ICSA Book Series in Statistics

Series Editors: Jiahua Chen · (Din) Ding-Geng Chen

Changbao Wu
Mary E. Thompson

Sampling Theory and Practice

ICSA
泛華統計協會

Springer

ICSA Book Series in Statistics

The ICSA Book Series in Statistics showcases research from the International Chinese Statistical Association that has an international reach. It publishes books in statistical theory, applications, and statistical education. All books are associated with the ICSA or are authored by invited contributors. Books may be monographs, edited volumes, textbooks and proceedings.

More information about this series at http://www.springer.com/series/13402

Changbao Wu • Mary E. Thompson

Sampling Theory and Practice

Springer

Changbao Wu
Department of Statistics
and Actuarial Science
University of Waterloo
Waterloo, ON, Canada

Mary E. Thompson
Department of Statistics
and Actuarial Science
University of Waterloo
Waterloo, ON, Canada

ISSN 2199-0980 ISSN 2199-0999 (electronic)
ICSA Book Series in Statistics
ISBN 978-3-030-44244-6 ISBN 978-3-030-44246-0 (eBook)
https://doi.org/10.1007/978-3-030-44246-0

This Springer imprint is published by the registered company Springer Nature Switzerland AG.
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

To Miranda, Domeny, and Jianchuan
Changbao Wu

To Elise, Amy, Abigail, and Adam
Mary E. Thompson

Foreword

Probability sampling designs and repeated sampling inference, also called design-based approach, have played a dominant role, especially in the production of official statistics, ever since the publication of the landmark paper by Neyman (1934). Neyman's paper laid theoretical foundations of the design-based approach which was almost universally accepted by practicing survey statisticians because of its model-free features. It also inspired various important theoretical contributions, mostly motivated by practical and efficiency considerations, thanks to the early pioneering contributions of P. C. Mahalanobis, Morris Hansen, P. V. Sukhatme, and others. Classical textbooks on sampling theory by Cochran (1953), Hansen et al. (1953), and Sukhatme (1954) expanded and elaborated the theory of the design-based approach and greatly benefited students as well as practitioners. I was very fortunate to have been a graduate student of the University of Bombay soon after those textbooks appeared. V. P. Godambe, a colleague of the present authors at the University of Waterloo and a former student of the University of Bombay, addressed foundational issues in the theory associated with the design-based approach in his landmark 1955 paper and subsequent papers. He studied the role of labels attached to the population units in making the design-based inference. He established the nonexistence of best linear design-unbiased estimator of a population mean in the general class of estimators induced by the labels, even for simple random sampling. This negative result prompted the development of alternative approaches, including the scale-load approach of Hartley and Rao (1968) and the model-dependent or prediction approach of Brewer (1963b) and Royall (1970). The scale-load approach is the forerunner of the currently popular empirical likelihood (EL) method proposed by Owen (1988) under the mainstream setup. Wu and Thompson have devoted Chap. 8 to the EL method taking account of survey design features.

The prediction approach can lead to conditional inferences more relevant and appealing than repeated sampling inference, but for large samples, it may perform poorly under model misspecification in the sense that even a small model deviation can cause serious problems (Hansen et al. 1983). A compromise approach, called the model-assisted approach, uses a working model but design-based inferences remain valid for large samples even under model deviations. The model-assisted approach leads to the currently popular generalized regression estimators based on covariates related to the study variable.

The first five chapters (Part I) of the book by Wu and Thompson cover the basic material of the design-based approach and the model-dependent and model-assisted approaches in a succinct manner. Various extensions of the basic theory are delegated to exercises at the end of each chapter. Chapters 6–11 (Part II) cover important theoretical developments in modern survey sampling. The topics covered include calibration weighting and estimation, survey-weighted regression analysis and estimating equations methods, empirical likelihood (EL) inference and analysis of survey data, methods for handling unit and item nonresponse, resampling and replication methods for variance estimation, and Bayesian methods for survey data analysis. The EL method provides a nonparametric approach to constructing likelihood-ratio type confidence intervals that are data-driven and range-respecting, unlike the customary large-sample normal theory intervals. Wu and Thompson provided R codes to facilitate the implementation of EL on survey data.

Different types of surveys are presented in Part III and the associated methodologies are related to the basic theory in Part I. This promotes appreciation of the basic theory and its extensions. Surveys covered include area frame household surveys, telephone and web surveys, natural resources inventory surveys, adaptive and network surveys, and multiple frame surveys. The authors were heavily involved in the design and analysis of several surveys covered in Part III. Some of the surveys studied require advanced theory to account for complexities in the survey design.

