3 Coefficient of variation (%) of the average number of trips per person per day; monthly and cumulative figures.

for a point estimate $\hat{\theta}$ is given by $\text{c.v}(\hat{\theta}) = \text{s.e}(\hat{\theta})/\hat{\theta}$. For making easier comparisons between variables, surveys and monthly data slots, the coefficient of variation is defined as a percentage by

$$\text{COEFFICIENT OF VARIATION (\%)} = \frac{\text{s.e}(\hat{\theta})}{\hat{\theta}} \times 100.$$

Figure 9.3 represents the monthly and cumulative values of the coefficient of variation of the average number of trips per person per day. The cumulative value clearly shows that the increase in the number of observations reduces the coefficient of variation.

The average monthly value is about 2.7%, showing only slight relative sampling error. As expected, the cumulative value of c.v(%) declines steadily when the number of monthly slots increases.

A Format for a Quality Report

The survey organization decided to provide two monthly quality reports so that the homogeneity of quality between successive data-collection waves could be monitored. The first form included 25 different indicators whose values were

Table 9.1 An example quality report: June 1999. Passenger Transport Survey 1998–1999.

Measurement	June	Cumulative	Remarks
Sample size	1500	18 250	Monthly/yearly
Telephone numbers identified	84.7%	84.2%	Coverage rate (%)
Eligible persons contacted	77.3%	77.5%	Contact rate I
Contacted somebody at home	78.2%	77.9%	Contact rate II
Responded by mobile phone	16.1%	12.5%	Contact rate III
Completed interviews	63.9%	64.2%	Response rate (%)
Unable to give answers	0.9%	1.3%	Cause of nonresponse
Linguistic problems	0.0%	0.2%	Cause of nonresponse
Refusals and reasons for refusal	12.5%	11.9%	Causes of nonresponse
• No time/busy	1.8%	2.0%	
• Don't cooperate on principle	5.5%	3.6%	
• Fearing the misuse of personal data	0.0%	0.0%	
• Useless survey	0.2%	0.1%	
• Uncertain about the use of study results	0.0%	0.0%	
• Not interesting survey	1.3%	1.6%	
• Other reason	3.7%	4.6%	
Interview interrupted	0.1%	0.1%	Cause of nonresponse
No contact	22.7%	22.5%	No contact rate
Known endpoint from total number of trips	76.3%	71.7%	Measurement error
Number of interviewers	10	19	Monthly/yearly
Completed interviews per interviewer	96	616	Workload/interviewer
Intra-class correlation of the numbers of trips	0.071	0.017	Interviewer effect
Intra-class correlation of daily kilometrage	0.0024	0.0016	Interviewer effect
Coefficient of variation of the number of trips	2.8%	0.8%	Sampling error
Coefficient of variation of daily kilometrage	9.9%	2.7%	Sampling error

Source: Pastinen (1999). Passenger Transport Survey 1998–1999 (in Finnish). Publications of the Ministry of Transport and Communications 43/99. Finland: Edita Ltd

calculated from monthly data and from cumulative data. An example of this type of format is reproduced in Table 9.1.

This report targeted the use of client and survey organizations. In Table 9.1, cumulative figures serve as benchmarks for monthly figures. For example, the coefficients of variations are presented on the two last rows. In practice, it is important to estimate coefficients of variation for all variables of interest, especially to check that the maximum acceptable level for releasing intermediate results is not exceeded.

9.2 ESTIMATION OF MEAN SALARY IN A BUSINESS SURVEY

The main concern in this case study is the estimation of average salaries of employees in different occupations within the commercial sector using data

collected from business firms. In the sampling design, the primary sampling unit is the individual firm, which implies that data on salaries at the employee level are clustered by firms and so, accordingly, this design should be taken into account in the estimation. The actual sampling design is stratified one-stage cluster sampling. In the estimation of the average salaries in the commerce sector as a whole, as well as in certain occupational groups within this sector, three other sampling design assumptions are also used for comparison.

Sampling Design

The sampling frame used is a business register, in which business firms in the commerce sector are divided into two subpopulations. The first comprises all the firms that are members of the Confederation of Commerce Employers (for convenience, CCE firms). From this subpopulation, the Confederation collects census data on salaries in different commercial occupations. The average salaries calculated on the basis of the complete data set will be used as a point of reference in subsequent comparisons.

The other subpopulation comprises firms that are not members of the Confederation of Commerce Employers. From this subpopulation, a stratified simple random sample has been selected, using the individual firm as the primary sampling unit. Our aim is to estimate the average salaries for different occupations in this subpopulation using the collected sample data.

For a sampling frame for the present sample, the smallest companies (those employing 1–2 people) have been first excluded from the business register. This leaves a population of 25 345 companies, which is stratified into five categories by the number of employees and into five categories by the branch of business, giving 25 strata. Sampling fractions vary by stratum; in some strata, all firms are included, and in others, only some firms. The order in which individual firms appear in the Business Register is then stratum-wise randomized. Next, starting from the top, the required number of units is sampled from each stratum. The initial sample size was 1572 business firms. Excluding the frame over-coverage of 165 CCE member firms, 76 non-eligible firms and 38 firm closures resulted in a final sample of 1369 business firms. The number of responding firms was 1100, thus the response rate was 80%.

Insofar as the sampling takes place at the firm level, the sampling design may be described as stratified simple random sampling without replacement. If conclusions were to be drawn for the firm level, then the analysis would be carried out within a stratified simple random sampling design. For example, this sort of sample design is well suited to the analysis of turnover and similar firm-level data.

However, the purpose here is to estimate the average salaries of employees in different occupations. This implies a different interpretation of the sampling design in that the individual employee who is the unit of analysis is not the primary sampling unit. The selection of a certain firm into the sample implies that all its

employees are also included. Each selected firm should therefore be interpreted as a cluster, the elements of which are all the firm's employees. This sample design is described as *stratified one-stage cluster sampling*. There is only one single stage in the sampling procedure; namely, the sampling of firms. Within each selected firm, then, data are collected on the salaries of all employees.

The specific concern here is the regular monthly salaries of commercial occupations at the time of measurement in August 1991. These occupations are grouped according to the classification used by Statistics Finland. The average salaries of 22 occupational groups are regularly published, but some of these categories are so small that for reasons of confidentiality only the job title can be indicated. The focus here is restricted to the occupational groups that occur in at least 50 sampling units or firms. One item obviously of special interest is the average salary for the whole commercial sector, which in the present sample design comprises 744 firms or clusters with a total of 13 987 employees. When weighted by the inverse of the sampling rate, the size of the corresponding population is estimated to be $\hat{N}_{\text{STATFIN}} = 57\,762$ employees. For comparison, the total number of employees in the CCE Register is $N_{CCE} = 190\,217$.

Weighting and Estimators of the Mean

For the present kind of sample data, it is possible to construct different types of mean estimators depending on the assumptions made in the sampling design. In the following text, four alternative sampling designs are presented with the corresponding mean and design-effect estimates. Appropriate variance estimators have been considered in Chapters 2, 3 and 5 and we omit them here.

Simple random sampling The firm level is omitted and the sample at the employee level is interpreted as a simple random sample taken directly from the employee population. Thus, the corresponding estimator of average salary is

$$\overline{y} = \frac{\hat{N}}{n} \sum_{k=1}^{n} y_k / \hat{N}, \tag{9.1}$$

where y_k is the salary of the kth employee in the sample and the joint sample size is $n = 13\,987$. The same weight $\hat{N}/n$ is used for all employees; this is the inverse of the approximate sampling rate. The weight is $\hat{N}/n = 57\,762/13\,987 = 4.13$. This coefficient could only be justified if the sampling had been carried out at the employee level and if neither stratification nor clustering had been done. In the present case, neither of these conditions holds. The variance of the mean estimator is useful in determining the estimate of the design effect, a measure that summarizes the effects of design complexities on variance estimation. As

defined in Chapter 2, the design-effect estimator for the mean is a ratio of two variance estimators:

$$\text{deff}\,(\overline{y}^*) = \frac{\hat{v}_{p(s)}(\overline{y}^*)}{\hat{v}_{srs}(\overline{y})}, \tag{9.2}$$

where $\overline{y}^*$ is an estimator of the mean under the actual sampling design $p(s)$ with a variance estimator $\hat{v}_{p(s)}(\overline{y}^*)$, and $\hat{v}_{srs}(\overline{y})$ is the variance estimator of $\overline{y}$ under SRSWOR. If the design effect is close to one, the actual sample design can be interpreted as an SRS design. In this case, the analysis does not require sampling-design identifiers. In situations in which cluster sampling is used, the design effect can be larger than one. Then, to obtain a proper analysis it is necessary to use specialized software with the appropriate design identifiers. Under the SRS design, the design effect is by definition equal to one.

Stratified simple random sampling Element-level sampling is assumed and each stratum is assigned its own weight. The estimator of the average salary is

$$\overline{y}_{str} = \sum_{h=1}^{H} \sum_{k=1}^{n_h} \frac{\hat{N}_h}{n_h} y_{hk} / \hat{N}. \tag{9.3}$$

The stratum-specific weights are $\hat{N}_h/n_h$ or the inverse of the sampling rate in stratum h where $\sum_{h=1}^{H} \hat{N}_h = \hat{N}$ and $\sum_{h=1}^{H} n_h = n$. It is worth noting that the weight remains constant for all employees in the same stratum even if (as indeed is the case in practice) they work in different companies.

Stratified cluster sampling with stratum-wise varying weights The estimator for the mean is equal to that of stratified simple random sampling. However, the designs involve different estimators for the standard error, which can be used to determine confidence intervals, for instance. In stratified cluster sampling, the design effect is usually larger than one (deff ≥ 1), depending on the internal homogeneity of the clusters with respect to the study variable.