I am very happy to see that the last chapter in the book (Chap. 17) addresses inferential issues involving non-probability samples. This topic is receiving a lot of attention these days because of steadily falling response rates and increasing costs and response burden associated with traditional surveys based on probability sampling, and the availability of low-cost non-probability samples of very large sizes. As a result, statistical agencies producing official statistics are now undertaking modernization initiatives to find new ways to integrate data from multiple sources. The methods proposed in Chap. 17, based on integrating non-probability samples with independent probability samples observing common covariates related to study variables, are very promising.

Wu and Thompson have done a great service in producing this comprehensive book on survey sampling theory and methods, covering both basic theory and new developments and applications to a variety of real surveys.

Carleton University
Ottawa, ON, Canada
January 2020

J.N.K. Rao

Preface

This book has been developed out of our many years of teaching an advanced course in survey sampling to fourth-year undergraduate students and graduate students in statistics and biostatistics at University of Waterloo and our many years of research and consulting activities over a wide range of topics in survey sampling. There exist several textbooks or research monographs in the field, such as the popular textbooks by Cochran (1977) and Lohr (2010), the advanced books on sampling statistics by Fuller (2009), and small area estimation by Rao and Molina (2015), and the applied survey sampling books by Valliant et al. (2013) and Heeringa et al. (2017). Why another book on survey sampling?

Finite population sampling and design-based inference only gained widespread acceptance in the 1940s and 1950s and were mainly used by official statistics in the early days. The field of survey sampling, however, has been developed at an amazing pace and has become a vital part of the modern statistical sciences. Surveys are conducted on human populations, animal and fish populations, businesses and establishments, and natural resources and the environment. Survey data are used for all sorts of scientific investigations and decision-making. While the fundamental framework for finite population sampling remains distinguished from other areas of statistics, many modern statistical advances have found adaptations for complex survey data analysis under suitable inferential frameworks. It becomes increasingly clear that senior and graduate students in statistics and biostatistics require a rigorous course in survey sampling, junior researchers need a running start to current methodological advances in the field, and survey practitioners want a reliable resource that addresses applicable theory and practical issues from different subject areas under a single connected roof.

This book attempts to achieve three major goals in one package: (i) a concise introduction to basic sampling theory and methods; (ii) a detailed coverage of selected topics that reflect the current state of the art in survey methodology research; and (iii) a useful resource and pragmatic guide to survey design and implementation as well as survey data production and analysis. The book is structured into three parts, with each part focusing on one goal but with all topics presented under a cohesive framework. The book starts with a journey through classical finite population sampling techniques and continues through a number of theoretical and practical problems in modern survey sampling; it ends with discussions on a topic increasingly important in recent years, dealing with non-probability survey samples.

The basic materials presented in Part I (Chaps. 1–5), plus a couple of selected topics from Parts II and III, can be used for a one-term introductory course on survey sampling for students with a solid background in statistics. An advanced one-term course on survey sampling theory and practice for graduate students in statistics and biostatistics can be taught out of the six chapters from Part II, supplemented by materials from Part I (Chaps. 4 and 5) and selected topics from Part III. A special topic course focusing on certain aspects of modern sampling theory and methods could also be organized using materials from Parts II and III. The book provides sufficient materials for a two-term course on survey sampling and additional readings on several specialized topics.

We would like to express our thanks to Professor J.N.K. Rao, who graciously agreed to write a foreword for the book, for his many thoughtful comments and suggestions. His continued support and collaborative research played a crucial role in the conception of this book project. We would also like to thank many other research collaborators, colleagues, and students who have been the driving forces behind many research projects that we have completed.

Special thanks are due to Dr. Geoffrey Fong and Dr. Jiang Yuan, the Co-Principal Investigators for the International Tobacco Control (ITC) China Survey project, for their support and their contributions to the materials presented in Chap. 12. Thanks are due also to Dr. Jiahua Chen for his collaborative work on the design and analysis of the FPI Survey used in Chap. 14 and his support and encouragement for the writing of the book.

And lastly, Changbao Wu would like to express his sincere gratitude to the late Professor Randy Sitter who was tragically lost at sea during a kayak trip in the summer of 2007. It was Professor Sitter who led Changbao Wu into the field of survey sampling and guided him through his PhD dissertation research at Simon Fraser University.