Stratified cluster sampling with cluster-wise varying weights This is a very realistic assumption in samples of business firms. The size of firms (i.e. the size of the cluster), measured in terms of the number of employees, usually varies considerably. In this case, the design can be taken into account by estimating the mean using the *Horvitz–Thompson* estimator and regarding the relative size of a cluster as the sampling weight. Here, the relative size of a cluster is measured by the number of employees N_{hi} in a firm divided by the total number of employees N_h in the corresponding stratum. This will yield a cluster weight for a certain firm, and the inverse of this figure is, accordingly, the sampling weight for that particular firm. To match the sum of the weights with the total number of employees within the

frame population, this figure must still be divided by the number m_h of sample firms in the stratum. Thus, the mean estimator is

$$\overline{y}_{clu} = \sum_{h=1}^{H} \sum_{i=1}^{m_h} \sum_{k=1}^{n_{hi}} \frac{\hat{N}_h}{m_h \times N_{hi}} y_{hik} / \hat{N}. \tag{9.4}$$

The estimator incorporates all the information concerning the sampling design: sampling weights that vary firmwise, and stratification.

Results

The sample data have been analysed so that the appropriate sample design can be properly taken into account. Estimations under the four sampling design assumptions differ in their weighting schemes and they take the same sampling design into account to varying extents. The most realistic of these design assumptions is obviously stratified cluster sampling with cluster-wise varying weights, which incorporates all the information concerning the sampling design, whilst the SRS design is the simplest one. The results on these sampling designs can also be compared with the statistics on average salaries obtained by the CCE from its census. In Table 9.2, these data are shown on the last line. The Statistics Finland sample specifies the estimated number of employees as 57 762, which means that the figure for the whole sector in August 1991 would have been 57 762 + 190 217 = 247 979 full-time employees.

The estimates from the SRS design give the largest average salary as EUR 1759. On the other hand, it also has the smallest standard error estimate of s.e = 7.4. In other designs, the average salary approximates the reference figure obtained from a census, which is EUR 1530. Since this is the exact figure for the corresponding subpopulation, it obviously contains no standard error. The design that estimates closest to the reference figure is stratified cluster sampling with cluster-wise weights. The estimated average salary from this design is EUR 1581.

Table 9.2 Average salary (EUR) of commercial sector employees in 1991 based on different sampling design assumptions and census data.

Sample design	Weighted sample size	Average salary	Standard error	deff
SRS	57 762	1759	7.4	1.00
STR (stratified)	57 762	1602	9.3	1.72
CLU (stratum weights)	57 762	1602	10.1	2.10
CLU (cluster weights)	57 762	1581	11.1	2.58
Census (CCE register)	190 217	1530	0	n.a.

n.a. not available

There, the primary sampling unit was the firm, but the weighting is done at the employee level.

Comparison of the Results

Moving on to look at average salaries in selected commercial occupational groups, Table 9.3 compares the figures from three sources: the Confederation of Commerce Employers register data, the Statistics Finland estimates based on the stratified one-stage simple random sampling and finally the estimates obtained from the stratified cluster sampling design with cluster-wise varying weights. The comparison covers the biggest occupational categories on which data have been obtained from at least 50 companies (Table 9.3).

There are certain differences between the figures based on the census data and the sample compiled by Statistics Finland. However, since these differences only occur in a small number of occupational groups, it would seem useful to look more closely at the internal compatibility of occupational classifications used in different statistical sources. On average, the estimates from stratified cluster sampling with cluster-wise weights come closer to the census figures than those of Statistics Finland, which are based on an assumption of stratified simple random sampling.

The use of complete design information significantly increases the standard errors of average salary estimates. One possible reason for this is that during

Table 9.3 Average salaries in different occupational groups in August 1991: census of CCE member companies and the Statistics Finland sample.

	Average salary in August 1991		
		STATFIN sample	
Occupational group	CCE census	CLU design	STR design
Shop managers	1612	1486	1430
Service station workers	1159	1173	1161
Cleaners	1150	911	906
Warehouse workers	1195	1196	1191
Van/lorry drivers	1313	1201	1216
Forwarders	1504	2164	2293
Other branches	1414	1288	1303
Upper white-collar	2545	2427	2421
Office management	3231	3306	3326
Office supervisors	2349	2523	2542
Clerical staff	1494	1708	1707
Motor-transport workers	1613	1332	1324
All occupational groups	1530	1581	1602

the time lag between the compilation of the sampling frame and the sampling date, firms have moved up or down from their original size category but have retained the weight of that stratum. This was evident in the design effects in the sample design employed by Statistics Finland (deff = 1.72). Firm-specific weights have two kinds of effects. Firstly, they lessen the above-mentioned frame-ageing problems by taking the actual size measure into account. Secondly, they introduce a clustering effect, which results in positive intra-class correlation. Therefore, the use of stratified sampling design with cluster-wise varying weights increases the standard errors of average salaries and, accordingly, the design effects.

Conclusions

This case study illustrated a data-collection strategy of mixed type. The target population of business firms comprised two subpopulations: the registered members of the employer's confederation and firms not registered. For producing reliable salary statistics, the information on paid salaries of each firm is needed, thus influencing a strong response burden on the business population. Here the main share of data was gathered from the available census-type administrative register. From the rest of the firms or from the population of unregistered firms, Statistics Finland selected a sample applying stratified simple random sampling using firms as sampling units. Thus, only sampled firms of the total business population should fill in a questionnaire. This procedure minimized the additional response burden created by this kind of survey. On the other hand, the data collected by this design should be analysed very carefully, as we showed, under different estimation strategies.

The relatively high design-effect estimates of the clustered designs ($2.10 \leq$ deff < 2.58) lend further support to the argument that there is a considerable clustering effect that should be taken into account in the calculation of average salaries in business firms. Clustering effect here means that employees working in a certain occupation within the same firm (say, shop assistants) have more or less the same salary, whereas their salary is clearly different from the average pay for their occupation in other firms. This observation also supports the view that the calculation of average salaries should use weights at the cluster level. Another factor that speaks in favour of cluster-level weights is the wide range of variation in firm (cluster) size. The most natural way to do this is to apply *Horvitz–Thompson* estimators. Recent developments in business survey methodology are summarized in Cox *et al.* (1995).

9.3 MODEL SELECTION IN A SOCIOECONOMIC SURVEY

We demonstrate in this case study not only that accounting for the clustering effect is crucial but also that the model formulation and assumptions on the

predictors can be important. For this, we use the generalized weighted least squares (GWLS) and pseudolikelihood (PML) methods introduced in Sections 8.3 and 8.4 for logit ANOVA and ANCOVA modelling on domain proportions. We use three analysis options in this exercise (see Section 8.2). The design-based analysis option (Option one) accounts for all the sampling-design complexities present in this case, that is, weighting and clustering. The weighted SRS option (Option 2) assumes simple random sampling but accounts for the weighting. The unweighted SRS option (Option 3) assumes simple random sampling and ignores all the sampling complexities. The study problem evaluates a sickness insurance scheme. The data make up a single selected regional stratum from the Finnish Health Security Survey sampling design, which involves clustering with households as clusters and weighting for nonresponse adjustment.

The Study Problem and the Data

An important aim of sickness insurance is to reduce differences between population subgroups in the utilization of health services, and to reduce the financial burden of illness on individuals and families. In Finland, a public sickness insurance scheme, covering the entire population, has been in force since 1964. In the 1980s, a supplemental sickness insurance scheme, supplied by private insurance companies, was increasingly used, e.g. in reimbursing in the private health-care sector, costs of visiting a physician because of sickness. We shall study variations in the proportion of privately insured persons in various income groups using data from the Finnish Health Security (FHS) Survey. The survey was conducted in 1987 by the Social Insurance Institution of Finland.

The FHS Survey was intended to produce reliable information for the evaluation of health and social security. Regionally stratified one-stage cluster sampling was used. Both substantive matters and economy of data collection motivated the use of households as the units of data collection. Of a sample of 6998 households, a total of 5858 (84%) took part in the survey. All eligible members in the sample households formed the element-level sample, consisting of a total of 16 269 interviewed non-institutionalized persons. Unit nonresponse was concentrated in urban regions, especially large towns such as Helsinki. Because of the nonignorability of the nonresponse, poststratification was used for adjusting so that the poststrata were formed by region, sex and age groups.

Personal interviews were conducted household-wise, but the main interest was on person-level inferences. It is obvious that many characteristics concerning health, use of health services and health behaviour, tend to be homogeneous within households. Owing to this, the corresponding study variables can be positively intra-cluster correlated. Design-effect estimates of means and proportions of such variables were often greater than one but less than two. The largest design-effect estimate (deff $= 1.7$) was found for a binary variable INSUR describing access to private sickness insurance.

A subsample of 2071 persons and 878 households living in the Helsinki Metropolitan Area, being one of the 35 strata, is considered in this case study. The estimated proportion of private sickness insured persons was relatively high, about 17% in the Helsinki Metropolitan Area, where the supply of private health-care services was high relative to other parts of the country. In rural areas, this proportion was noticeably smaller.

Examining the association of INSUR with household incomes was seen to be relevant to the evaluation of the public sickness insurance scheme. The preliminary analysis, however, does not lend support to the hypothesis that having private sickness insurance depends on high incomes. Estimated INSUR proportions in three household-income categories (low, medium, high) are 15.2%, 17.3% and 18.1%, respectively. In a homogeneity test on these proportions, an observed value $X_P^2 = 2.15$ of the Pearson test statistic was obtained, with a p-value 0.342, clearly indicating nonsignificant variation. Further, a logit regression with INSUR as the response and household income as the quantitative predictor, with integer scores from 1 to 3, has a p-value 0.148, indicating a nonsignificant linear trend.

But, having private health insurance depends strongly on age. Private insurance appears to be a form of sickness insurance used especially for children. In the Helsinki Metropolitan Area, 43% of children were covered, whereas the proportion for adults was only 9%. Moreover, the need to visit a physician because of a chronic or acute illness tends to increase the probability of being privately insured. Of those who had visited a doctor at least once in a given time interval, 27% had access to private sickness insurance. The proportion was 14% in the other group. Possible causal relationships (if any) can of course also work the other way round. Taking the age of the respondent and visiting a private physician as confounding factors can thus be informative when studying more closely the relationship of a household member being privately insured with the income of the household.