Waterloo, ON, Canada Changbao Wu
Waterloo, ON, Canada Mary E. Thompson
January 2020

Contents

Part I Basic Concepts and Methods in Survey Sampling

1 Basic Concepts in Survey Sampling 3
1.1 Survey Populations 3
1.1.1 Eligibility Criteria for Survey Populations 4
1.1.2 Three Versions of Survey Populations 5
1.2 Survey Samples 5
1.2.1 Non-probability Survey Samples 6
1.2.2 Justifications for Using Survey Samples 7
1.3 Population Structures and Sampling Frames 7
1.3.1 Stratification 8
1.3.2 Clustering 8
1.3.3 Sampling Frames 9
1.4 Descriptive Population Parameters 10
1.5 Probability Sampling and Design-Based Inference 11
1.5.1 Probability Sampling Designs 12
1.5.2 Design-Based Inference in Survey Sampling 13
1.5.3 Principal Steps in Survey Sampling 13
1.6 Problems 14

2 Simple Single-Stage Sampling Methods 17
2.1 Simple Random Sampling Without Replacement 17
2.2 Simple Random Sampling with Replacement 21
2.3 Simple Systematic Sampling 22
2.4 Central Limit Theorems and Confidence Intervals 24
2.5 Sample Size Calculation 25
2.5.1 Accuracy Specified by the Absolute Tolerable Error 26
2.5.2 Accuracy Specified by the Relative Tolerable Error 27
2.6 Problems 28

3 Stratified Sampling and Cluster Sampling 33
3.1 Stratified Simple Random Sampling 33
3.1.1 Population Parameters 34
3.1.2 Sample Data 35
3.1.3 Estimation of the Overall Population Mean μ_y 35
3.1.4 Confidence Intervals 37
3.1.5 Justifications for Using Stratified Sampling 37
3.2 Sample Size Allocation Under Stratified Sampling 38
3.2.1 Proportional Allocation 38
3.2.2 Neyman Allocation 39
3.2.3 Optimal Allocation with Pre-specified Cost or Variance 41
3.3 Single-Stage Cluster Sampling 42
3.3.1 Notation for Cluster Sampling 42
3.3.2 Single-Stage Cluster Sampling with Clusters Selected by SRSWOR 43
3.3.3 Estimation of the Population Total T_y 44
3.3.4 Estimation of the Population Mean μ_y 45
3.3.5 A Comparison Between SRSWOR and Cluster Sampling 45
3.4 Two-Stage Cluster Sampling 46
3.4.1 Two-Stage Cluster Sampling with SRSWOR at Both Stages 47
3.4.2 Estimation of the Population Total T_y 48
3.4.3 Estimation of the Population Mean μ_y 50
3.5 Stratified Two-Stage Cluster Sampling 50
3.6 Problems 52

4 General Theory and Methods of Unequal Probability Sampling 57
4.1 Sample Inclusion Probabilities 57
4.2 The Horvitz-Thompson Estimator 59
4.2.1 The Yates-Grundy-Sen Variance Formula for the HT Estimator 62
4.2.2 The Hájek Variance Estimator 63
4.2.3 Estimation of the Population Mean μ_y and the Hájek Estimator 64
4.2.4 Estimation of the Population Distribution Function and Quantiles 65
4.3 PPS Sampling and the HT Estimator: An Optimal Strategy 66
4.3.1 PPS Sampling Designs 66
4.3.2 An Optimal Strategy 68
4.3.3 Basu's Elephant Example 70

4.4 PPS Sampling Procedures 71
4.4.1 PPS Sampling with $n = 2$ 71
4.4.2 The Randomized Systematic PPS Sampling Method 73
4.4.3 The Rao-Sampford Method 74
4.4.4 The Rao-Hartley-Cochran Method 75
4.5 PPS Sampling with Replacement 78
4.6 Poisson Sampling 83
4.7 Problems 85

5 Model-Based Prediction and Model-Assisted Estimation 89
5.1 Model-Based Prediction Methods 90
5.1.1 Model-Unbiased Prediction Estimator for μ_y 91
5.1.2 Variance Estimation for Model-Based Prediction 93
5.2 Model-Assisted Estimation Methods 94
5.2.1 The Simple Ratio Estimator 95
5.2.2 The Simple Regression Estimator 98
5.3 The Generalized Ratio Estimator 100
5.4 The Generalized Regression Estimator 103
5.4.1 GREG Under SRSWOR 104
5.4.2 GREG Under a General Sampling Design 106
5.4.3 The Generalized Difference Estimator 107
5.5 Problems 108