An ANOVA-type logit model on cross-classified data provides the simplest modelling approach for studying the association further. For simplicity, we choose the binary variables VISITS (visiting a private physician at least once during a fixed time interval), AGE (0–17-year-old child or over-17-year-old adult) and a three-category variable INCOME (household net income per OECD consumer unit, one-third parts) as the predictors in the ANOVA model. With these predictors, a total of 12 population subgroups or domains are produced. Because INCOME can also be taken as a quantitative predictor, we fit a logit ANCOVA model for these proportions to further examine the possible linear trend for household incomes.

Domain proportions of INSUR are displayed in Table 9.4. The proportions $\hat{p}_j^U = n_{j1}/n_j$ and the domain sample sums n_j of INSUR and the domain sample sizes n_j are the original unweighted quantities used under the SRS option that ignores the weighting. Under the other two options, the proportions $\hat{p}_j = \hat{n}_{j1}/\hat{n}_j$ are used, which are reweighted for the unit nonresponse. The proportion estimators are

Table 9.4 Unweighted and weighted proportion estimates $\hat{p}_j^U$ and $\hat{p}_j$ (%) of privately sickness insured persons (INSUR) by VISITS, AGE and INCOME in the Helsinki Metropolitan Area (the FHS Survey).

Domain	VISITS	AGE	INCOME	$\hat{p}_j^U$	n_j	$\hat{p}_j$	$\hat{d}_j$	$\hat{n}_j$	m_j
1	None	Child	Low	27.6	145	29.0	1.7	140	86
2			Medium	33.3	135	33.6	1.7	125	93
3			High	41.3	75	41.2	1.3	69	57
4		Adult	Low	6.7	400	6.5	1.5	422	258
5			Medium	8.9	427	8.6	1.5	425	245
6			High	11.6	423	11.3	1.6	422	256
7	Some	Child	Low	60.5	43	60.3	1.4	44	33
8			Medium	74.4	39	75.2	1.4	37	30
9			High	75.6	41	75.4	1.3	41	35
10		Adult	Low	12.6	103	12.9	1.3	110	92
11			Medium	12.5	88	11.4	1.0	87	83
12			High	11.2	152	10.5	1.3	149	127
Total sample				17.2	2071	16.8	1.8	2071	878

INSUR Access to private sickness insurance (binary response)
VISITS Visiting a private physician at least once in a given time interval
AGE Age (children 0–17 years/adults 18 years and above)
INCOME Household net income 1986/87 per OECD consumer unit (one-third parts)

thus consistent ratio estimators where $\hat{n}_{j1}$ and $\hat{n}_j$ are weighted domain sample sums and weighted domain sample sizes respectively. The design-effect estimates $\hat{d}_j$ are for the weighted proportion estimates $\hat{p}_j$. The number of sample clusters m_j, i.e. households covered by each subgroup, is also displayed.

With VISITS and AGE fixed, the INSUR proportions increase with increasing income, except in the last three income groups. The proportions tend to be larger on average in the second VISITS group and in the first AGE group. The largest proportions are for children with at least one doctor's visit. The design-effect estimates indicate a slight clustering effect; their average is 1.4.

Methods

A logit ANOVA model is first fitted by the GWLS method to the INSUR proportions $\hat{p}_j$ and $\hat{p}_j^U$ with VISITS, AGE and INCOME as the qualitative predictors. Then, a logit ANCOVA model is fitted by the PML method for the same table, but the predictor INCOME is taken as quantitative with scores from 1 to 3. We use the GWLS and PML methods under the three analysis options introduced in Section 8.2. Under the unweighted SRS option, all design complexities are ignored, and only the

weighting is accounted for under the weighted SRS option. Under the design-based option, the extra-binomial variation and the correlations between separate proportion estimates are allowed in addition. This option uses the actual cluster-sampling design, whereas the other two options assume simple random sampling.

There are obvious reasons for supporting design-based analysis. The response variable INSUR appears positively intra-cluster correlated in such a way that if a household member, especially a child, is insured, then the others tend to be as well. This clustering effect is indicated in the design-effect estimate deff = 1.8 of the overall INSUR proportion and in the domain design effects, which clearly indicate extra-binomial variation.

There is another important issue concerning the intra-cluster correlations with respect to the domain structure. VISITS and AGE obviously constitute cross-classes in that they cut across the clusters, i.e. the households. INCOME constitutes segregated classes because it is a household-level predictor. These predictors together thus produce a structure that is of a mixed-classes type. This causes pair-wise correlations between separate proportions $\hat{p}_j$. Not all proportions are allowed to be correlated, but only those corresponding to the respective INCOME groups, i.e. every third domain. So, in addition to the extra-binomial variation, positive covariances can be expected between the proportion estimates in these domains, also supporting the use of the design-based analysis option.

The structure of the intra-cluster correlation is reflected in the 12×12 design-based covariance-matrix estimate $\hat{\mathbf{V}}_{des}$ of domain proportions $\hat{p}_j$. This estimate is displayed in Figure 9.4, in which the corresponding binomial estimate $\hat{\mathbf{V}}_{bin}$ is

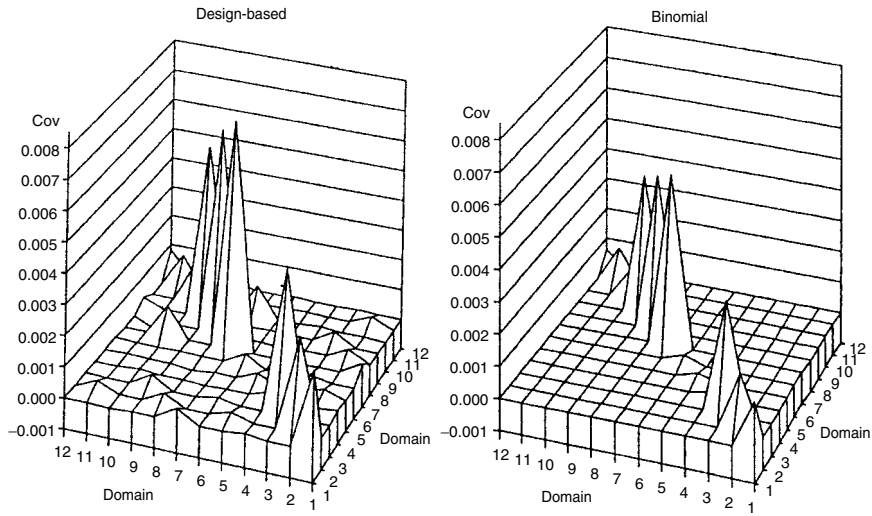


Figure 9.4 Covariance-matrix estimates for INSUR proportions $\hat{p}_j$. The design-based estimate $\hat{\mathbf{V}}_{des}$ and the binomial estimate $\hat{\mathbf{V}}_{bin}$ (the FHS Survey).

shown for comparison. The estimate $\hat{\mathbf{V}}_{des}$, obtained by the linearization method, appears quite stable owing to the large number of degrees of freedom, $f = m - H = 877$, and the condition number of $\hat{\mathbf{V}}_{des}$ is not large (37.4). It can thus be expected that the GWLS and PML methods work adequately under the design-based option. Because the variance estimates on the diagonal of $\hat{\mathbf{V}}_{des}$ are larger than the corresponding binomial variance estimates, liberal test results can be expected under the SRS options, relative to those obtained under the design-based option.

As was shown in Chapter 8, the vector of proportion estimates and its covariance-matrix estimate, depending on the analysis option considered, are required for logit modelling with the GWLS and PML methods. In the GWLS analysis, equations (8.5) to (8.13) in Section 8.3 were used, and in the PML analysis, equations (8.24) to (8.27) in Section 8.4 were used. Under the design-based option, the estimates $\hat{p}_j$ and $\hat{\mathbf{V}}_{des}(\hat{\mathbf{p}})$ were used. Under the weighted SRS option, the binomial estimate $\hat{\mathbf{V}}_{bin}(\hat{\mathbf{p}})$ was used in addition to $\hat{p}_j$. Under the unweighted SRS option, the unweighted estimates $\hat{p}_j^U$ and $\hat{\mathbf{V}}_{bin}(\hat{\mathbf{p}}^U)$ were used.

Results

Let us first consider the test results for the logit ANOVA model. We wish to study the dependence of being privately insured on incomes of the household with adjustment for the confounding effects of visiting a doctor and age of respondent. In addition to the corresponding main effects, possible interactions should be examined as well. Thus, the relevant saturated logit model is of the form $\log(\mathrm{P}/(1-\mathrm{P})) = \mathrm{V} + \mathrm{A} + \mathrm{I} + \mathrm{V}^*\mathrm{A} + \mathrm{V}^*\mathrm{I} + \mathrm{A}^*\mathrm{I} + \mathrm{V}^*\mathrm{A}^*\mathrm{I}$, where V refers to VISITS, A refers to AGE, I refers to INCOME and P stands for the domain proportions of being privately insured. Note that in this expression, all the predictors are taken to be qualitative.

An ANOVA model with all the main effects, and an interaction of VISITS and AGE, appeared to fit reasonably well and could not be further reduced. Results on goodness of fit of the model are displayed in the left-most part of Table 9.5, including the observed values of the Wald statistics based on the SRS-based and design-based Wald statistics. There is no need for F-corrections for unstability because of the large number of sample clusters. The reduced ANOVA model fits well according to the test results, under any of the analysis options.

The main interest in the analysis is the importance of the INCOME effect in the ANOVA model as a predictor of being privately insured. The Wald test results under the selected analysis options, using the statistic $X^2(\mathbf{b})$, are given in the middle part of the table. The test results indicate that under the SRS-based options, the INCOME effect clearly remains significant. The most liberal test, significant at the 1% level, is under the unweighted SRS option. Under the weighted SRS

Table 9.5 Wald-test results of goodness of fit of the logit ANOVA model, and of significance of the INCOME effect and the INCOME contrast 'low versus high', under the design-based and SRS-based analysis options (the FHS Survey).

	Model fit			Significance of INCOME effect			Significance of contrast low vs high
Option	X^2	df	p-value	$X^2(\mathbf{b})$	df	p-value	p-value
Option 1	4.23	6	0.6450	4.35	2	0.1138	0.0372
Option 2	4.52	6	0.6063	7.95	2	0.0188	0.0048
Option 3	3.61	6	0.7290	9.31	2	0.0095	0.0023

Option 1: Design-based analysis under the actual cluster-sampling design
Option 2: Simple random sampling assumption, weighted analysis
Option 3: Simple random sampling assumption, unweighted analysis

option, the test is significant at the 5% level. In both of these tests, the clustering effect is ignored. But, the INCOME effect turns out to be nonsignificant as soon as the extra-binomial variation and the correlations of the domain proportions are accounted for using the design-based option. Then, the INCOME effect becomes nonsignificant even at the 10% level.

For more detailed inferences, we separately test the hypothesis that the model parameters for the low and high INCOME groups were equal. The test results for the corresponding contrast 'low versus high' are given in the right-most part of Table 9.5. All the tests indicate a significant difference at least at the 5% level, and the pattern of the p-value follows that of the previous tests.

We next calculate the corresponding adjusted odds ratios and their 95% confidence intervals using the estimated model coefficients and their standard errors (Table 9.6). This is done for the two extreme options 1 and 3.

Under both options, the adjusted odds ratios for the first INCOME group differ significantly (at the 5% level) from one, which is the odds ratio for the highest INCOME group.

The results from the logit ANOVA model give some support to the conclusion that access to private sickness insurance might not be equally likely in the two extreme income groups, although the overall effect of household incomes appeared nonsignificant when the clustering effect was accounted for. It is thus reasonable to model the variation further so that the possible linear trend in the proportions in the INCOME groups, adjusted for the confounding factors, can be tested more explicitly. This is carried out by a logit ANCOVA model, where INCOME is taken as a quantitative predictor so that integer scores from 1 to 3 are assigned to the classes. Hence, we increase the use of the information inherent in the variable INCOME.

A logit ANCOVA model is fitted by the PML method. A model with identical model terms as in the previous ANOVA model appears reasonable for further

Table 9.6 Adjusted odds ratio statistics for INSUR under the design-based analysis option and the unweighted SRS option (the FHS Survey).

		95% confidence interval for OR	
Option	Odds ratio	Lower	Upper
Option 1			
INCOME class			
1	1	1	1
2	1.22	0.81	1.85
3	1.56	1.03	2.38
Option 3			
INCOME class			
1	1	1	1
2	1.23	0.91	1.69
3	1.64	1.19	2.22

Option 1: Design-based analysis under the actual cluster-sampling design
Option 3: Simple random sampling assumption, unweighted analysis

examination. Let us consider more closely the test results on the regression coefficient b_4 for INCOME in this model. The results obtained under the design-based and unweighted SRS are given in Table 9.7. In fact, the unweighted SRS results are based on the ML method because the weighting is ignored. In the table, the t-test results under both options indicate significant deviation from zero (at least at the 5% level) for the regression coefficient of INCOME. Here also, the SRS-based test is liberal relative to the design-based test. The test result under the weighted SRS option would be intermediate. Note also that the estimates $\hat{b}_4$ somewhat differ; under the weighted SRS option, an equal estimate to the design-based counterpart would have been obtained.

Summary

We studied whether access to private sickness insurance depends on household incomes when the confounding effects of visiting a private physician and age of respondent are adjusted for. For the analysis, the data were arranged in a multidimensional table of domain proportions. The proportions indicated slight clustering effects. Logit ANOVA modelling provided the simplest approach to studying the variation of the proportions. The effect of household incomes appeared

Table 9.7 Estimation and test results on the regression coefficient b_4 for INCOME in a logit ANCOVA model fitted by the PML method under the design-based and unweighted SRS options (the FHS Survey).

Option	$\hat{b}_4$	$\hat{d}(\hat{b}_4)$	s.e $(\hat{b}_4)$	t-test	p-value
Option 1	0.229	1.77	0.109	2.10	0.0357
Option 3	0.246	1.00	0.081	3.02	0.0026

Option 1: Design-based analysis under the actual cluster-sampling design
Option 3: Simple random sampling assumption, unweighted analysis

significant when the clustering effects were ignored, but it lost its significance when these effects were accounted for. In the test of a contrast, and in the odds ratio estimates, some evidence, however, was present on differences between the extreme income groups with respect to the coverage of private sickness insurance, thus supporting the need for further modelling. A logit ANCOVA model, where a linear trend on household incomes was more explicitly tested, provided results giving more evidence of having access to private insurance depending on high incomes. This result indicates that a private insurance scheme, as a supplement to a public insurance scheme, can involve inequality with respect to access to, and use of, health-care services.

In the preceding analysis, the variable describing access to a private sickness insurance scheme was the binary response. This was used mainly for illustrative purposes; the intra-cluster correlation of that variable was relatively strong. It would also be reasonable to take the variable describing use of health services as the response, with the insurance variable as one of the predictors. Then, a different view of the problem would be possible.

Methodological Conclusion

Positive intra-cluster correlation of a response variable can severely distort the test results in a multivariate analysis even if the correlations were relatively weak, as in the case demonstrated. In both logit ANOVA and ANCOVA modelling, ignoring the clustering effects resulted in overly liberal tests relative to those in which the clustering effects were properly accounted for. This was because the standard errors of model coefficients were underestimated by ignoring the clustering effects. Hence the results indicate a warning against relying on results from standard analyses when working with data from a clustered design. For the nuisance approach, which appeared to be relevant in the analysis considered, the design-based methods using least-squares or likelihood-based estimation with element weights provide a safe and easily manageable approach for modelling intra-cluster correlated responses. There also, the results should be carefully

compared with alternative model formulations in order to reach valid inferences on the subject matter.

9.4 MULTI-LEVEL MODELLING IN AN EDUCATIONAL SURVEY

Multi-level modelling on hierarchically structured data with a continuous response variable is used in a study problem concerning students' literacy in a multinational educational survey. Cluster sampling has been used with schools as clusters, reflecting the hierarchical structure of the population. The sampling design introduces strong intra-cluster correlation for the response variable, and this is a property that should be taken into account in the analysis. The disaggregated approach introduced here provides an alternative to the methods for the nuisance or aggregated approach, which is the main approach in this book. We apply the disaggregated approach by fitting a two-level linear model separately for data from a number of countries. The results are also compared with those from an analysis ignoring the design complexities.

PISA: An International Educational Survey

The data are from the OECD's Programme for International Student Assessment (PISA). The first PISA Survey was conducted in 2000 in 28 OECD member countries and 4 non-OECD countries. The PISA 2000 Survey covered three subject-matter areas: reading literacy, mathematical literacy and scientific literacy. We discuss here the area of reading literacy. We selected from the PISA database the following countries: Brazil, Finland, Germany, Hungary, Republic of Korea, United Kingdom and United States. Our selection of countries is deliberate; countries with varying clustering effects were chosen, keeping, however, in mind a good regional representativeness. The survey data set from these 7 countries comprised a total of 1388 schools and 32 101 pupils.

A highly standardized survey design was used in the PISA 2000 Survey, including standardization of basic concepts, procedures and tools, such as measurement instruments, sampling design, data-collection procedures and estimation and analysis procedures. This was to guarantee as far as possible the international comparability of results.

Sampling of Schools and Students

In the sampling design for an educational survey, it is natural to utilize the existing administrative and functional structures of the school system. There, the schools can be taken as basic units, which are grouped by areas of school administration or

similar administrative criteria. On the other hand, the teaching is organized into teaching groups or school classes, composed of the students and the teacher. In educational surveys, a school is often taken as the primary unit of data collection because of economical and other practical reasons. From the sampled schools, students are selected as the secondary units. There is thus a natural hierarchy in the population, which is a property that is utilized both in the sampling design and in the modelling procedures for this case study.

Stratified two-stage cluster sampling was used in most PISA countries. The first stage consisted of sampling individual schools in which 15-year-old students were enrolled. Schools were sampled with systematic PPS sampling (see Section 3.2), the measure of size being a function of the estimated number of eligible (15-year-old) students enrolled. In most cases, the population of schools was stratified before sampling operations. A minimum of 150 schools was selected in each country (where this number existed), although the requirements for national analyses often required a somewhat larger sample.

In the second stage, samples of students were selected within the sampled schools. Once the schools were selected, a frame list of each sampled school's 15-year-old students was prepared. From this list, 35 students were then selected with equal probability. All 15-year-old students were selected if fewer than 35 were enrolled.

A minimum response rate of 85% was required for the schools initially selected. A minimum participation rate of 80% of students within participating schools was required. This minimum participation rate had to be met at the national level, not necessarily by each participating school (OECD 2001, 2002a).

Weighting Schemes

Appropriate sampling weights were constructed for each national sample data set. The element weight consisted of factors reflecting school selection probabilities, student selection probabilities within schools and school and student nonresponse adjustments. For each country, the weight w_{ik} for student k in school i can be expressed as follows:

$$w_{ik} = w_{1i} \times w_{2ik} \times f_i, \qquad i = 1, \ldots, m \text{ and } k = 1, \ldots, n_i,$$

where

$w_{1i} = 1/(\pi_i \hat{\theta}_i)$ is the reciprocal of the product of the inclusion probability π_i and the estimated participation probability $\hat{\theta}_i$ of school i;

$w_{2ik} = 1/(\pi_{k|i} \hat{\theta}_{k|i})$ is the reciprocal of the product of the conditional inclusion probability $\pi_{k|i}$ and estimated conditional response probability $\hat{\theta}_{k|i}$ of student k from within the selected school i;

f_i is an adjustment factor for school i to compensate any country-specific refinements in the survey design, and

m is the number of sample schools in a given country and n_i is the number of sample students in school i.

The student-level element weights, rescaled to sum up to the actual size of the available sample data set in each country, were used in the analyses. In a given country, the mean of the rescaled weights is one, but there are differences between countries in the variation of the weights. The smallest standard deviation of the rescaled weights is 0.143 and the largest is 0.983. A more detailed description of weighting procedures is given in OECD (2002b).