Part II Advanced Topics on Analysis of Probability Survey Samples

6 Calibration Weighting and Estimation 115
6.1 Calibration Estimators 116
6.1.1 Definition of Calibration Estimators 116
6.1.2 Calibration Estimators Under the Chi-Square Distance 118
6.2 Calibration Weighting 121
6.2.1 Generalized Exponential Tilting 121
6.2.2 Generalized Pseudo Empirical Likelihood 123
6.2.3 Comparisons Among Three Calibration Methods 124
6.2.4 Range-Restricted Calibration 125
6.3 Model-Calibration Estimators 127
6.3.1 Model-Calibration Estimators of the Population Total 128
6.3.2 Model-Calibration Estimators of Other Parameters 130
6.4 Raking Ratio Estimation 132
6.5 Additional Remarks 134
6.6 Problems 136

7 Regression Analysis and Estimating Equations 139
7.1 Parameters of Superpopulations and Survey Populations 140
7.1.1 Model Parameters and Estimating Functions 140
7.1.2 Survey Population Parameters 142

7.2 Survey Weighted Estimating Equations 144
7.2.1 Point Estimators 144
7.2.2 Design-Based Variance Estimation 147
7.2.3 Confidence Intervals and Regions 149
7.3 Regression Analysis with Survey Data 150
7.3.1 Linear Regression Analysis 151
7.3.2 Logistic Regression Analysis 153
7.4 Longitudinal Surveys and Generalized Estimating Equations 155
7.5 Problems 158

8 Empirical Likelihood Methods 161
8.1 Pseudo Empirical Likelihood and Sample Empirical Likelihood 162
8.2 Pseudo Empirical Likelihood for Non-stratified Sampling 164
8.2.1 Pseudo EL Without Auxiliary Information 164
8.2.2 Pseudo EL with Auxiliary Information 167
8.2.3 Examples of Pseudo EL Ratio Confidence Intervals 169
8.3 Pseudo Empirical Likelihood for Stratified Sampling 171
8.4 Computational Procedures 173
8.4.1 Constrained Maximization for Non-stratified Samples 173
8.4.2 Constrained Maximization for Stratified Samples 174
8.4.3 Profile Pseudo EL Ratio Confidence Intervals 176
8.5 Generalized Pseudo Empirical Likelihood 177
8.6 Pseudo Empirical Likelihood and Estimating Equations 180
8.6.1 General Results on Point and Variance Estimation 180
8.6.2 General Results on Hypothesis Tests 183
8.7 Sample Empirical Likelihood and Estimating Equations 185
8.7.1 General Results on Point and Variance Estimation 185
8.7.2 General Results on Hypothesis Tests 186
8.8 Design-Based Variable Selection Methods 187
8.8.1 Penalized Pseudo Empirical Likelihood 188
8.8.2 Penalized Sample Empirical Likelihood 189
8.9 Problems 190

9 Methods for Handling Missing Data 193
9.1 Issues with Missing Survey Data 193
9.1.1 Missing Data Mechanisms 194
9.1.2 Frameworks for Statistical Inference 195
9.1.3 Validity and Efficiency 196
9.2 Methods for Handling Unit Nonresponse 196
9.3 Methods for Handling Item Nonresponse 198
9.3.1 Complete-Case Analysis 198
9.3.2 Propensity Score Adjusted Analysis 199
9.3.3 Analysis with Imputation for Missing Values 200
9.3.4 Doubly Robust Estimation 201

9.4 Imputation for Missing Values 202
9.4.1 Single Imputation 203
9.4.2 Multiple Imputation 206
9.4.3 Fractional Imputation 207
9.5 Estimation with Missing Survey Data 209
9.5.1 Estimation of the Population Mean 210
9.5.2 Regression Analysis: Scenario I 211
9.5.3 Regression Analysis: Scenario II 213
9.6 Variance Estimation 214
9.6.1 PSA Estimators 214
9.6.2 Single Imputation Based Estimators 216
9.6.3 Multiple Imputation Estimators 217
9.7 Some Additional Notes on Missing Data Problems 218
9.8 Problems 219

10 Resampling and Replication Methods 223
10.1 The With-Replacement Bootstrap 223
10.1.1 Single-Stage Sampling 224
10.1.2 Stratified Sampling 225
10.1.3 Multi-Stage Cluster Sampling 226
10.2 The Pseudo-Population Bootstrap 227
10.3 The Multiplier Bootstrap 229
10.3.1 The Multiplier Bootstrap with Estimating Functions 230
10.3.2 Variance Estimation in the Presence of Nuisance Functionals 231
10.4 Replication Weights for Public-Use Survey Data 233
10.5 An Algebraic Construction of Replication Weights 237
10.6 Bootstrap Methods for Imputed Survey Data 238
10.7 Additional Remarks 241
10.8 Problems 242