Reading Literacy in Selected Countries

The outcome variable y is the student's combined reading literacy score (or to be exact, the first of five plausible values of combined reading literacy), scaled so that the common mean over the participating OECD countries is 500 and the standard deviation is 100. We call the response variable the combined reading literacy score. Descriptive statistics on reading literacy in the selected countries are presented in Table 9.8. Means and standard errors of the combined reading literacy score have been calculated by techniques presented in Chapter 5. Therefore, the estimates are design-based and account properly for the complexities (weighting, stratification and clustering) of the sampling design used in a given country. There are two different design effects in the table. The overall design effect accounts for weighting, stratification and clustering. The second design effect

Table 9.8 Descriptive statistics for combined reading literacy score in the PISA 2000 Survey by country (in alphabetical order).

	Combined reading literacy score					Number of observations in data set	
Country	Mean	Standard error	Overall design effect	Design-effect accounting for stratification and clustering	Effective sample size of students	Students	Schools
Brazil	402.9	3.82	8.33	5.17	476	3961	290
Finland	550.7	2.15	2.79	2.74	1600	4465	147
Germany	497.4	5.68	13.47	11.68	305	4108	183
Hungary	485.7	6.02	20.00	16.20	231	4613	184
Republic of Korea	526.6	3.66	12.99	11.67	351	4564	144
United Kingdom	531.4	4.08	14.08	7.16	564	7935	328
United States	517.0	5.16	6.93	5.46	354	2455	112

Data source: OECD PISA database, 2001.

accounts for stratification and clustering and allows for a comparison with the weighted SRS analysis option. Both design effects indicate a strong clustering effect for most countries. In some cases, the difference between the first and second design-effect estimates is substantial, indicating a large variation in the weights.

The effective sample sizes of students are calculated by dividing the number of students by the overall design effect. The effective sample size is the equivalent sample size needed to achieve the same precision in estimation if simple random sampling from a student population without any clustering were used. If the observations are not independent from each other, the effective sample size decreases: the higher the design effect, the smaller the effective sample size. Though the nominal sample sizes of students are large (several thousands) in all countries, some of the effective sample sizes are quite small (only a few hundred).

Design-effect estimates also indicate that standard errors calculated under an erroneous assumption of simple random sampling would be much smaller than the design-based standard error estimates for most countries.

Fitting a Two-level Hierarchical Linear Model

In the analysis, the outcome variable y is the combined reading literacy score. The variation of the outcome variable is explained with two school-level and four student-level variables. The school-level explanatory variables are school size (SSIZE) and teacher autonomy (AUTONOMY). School size is a measure formed from the actual number of students in the school, divided by 100. School principals were asked to report who had the main responsibility for several tasks in the school. Teacher autonomy was derived from the number of categories that principals identified as being mainly the responsibility of teachers. Both variables were standardized so that the common mean over the participating OECD countries was zero and the standard deviation was one.

The student-level explanatory variables are student's gender (recoded so that one is for females and zero is for males, and named FEMALE), socioeconomic background (SEB), engagement in reading (ENGAGEMENT) and achievement press (ACHPRESS). The index of SEB was derived from students' responses on parental occupation. The index of engagement in reading was derived from students' level of agreement with several statements concerning reading habits and attitudes, and the index of achievement press was derived from students' reports of the pressure they feel from their teacher. These three indices were again standardized so that the common mean over the participating OECD countries was zero and the standard deviation was one.

The two-level regression model for the combined reading literacy score y, with explanatory variables and random variation at both levels, is given by

$$\begin{aligned} y_{ik} = {} & \text{INTERCEPT} + \gamma_1 \times \text{SSIZE}_i + \gamma_2 \times \text{AUTONOMY}_i \\ & + \beta_1 \times \text{FEMALE}_{ik} + \beta_2 \times \text{SEB}_{ik} + \beta_3 \times \text{ENGAGEMENT}_{ik} \\ & + \beta_4 \times \text{ACHPRESS}_{ik} + u_i + e_{ik}, \end{aligned}$$

where the index k refers to the level-1 unit (student) and i to the level-2 unit (school). The fixed effects γ and β denote regression coefficients of the school- and student-level variables respectively. Residual u_i is the random effect of school i assumed normally distributed with mean zero and variance σ_u^2, whereas e_{ik} is the student-level residual assumed normally distributed with mean zero and variance σ_e^2. The random effects u_i and e_{ik} are assumed independent. The student-level rescaled weights were used in the analyses.

Units within naturally existing clusters, such as schools, tend to be more similar or homogeneous with respect to the variable of interest than units selected at random from the population. This means that the level-1 units (students) cannot be assumed statistically independent within schools, and the study variable tends to be positively intra-cluster correlated. In the context of multi-level modelling, the intra-cluster correlation is estimated by (Skinner *et al.* 1989; Goldstein 2002; Snijders and Bosker 2002) as

$$\hat{\rho}_{\text{int}} = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \hat{\sigma}_e^2} = \frac{\hat{\sigma}_u^2}{\hat{\sigma}^2},$$

where the estimated total variance $\hat{\sigma}^2$ of the study variable is divided into two components, the between-school variance $\hat{\sigma}_u^2$ and the within-school variance $\hat{\sigma}_e^2$. The intra-cluster correlation coefficient measures the pair-wise correlation between values of level-1 units (students) in the same level-2 group (school) and is called the intra-school correlation coefficient. In a model-based context, the coefficient is estimated from the variance components of the null model, i.e. the multi-level model with only intercept and residuals at both levels. For example, the estimated intra-school correlation coefficient for Hungary in Table 9.9 is $6093.7/(6093.7 + 3148.3) = 0.659$. The coefficient can also be estimated from the variance components of the model including explanatory variables, in which case it is called the residual intra-school correlation coefficient. The residual intra-school correlation coefficient for Hungary in Table 9.10 is $4744.2/(4744.2 + 2897.4) = 0.621$. Note that the concept of intra-cluster correlation is used in a design-based context earlier in this book (see Section 3.2).

Variance components were estimated by restricted maximum likelihood (REML), and the fixed effects were estimated by generalized least squares (GLS) given these variance estimates (Bryk and Raudenbush 1992). These estimates are accompanied by standard error estimates that account for the clustering effect (see, for example, the 'sandwich' form in Section 8.4).

Table 9.9 Estimates of two-level variance component models (null models) for combined reading literacy score in the PISA 2000 Survey by country (ordered by the size of the estimated intra-school correlation coefficient).

Country	Intra-school correlation coefficient	Variance components		Intercept	Standard error
		School level	Student level		
Hungary	0.659	6093.7	3148.3	464.1	5.84
Germany	0.553	5572.2	4507.8	496.1	5.61
Brazil	0.428	3146.9	4201.4	387.9	3.61
Republic of Korea	0.375	1828.6	3043.0	520.9	3.74
United States	0.241	2318.2	7315.5	503.3	4.97
United Kingdom	0.212	1917.5	7126.5	529.0	2.88
Finland	0.063	470.7	6960.9	550.6	2.18

Data source: OECD PISA database, 2001.

Table 9.9 presents results for basic two-level variance component models, i.e. null models without explanatory variables. In these models, one fixed effect, the intercept, and the school-level random intercepts are estimated. The total variance is divided into between-schools and within-schools variance components, which are used to calculate the intra-school correlation coefficient. Estimated coefficients vary considerably between the selected countries, with a minimum value of 0.063 and a maximum value of 0.659.

In a given country, the intercept in Table 9.9 is the estimated average of school intercepts. The intercepts are somewhat different from the country means in Table 9.8. Standard error estimates of estimated intercepts are also different because they are calculated using the estimated multi-level model.

Estimated two-level models for combined reading literacy score are presented in Table 9.10. In school-level variables, the effect of school size is statistically significant in some countries. The second school-level variable, teacher autonomy, does not have statistically significant effects in any of the countries.

In student-level explanatory variables, the effects of socioeconomic background and engagement in reading are statistically significant at least at the 5% level in every country. The effect of socioeconomic background varies greatly between countries. The higher the socioeconomic background score, and the more he or she is engaged in reading, the better tends to be his or her reading proficiency score. The strength and direction of the effect of achievement press varies greatly. In most cases, the gender effect was statistically significant.

The estimated models explain a considerable amount of school- and student-level variation in reading literacy as is indicated by the proportional reduction figures. However, there is substantial variation in the degree of reduction gained by the fitted model, when compared to the null model. In most countries, the

Table 9.10 Estimates of two-level models for combined reading literacy score in the PISA 2000 Survey by country.

		Hungary	Germany	Brazil	Republic of Korea	United States	United Kingdom	Finland
Fixed effects:								
Coefficient								
Intercept	γ_0	471.2	496.4	382.0	506.8	496.6	524.9	531.6
	s.e	6.36	4.58	4.56	6.29	6.05	3.38	4.91
	t-test	74.14	108.37	83.75	80.53	82.12	155.06	108.27
	p-value	0.000	0.000	0.000	0.000	0.000	0.000	0.000
School-level variables:								
	γ_1	30.6	27.4	2.4	7.1	1.0	3.8	5.9
School size	s.e	9.00	9.22	1.47	3.44	2.54	3.14	7.35
	t-test	3.41	2.97	1.64	2.07	0.38	1.20	0.80
	p-value	0.001	0.003	0.100	0.039	0.705	0.232	0.426
Teacher	γ_2	4.8	−7.1	−3.1	2.5	4.1	−2.3	2.8
autonomy	s.e	5.62	5.22	4.24	5.39	3.63	2.61	2.68
	t-test	0.86	−1.37	−0.74	0.47	1.14	−0.89	1.06
	p-value	0.392	0.171	0.459	0.641	0.256	0.374	0.291
Student-level variables:								
Female	β_1	6.4	3.6	3.1	15.9	14.9	9.8	19.6
	s.e	2.22	2.41	2.54	2.49	3.71	2.64	2.43
	t-test	2.89	1.50	1.21	6.38	4.00	3.71	8.09
	p-value	0.004	0.133	0.228	0.000	0.000	0.000	0.000
Socioeconomic	β_2	6.0	11.5	9.9	2.2	16.7	23.3	15.8
background	s.e	1.09	1.53	1.35	0.92	2.22	1.32	1.34
	t-test	5.56	7.50	7.34	2.40	7.51	17.70	11.78
	p-value	0.000	0.000	0.000	0.016	0.000	0.000	0.000
Engagement in	β_3	19.5	19.0	19.5	16.6	28.9	31.5	33.9
reading	s.e	1.04	0.98	1.51	1.04	1.99	1.40	1.26
	t-test	18.68	19.36	12.87	15.94	14.49	22.59	27.05
	p-value	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Achievement	β_4	0.9	−1.6	3.4	3.4	−3.3	−7.2	−3.7
press	s.e	0.93	1.16	1.44	0.89	2.04	1.59	1.40
	t-test	0.92	−1.35	2.36	3.85	−1.62	−4.52	−2.65
	p-value	0.356	0.176	0.018	0.000	0.106	0.000	0.008
Random effects:								
Variance component								
School level		4744.2	3501.6	2730.5	1387.3	1770.6	999.6	394.8
Student level		2897.4	3981.9	3830.6	2809.6	6094.1	5779.0	4984.3
Residual intra-school correlation coefficient		0.621	0.468	0.416	0.331	0.225	0.147	0.073
Proportional reduction in variance components, compared to null model (%)								
School level		22.1	37.2	13.2	24.1	23.6	47.9	16.1
Student level		8.0	11.7	8.8	7.7	16.7	18.9	28.4
Total		17.3	25.8	10.7	13.8	18.4	25.0	27.6

Data source: OECD PISA database, 2001.

unexplained school-level variation is still large, compared to the unexplained total variation, which can be seen from the residual intra-school correlation coefficient figures.

Only linear effects of explanatory variables were included in the models. The possible quadratic effects could also be studied for some variables (e.g. school size). All the coefficients of the level-1 explanatory variables are also considered as fixed effects, although there may exist between-school variation in the coefficients, in which case also random coefficient regression models could be used.

Comparison with Weighted SRS Analysis

We finally compare the results of the multi-level modelling exercise with those obtained ignoring the clustering effects. We use the weighted SRS analysis option (see Section 8.2) corresponding to an assumption of independence of the observations. Under this option, a fixed-effects linear model is fitted for the outcome variable, using similar explanatory variables as for the two-level model. Estimation under the weighted SRS option uses the weighted least squares method (see Section 8.4). We selected the German data for comparison (Table 9.11).

The response variable in the German data is highly intra-school correlated, and, as a consequence, the standard-error estimates of the estimated fixed level-2 effects are too small in the model fitted under the weighted SRS option. One of the two school-level effects, teacher autonomy, would be mistakenly considered as statistically significant if the weighted SRS analysis option were used, and the effect of school size would be estimated as being too small. From the level-1 explanatory variables, the effects of socioeconomic background and engagement in reading are much larger compared to the estimates from the two-level model. Achievement press would also appear as a statistically significant effect.

Summary

This case study shows that for data obtained by cluster sampling, an analysis assuming independent observations may be grossly misleading, since the positive intra-cluster correlation of observations will be ignored. Only if the clustering effect were not indicated would the results of an analysis with a two-level model and a weighted SRS-based analysis be similar.

We used here a 'disaggregated' approach in which the hierarchical structure of the population was explicitly modelled by a two-level model. An alternative way to analyse hierarchically structured data is to use design-based methods, as described in Chapter 8. There, instead of modelling the hierarchical structure, the clustering effect induced by the data structure was considered as a nuisance.

Table 9.11 Comparison of estimated coefficients of a two-level model for combined reading literacy score and a fixed-effects model fitted under the weighted SRS analysis option (the German data are used as an example).

Coefficient		Two-level model	Weighted SRS option
Intercept	γ_0	496.4	497.5
	s.e	4.58	1.93
	t-test	108.37	258.08
	p-value	0.000	0.000
School size	γ_1	27.4	20.1
	s.e	9.22	1.74
	t-test	2.97	11.52
	p-value	0.003	0.000
Teacher autonomy	γ_2	−7.1	−7.3
	s.e	5.22	1.38
	t-test	−1.37	−5.26
	p-value	0.171	0.000
Female	β_1	3.6	3.3
	s.e	2.41	2.74
	t-test	1.50	1.20
	p-value	0.133	0.229
Socioeconomic background	β_2	11.5	31.5
	s.e	1.53	1.38
	t-test	7.50	22.9
	p-value	0.000	0.000
Engagement in reading	β_3	19.0	28.9
	s.e	0.98	1.17
	t-test	19.36	24.6
	p-value	0.000	0.000
Achievement press	β_4	−1.6	−4.7
	s.e	1.16	1.31
	t-test	−1.35	−3.64
	p-value	0.176	0.000

Data source: OECD PISA database, 2001.

Thus, in a design-based analysis, we try to 'clean out' the clustering effect from the estimation and testing results to obtain valid inferences.

From a substance matter point of view, the extra contribution of multi-level modelling is that it provides explicit information about the differences between clusters, and thus more information is obtained for the interpretation of the results.

References

Bean J. A. (1975) Distribution and properties of variance estimators for complex multistage probability samples *Vital and Health Statistics* Series 2, No. 65.

Biemer P. P., Groves R. M., Lyberg L. E., Mathiowetz N. A. and Sudman S. (eds) (1991) *Measurement Errors in Surveys* Chichester: Wiley.

Biemer P. P. and Lyberg L. E. (2003) *Introduction to Survey Quality* New York: Wiley.

Binder D. A. (1983) On the variances of asymptotically normal estimators from complex surveys *International Statistical Review* **51** 279–292.

Binder D. A. (1991) A framework for analyzing categorical survey data with non-response *Journal of Official Statistics* **7** 393–404.

Binder D. A. (1992) Fitting Cox's proportional hazards models from survey data *Biometrika* **79** 139–147.

Breslow N. E. and Clayton D. G. (1993) Approximate inference in generalized linear mixed models *Journal of the American Statistical Association* **88** 9–25.

Brewer K. R. W. (1963) A model of systematic sampling with unequal probabilities *Australian Journal of Statistics* **5** 5–13.

Brewer K. R. W. and Hanif M. (1983) *Sampling with Unequal Probabilities* New York: Springer.

Brier S. S. (1980) Analysis of contingency tables under cluster sampling *Biometrika* **67** 591–596.

Bryk A. S. and Raudenbush S. W. (1992) *Hierarchical Linear Models: Applications and Data Analysis Methods* Newbury Park: Sage Publications.

Chambers R. and Skinner C. (eds) (2003) *Analysis of Survey Data* Chichester: Wiley.

Clayton D., Spiegelhalter D., Dunn G. and Pickles A. (1998) Analysis of longitudinal binary data from multiphase sampling *Journal of the Royal Statistical Society, B* **60** 71–87.

Cochran W. G. (1977) *Sampling Techniques* Third Edition. New York: Wiley.

Couper M., Baker R., Bethlehem J., Clark C., Martin J., Nicholls II W. and O'Reilly J. (eds) (1998) *Computer Assisted Survey Information Collection* New York: Wiley.

Cox B. G., Binder D. A., Chinnappa B. N., Christiansson A., Colledge M. J. and Kott P. S. (eds) (1995) *Business Survey Methods* New York: Wiley.

Datta G. S., Lahiri P., Maiti T. and Lu K. L. (1999) Hierarchical Bayes estimation of unemployment rates for the states of the U.S. *Journal of the American Statistical Association* **94** 1074–1082.

Practical Methods for Design and Analysis of Complex Surveys Risto Lehtonen and Erkki Pahkinen

ISBN: 0-470-84769-7

Dempster A. P., Rubin D. B. and Tsutakawa R. K. (1981) Estimation in covariance component models *Journal of the American Statistical Association* **76** 341–353.

Deville J.-C. and Särndal C. E. (1992) Calibration estimators in survey sampling *Journal of the American Statistical Association* **87** 376–382.

Deville J.-C., Särndal C. E. and Sautory O. (1993) Generalized raking procedures in survey sampling *Journal of the American Statistical Association* **88** 1013–1020.

Diggle P. J., Heagerty P. J., Liang K.-Y. and Zeger S. L. (2002) *Analysis of Longitudinal Data* Second Edition Oxford: Oxford University Press.

Dillman D. (1999) *Mail and Internet Surveys: The Tailored Design Method* Second Edition New York: Wiley.

Efron B. (1982) *The Jackknife, The Bootstrap and Other Resampling Plans* Philadelphia: Society for Industrial and Applied Mathematics.

Estevao V., Hidiroglou M. A. and Särndal C.-E. (1995) Methodological principles for a generalized estimation system at Statistics Canada *Journal of Official Statistics* **11** 181–204.

Estevao V. M. and Särndal C.-E. (1999) The use of auxiliary information in design-based estimation for domains *Survey Methodology* **25** 213–221.

Feder M., Nathan G. and Pfeffermann D. (2000) Multilevel modelling of complex survey longitudinal data with time varying random effects *Survey Methodology* **26** 53–65.

Federal Committee on Statistical Policy (2001) Measuring and Reporting Sources of Error in Surveys *Statistical Policy Working Paper 31*, Washington DC: Statistical Policy Office, Office of Management and Budget.

Fellegi I. P. (1980) Approximate tests of independence and goodness of fit based on stratified multistage samples *Journal of the American Statistical Association* **75** 261–268.

Francisco C. A. and Fuller W. A. (1991) Quantile estimation with a complex survey design. *Annals of Statistics* **19** 454–469.

Frankel M. R. (1971) *Inference from Survey Samples* Ann Arbor: Institute for Social Research, The University of Michigan.

Freeman D. H. (1988) Sample survey analysis: analysis of variance and contingency tables. In: Krishnaiah P. R. and Rao C. R. (eds) *Handbook of Statistics 6. Sampling*. Amsterdam: North Holland, 415–426.

Ghosh M. (2001) Model-dependent small area estimation: theory and practice. In: Lehtonen R. and Djerf K. (eds) *Lecture Notes on Estimation for Population Domains and Small Areas* Helsinki: Statistics Finland Reviews 2001/5 51–108.

Ghosh M. and Natarajan K. (1999) Small area estimation: a Bayesian perspective. In: Ghosh S. (ed.) *Multivariate Analysis, Design of Experiments, and Survey Sampling* New York: Marcel Dekker, 69–92.

Ghosh M., Natarajan K., Stroud T. W. F. and Carlin B. (1998) Generalized linear models for small area estimation *Journal of the American Statistical Association* **93** 273–282.

Ghosh M. and Rao J. N. K. (1994) Small area estimation: an appraisal *Statistical Science* **9** 55–93.

Glynn R. J., Laird N. M. and Rubin D. B. (1993) Multiple imputation in mixture models for nonignorable nonresponse with follow-ups *Journal of the American Statistical Association* **88** 984–993.

Goldstein H. (1987) *Multilevel Models in Educational and Social Research* London: Griffin.

Goldstein H. (1991) Nonlinear multilevel models, with an application to discrete response data *Biometrika* **78** 45–51.

Goldstein H. (2002) *Multilevel Statistical Models* Third Edition London: Edward Arnold.

Goldstein H. and Rasbash J. (1992) Efficient computational procedures for the estimation of parameters in multilevel models based on iterative generalized least squares *Computational Statistics and Data Analysis* **13** 63–71.

Grizzle J. E., Starmer C. F. and Koch G. G. (1969) Analysis of categorical data by linear models *Biometrics* **25** 489–504.

Groves R. M. (1989) *Survey Errors and Survey Costs* New York: Wiley.

Groves R. M., Dillman D. A., Eltinge J. L. and Little R. J. A. (2001) *Survey Nonresponse* New York: Wiley.

Hansen M. H. and Hurwitz W. N. (1943) On the theory of sampling from a finite population *Annals of Mathematical Statistics* **14** 333–362.

Hedayat A. S. and Sinha B. K. (1991) *Finite Population Sampling* New York: Wiley.

Heliövaara M., Aromaa A., Klaukka T., Knekt P., Joukamaa M. and Impivaara O. (1993) Reliability and validity of interview data on chronic diseases *Journal of Clinical Epidemiology* **46** 181–191.

Hidiroglou M. A. and Rao J. N. K. (1987a) Chi-squared tests with categorical data from complex surveys: Part I *Journal of Official Statistics* **3** 117–132.

Hidiroglou M. A. and Rao J. N. K. (1987b) Chi-squared tests with categorical data from complex surveys: Part II *Journal of Official Statistics* **3** 133–140.

Holt D., Scott A. J. and Ewings P. D. (1980) Chi-squared tests with survey data *Journal of the Royal Statistical Society, A* **143** 303–320.

Holt D. and Smith T. M. F (1979) Post stratification *Journal of the Royal Statistical Society, A* **142** 33–46.

Holt D., Smith T. M. F. and Tomberlin T. J. (1979) A model-based approach to estimation for small subgroups of population *Journal of the American Statistical Association* **74** 405–410.

Horton N. J. and Lipsitz S. R. (1999) Review of software to fit generalized estimating equation regression models *The American Statistician* **53** 160–169.

Horvitz D. G. and Thompson D. J. (1952) A generalization of sampling without replacement from a finite universe *Journal of the American Statistical Association* **47** 663–685.

Judkins D. (1990) Fay's method for variance estimation *Journal of Official Statistics* **6** 223–240.

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

Keyfitz N. (1957) Estimates of sampling variance where two units are selected from each stratum *Journal of the American Statistical Association* **52** 503–510.

Kish L. (1962) Studies of interviewer variance for attitudinal variables *Journal of the American Statistical Association* **57** 92–115.

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

Kish L. (1992) Weighting for unequal P_i *Journal of Official Statistics* **8** 183–200.

Kish L. (1995) Methods for design effects *Journal of Official Statistics* **11** 55–77.

Kish L. and Frankel M. R. (1970) Balanced repeated replications for standard errors *Journal of the American Statistical Association* **65** 1071–1094.

Kish L. and Frankel M. R. (1974) Inference from complex samples (with discussion) *Journal of the Royal Statistical Society, B* **36** 1–37.

Koch G. G., Freeman D. H. and Freeman J. L. (1975) Strategies in the multivariate analysis of data from complex surveys *International Statistical Review* **43** 59–78.

Korn E. L. and Graubard B. I. (1999) *Analysis of Health Surveys* New York: Wiley.

Krewski D. and Rao J. N. K. (1981) Inference from stratified samples: properties of the linearization, Jackknife and balanced repeated replication methods *Annals of Statistics* **9** 1010–1019.

Kumar S. and Singh A. C. (1987) On efficient estimation of unemployment rates from labour force survey data *Survey Methodology* **13** 75–83.

Kuusela V. (2000) *Telephone coverage situation in Finland.* (In Finnish). Helsinki: Statistics Finland, Reviews 3/2000.

Lawson A. B., Biggeri A., Böhning D., Lesaffre E., Viel J.-F. and Bertollini R. (eds) (1999) *Disease Mapping and Risk Assessment for Public Health* Chichester: Wiley.

Levy P. S. and Lemeshow S. (1991) *Sampling of Populations: Methods and Applications* New York: Wiley.

Liang K.-Y. and Zeger S. L. (1986) Longitudinal data analysis using generalized linear models *Biometrika* **73** 13–22.

Liang K.-Y., Zeger S. L. and Qaqish B. (1992) Multivariate regression analyses for categorical data (with discussion) *Journal of the Royal Statistical Society, B* **54** 3–40.

Little R. J. A. (1991) Inference with survey weights *Journal of Official Statistics* **7** 405–424.

Little R. J. A. (1993) Post-stratification: a modeler's perspective *Journal of the American Statistical Association* **88** 1001–1012.

Little R. J. A. and Rubin D. B. (1987) *Statistical Analysis with Missing Data* New York: Wiley.

Lehtonen R. (1988) The Execution of the National Occupational Health Care Survey Helsinki: Publications of the Social Insurance Institution, Finland, M:64. (In Finnish with English summary.)

Lehtonen R. (1990) On Modified Wald Statistics (Doctoral Dissertation). Their application to a Goodness of Fit Test of Logit Models under Complex Sampling Involving Ill-Conditioning Helsinki: Publications of the Social Insurance Institution, Finland, M:74.

Lehtonen R. and Kuusela V. (1986) Statistical efficiency of the Mini-Finland health survey's sampling design. Part 5. In: Aromaa A., Heliövaara M., Impivaara O., Knekt P. and Maatela J. (eds) *The Execution of the Mini-Finland Health Survey* Helsinki, Turku: Publications of the Social Insurance Institution, Finland, ML:65. (In Finnish with English summary.)

Lehtonen R., Särndal C.-E. and Veijanen A. (2003) The effect of model choice in estimation for domains, including small domains *Survey Methodology* **29** 33–44.

Lehtonen R. and Veijanen A. (1998) Logistic generalized regression estimators *Survey Methodology* **24** 51–55.

Lehtonen R. and Veijanen A. (1999) Domain estimation with logistic generalized regression and related estimators. *Proceedings, IASS Satellite Conference on Small Area Estimation*, Riga, August 1999; Riga: Latvian Council of Science, 121–128.

Lohr S. L. (1999). *Sampling: Design and Analysis* New York: Duxbury Press.

Lundström S. and Särndal C.-E. (2002) *Estimation in the presence of Nonresponse and Frame Imperfections* Statistics Sweden. Örebro: SCB-Tryck.

Marker D. (1999) Organization of small area estimators using a generalized linear regression framework *Journal of Official Statistics* **15** 1–24.

McCarthy P. J. (1966) Replication. An approach to the analysis of data from complex surveys *Vital and Health Statistics* Series 2, No. 14.

McCarthy P. J. (1969) Pseudoreplication: further evaluation and application of the balanced half-sample technique *Vital and Health Statistics* Series 2, No. 31.

McCarthy P. J. and Snowden C. B. (1985) The bootstrap and finite population sampling *Vital and Health Statistics* Series 2, No. 95.

McCullagh P. and Nelder J. A. (1989) *Generalized Linear Models* Second Edition London: Chapman & Hall.

McCulloch C. E. and Searle S. R. (2001) *Generalized, Linear, and Mixed Models* New York: Wiley.

Morel J. G. (1989) Logistic regression under complex survey designs *Survey Methodology* **15** 203–223.

Moura F. A. S. and Holt D. (1999) Small area estimation using multilevel models *Survey Methodology* **25** 73–80.

Murthy M. N. (1957) Ordered and unordered estimators in sampling without replacement *Sankhya* **18** 379–390.

Nathan G. (1988) Inference based on data from complex sample designs. In: Krishnaiah P. R. and Rao C. R. (eds) *Handbook of Statistics 6. Sampling* Amsterdam: North Holland, 247–266.

Nelder J. A. and Wedderburn R. W. M. (1972) Generalized linear models *Journal of the Royal Statistical Society, A* **135** 370–384.

OECD (2001) *Knowledge and Skills for Life* First results from the OECD Programme for International Student Assessment (PISA) 2000. Paris: OECD.

OECD (2002a) PISA 2000 Technical Report Paris: OECD (http://www.pisa.oecd.org/).

OECD (2002b) Manual for the PISA 2000 Database Paris: OECD.

Ohlsson E. (1998) Sequential Poisson sampling *Journal of Official Statistics* **14** 149–162.

Pastinen V. (1999) *Passenger Transport Survey 1998–1999 (In Finnish)* Helsinki: Publications of the Ministry of Transport and Communications, 43/99.

Pfeffermann D. (1993) The role of sampling weights when modeling survey data *International Statistical Review* **61** 317–337.

Pfeffermann D., Skinner C. J., Goldstein H., Holmes D. J. and Rasbash J. (1998) Weighting for unequal selection probabilities in multilevel models (With discussion) *Journal of the Royal Statistical Society, B* **60** 23–40.

Plackett R. L. and Burman J. P. (1946) The design of optimum multifactorial experiments *Biometrika* **33** 305–325.

Platek R. and Särndal C.-E. (2001) Can a Statistician Deliver? (With discussion) *Journal of Official Statistics* **17**, 1–127.

Prasad N. G. N. and Rao J. N. K. (1999) On robust small area estimation using a simple random effects model *Survey Methodology* **25** 67–72.

Quenouille M. H. (1956) Notes on bias in estimation *Biometrika* **43** 353–360.

Rao J. N. K. (1997) Developments in sample survey theory: an appraisal *The Canadian Journal of Statistics* **25** 1–21.

Rao J. N. K. (1999) Some recent advances in model-based small area estimation *Survey Methodology* **25** 175–186.

Rao J. N. K. (2003) *Small Area Estimation* New York: Wiley.

Rao J. N. K., Hartley H. O. and Cochran W. G. (1962) A simple procedure of unequal probability sampling without replacement *Journal of the Royal Statistical Society, B* **24** 482–491.

Rao J. N. K. and Scott A. J. (1981) The analysis of categorical data from complex sample surveys: chi-squared tests for goodness of fit and independence in two-way tables *Journal of the American Statistical Association* **76** 221–230.

Rao J. N. K. and Scott A. J. (1984) On chi-squared tests for multiway contingency tables with cell proportions estimated from survey data *Annals of Statistics* **12** 46–60.

Rao J. N. K. and Wu C. F. J. (1985) Inference from stratified samples: second-order analysis of three methods for nonlinear statistics *Journal of the American Statistical Association* **80** 620–630.

Rao J. N. K. and Scott A. J. (1987) On simple adjustments to chi-square tests with sample survey data *Annals of Statistics* **15** 385–397.

Rao J. N. K. and Thomas D. R. (1988) The analysis of cross-classified categorical data from complex sample surveys *Sociological Methodology* **18** 213–269.

Rao J. N. K. and Wu C. F. J. (1988) Resampling inference with complex survey data *Journal of the American Statistical Association* **83** 209–241.

Rao J. N. K. and Thomas D. R. (1989) Chi-squared tests for contingency tables. In: Skinner C. J., Holt D. and Smith T. M. F. (eds) *Analysis of Complex Surveys* Chichester: Wiley, 89–114.

Rao J. N. K., Kumar S. and Roberts G. (1989) Analysis of sample survey data involving categorical response variables: methods and software (With discussion) *Survey Methodology* **15** 161–186.

Rao J. N. K. and Scott A. J. (1992) A simple method for the analysis of clustered binary data *Biometrics* **48** 577–585.

Rao J. N. K., Wu C. F. J. and Yue K. (1992) Some recent work on resampling methods for complex surveys *Survey Methodology* **18** 209–217.

Rao J. N. K. and Shao J. (1993) Jackknife variance estimation with survey data under hot deck imputation *Biometrika* **79** 811–822.

Rao J. N. K., Sutradhar B. C. and Yue K. (1993) Generalized least squares *F* test in regression analysis with two-stage cluster samples *Journal of the American Statistical Association* **88** 1388–1391.

Rao J. N. K. and Thomas D. R. (2003) Analysis of categorical response data from complex surveys: an appraisal and update. In: Chambers R. and Skinner C. (eds) *Analysis of Survey Data* Chichester: Wiley.

Roberts G., Rao J. N. K. and Kumar S. (1987) Logistic regression analysis of sample survey data *Biometrika* **74** 1–12.

Rubin D. B. (1987) *Multiple Imputation for Nonresponse in Surveys* New York: Wiley.

Rubin D. B. (1996) Multiple Imputation After 18+ Years *Journal of the American Statistical Association* **91** 473–489.

Särndal C.-E. (1996) For a better understanding of imputation. In Laaksonen S. (ed.) (1996). *International perspectives on nonresponse*. Proceedings of the sixth international workshop on household survey nonresponse. Helsinki: Statistics Finland, Research reports 219.

Särndal C.-E. (2001) Design-based methodologies for domain estimation. In: Lehtonen R. and Djerf K. (eds) *Lecture Notes on Estimation for Population Domains and Small Areas* Helsinki: Statistics Finland Reviews 2001/5 5–49.

Särndal C.-E., Swensson B. and Wretman J. (1992) *Model Assisted Survey Sampling* New York: Springer.

Satterthwaite F. E. (1946) An approximate distribution of estimates of variance components *Biometrics* **2** 110–114.

Schafer J. L. (2000) *Analysis of Incomplete Multivariate Data* New York: Chapman & Hall.

Schaible, W. L. (ed.) (1996) *Indirect Estimators in U.S. Federal Programs* New York: Springer.

Scott A. J. (1986) Logistic regression with survey data. *Proceedings of the Section on Survey Research Methods* American Statistical Association, 25–30.

Scott A. J., Rao J. N. K. and Thomas D. R. (1990) Weighted least-squares and quasilikelihood estimation for categorical data under singular models *Linear Algebra and its Applications* **127** 427–447.

Shao J. and Tu D. (1995) *The Jackknife and Bootstrap* New York: Springer.
Silva P. L. N. and Skinner C. J. (1997) Variable selection for regression estimation in finite populations. *Survey Methodology* **23**, 23–32.
Singh A. C. (1985) On Optimal Asymptotic Tests for Analysis of Categorical Data from Sample Surveys Working Paper No. SSMD 86–002, Social Survey Methods Division, Statistics Canada.
Singh M. P., Gambino J. and Mantel H. J. (1994) Issues and strategies for small area data *Survey Methodology* **20** 3–22.
Singh A. C., Stukel D. M. and Pfeffermann D. (1998) Bayesian versus frequentist measures of error in small area estimation *Journal of the Royal Statistical Society, B* **60** 377–396.
Sitter R. R. (1992) A resampling procedure for complex survey data *Journal of the American Statistical Association* **87** 755–765.
Sitter R. R. (1997) Variance estimation for the regression estimator in two-phase sampling *Journal of the American Statistical Association* **92** 780–787.
Skinner C. J., Holt D. and Smith T. M. F. (eds) (1989) *Analysis of Complex Surveys* Chichester: Wiley.
Snijders T. A. B. and Bosker R. J. (2002) *Multilevel Analysis: an Introduction to Basic and Advanced Multilevel Modelling* London: Sage Publications.
Sudman S. (1976) *Applied Sampling* New York: Academic Press.
Tepping B. J. (1968) Variance estimation in complex surveys *Proceedings of the Social Statistics Section* American Statistical Association 11–18.
Thomas D. R. and Rao J. N. K. (1987) Small-sample comparisons of level and power for simple goodness-of-fit statistics under cluster sampling *Journal of the American Statistical Association* **82** 630–636.
Thomas D. R., Singh A. C. and Roberts G. R. (1996) Tests of independence on two-way tables under cluster sampling: an evaluation *International Statistical Review* **64** 295–311.
Valliant R., Dorfman A. H. and Royall R. M. (2000) *Finite Population Sampling and Inference* New York: Wiley.
Verma V., Scott C. and O'Muircheartaigh C. (1980) Sample designs and sampling errors for the World Fertility Survey *Journal of the Royal Statistical Society A* **143** 431–473.
Wald A. (1943) Tests of statistical hypotheses concerning several parameters when the number of observations is large *Transactions of the American Mathematical Society* **54** 426–482.
Williams D. A. (1982) Extra-binomial variation in logistic linear models *Applied Statistics* **31** 144–148.
Wilson J. R. (1989) Chi-square tests for overdispersion with multiparameter estimates *Applied Statistics* **38** 441–453.
Wolter K. M. (1985) *Introduction to Variance Estimation* New York: Springer.
Woodruff R. S. (1971) A simple method for approximating the variance of a complicated estimate *Journal of the American Statistical Association* **66** 411–414.
You Y. and Rao J. N. K. (2000) Hierarchical Bayes estimation of small area means using multi-level models *Survey Methodology* **26** 173–181.
You Y. and Rao J. N. K. (2002) A pseudo-empirical best linear unbiased prediction approach to small-area estimation using survey weights *The Canadian Journal of Statistics* **30** 431–439.

Yung W. and Rao J. N. K. (2000) Jackknife variance estimation under imputation for estimators using poststratification information *Journal of the American Statistical Association* **95** 903–915.

Ziegler A., Kastner C. and Blettner M. (1998) The generalized estimating equations: an annotated bibliography *Biometrical Journal* **40** 115–139.

Author Index

Practical Methods for Design and Analysis of Complex Surveys Risto Lehtonen and Erkki Pahkinen

ISBN: 0-470-84769-7

Subject Index

Practical Methods for Design and Analysis of Complex Surveys Risto Lehtonen and Erkki Pahkinen

ISBN: 0-470-84769-7

Statistics in Practice

Human and Biological Sciences

Brown and Prescott—Applied Mixed Models in Medicine
Ellenberg, Fleming and DeMets—Data Monitoring Committes in Clinical Trials: A Practical Perspective
Lawson, Browne and Vidal Rodeiro—Disease Mapping with WinBUGS and MLwiN
Lui—Statistical Estimation of Epidemiological Risk
Marubini and Valsecchi—Analysing Survival Data from Clinical Trials and Observation Studies
Parmigiani—Modeling in Medical Decision Making: A Bayesian Approach
Senn—Cross-over Trials in Clinical Research, Second Edition
Senn—Statistical Issues in Drug Development
Spiegelhalter, Abrams and Myles—Bayesian Approaches to Clinical Trials and Health-Care Evaluation
Whitehead—Design and Analysis of Sequential Clinical Trials, Revised Second Edition
Whitehead—Meta-Analysis of Controlled Clinical Trials

Earth and Environmental Sciences

Buck, Cavanagh and Litton—Bayesian Approach to Interpreting Archaeological Data
Glasbey and Horgan—Image Analysis for the Biological Sciences
Webster and Oliver—Geostatistics for Environmental Scientists

Industry, Commerce and Finance

Aitken—Statistics and the Evaluation of Evidence for Forensic Scientists
Lehtonen and Pahkinen—Practical Methods for Design and Analysis of Complex Surveys, Second Edition
Ohser and Mücklich—Statistical Analysis of Microstructures in Materials Science