11 Bayesian Empirical Likelihood Methods 245
11.1 Bayesian Inference with Survey Data 245
11.2 Bayesian Inference Based on Pseudo Empirical Likelihood 247
11.3 Bayesian Inference Based on Profile Pseudo Empirical Likelihood 249
11.4 Bayesian Inference for General Parameters 251
11.4.1 Bayesian Inference with a Fixed Prior 251
11.4.2 Bayesian Inference with an n-Dependent Prior 252
11.4.3 Computational Procedures 253
11.5 Additional Remarks 254
11.6 Problems 255

Part III Practical Issues and Special Topics in Survey Sampling

12 Area Frame Household Surveys 259
12.1 Household Surveys 259
12.2 The ITC China Survey Design 261
12.2.1 The Target Population and Method for Data Collection 261
12.2.2 Sample Size Determination 262
12.2.3 Frame Construction and Sample Selection 262
12.2.4 Survey Measures and Questionnaire Development 264
12.3 The ITC China Survey Procedures 264
12.3.1 Survey Team 265
12.3.2 Training 265
12.3.3 Quality Control 266
12.3.4 Measures of Data Quality 266
12.4 Weight Calculation 267
12.4.1 Dealing with Cluster Level Unit Nonresponse 267
12.4.2 Cross-Sectional Weights 267
12.4.3 Longitudinal Weights 269
12.5 Additional Remarks 269

13 Telephone and Web Surveys 271
13.1 Telephone Survey Frames and Sampling Methods 271
13.2 Components of Survey Error for Telephone Surveys 273
13.3 Web Survey Frames and Sampling Methods 274
13.4 Mixed Mode Surveys 275
13.5 Construction of Survey Weights 275
13.5.1 Survey Weights for RDD Surveys 276
13.5.2 Survey Weights for Web Surveys 278
13.6 Problems 278

14 Natural Resource Inventory Surveys 281
14.1 Surveys of Non-human Populations 281
14.1.1 Non-human Population Surveys 281
14.1.2 Features of Natural Resource Inventory Surveys 283
14.2 Fish Abundance Surveys 283
14.2.1 The Fish Abundance Index 284
14.2.2 The FPI Survey Design 285
14.3 Estimation of Fish Abundance Indices 287
14.3.1 Models for the Catching Process 287
14.3.2 Point Estimation 289
14.3.3 Variance Estimation 290
14.3.4 Pseudo Empirical Likelihood Methods 291

15 Adaptive and Network Surveys 295
15.1 Finite Populations with Network Structure 296
15.1.1 Terminology and Notation 296
15.1.2 Estimation Problems 297
15.1.3 Probability Models for Networks 297
15.2 Link Tracing Sampling Designs 298
15.2.1 Snowball Sampling 299
15.2.2 Respondent Driven Sampling 300
15.2.3 Indirect Sampling 300
15.2.4 Adaptive Cluster Sampling 302
15.3 Problem 303

16 Dual Frame and Multiple Frame Surveys 305
16.1 Estimation with Dual Frame Surveys 306
16.2 Pseudo Empirical Likelihood for Dual Frame Surveys 308
16.2.1 Point Estimation 309
16.2.2 Confidence Intervals 310
16.2.3 Auxiliary Information and Computational Procedures ... 311
16.3 A Multiplicity-Based Approach for Multiple Frame Surveys 312
16.3.1 The Single-Frame Multiplicity-Based Estimator 312
16.3.2 Multiplicity-Based Pseudo Empirical Likelihood Approach 314
16.4 A Bootstrap Procedure for Multiple Frame Surveys 316
16.5 Additional Remarks 317

17 Non-probability Survey Samples 319
17.1 Non-probability Samples 319
17.2 General Setting and Basic Assumptions 320
17.3 Sample Matching and Mass Imputation 322
17.4 Estimation of Propensity Scores 323
17.5 Estimation of the Population Mean 325
17.6 Variance Estimation 326
17.7 An Example and Additional Remarks 327

A R Code for PPS Sampling, Empirical Likelihood and Regression Analysis 333
A.1 Randomized Systematic PPS Sampling 333
A.2 Rao-Sampford PPS Sampling 334
A.2.1 Select a PPS Sample 334
A.2.2 Compute Second Order Inclusion Probabilities 335
A.3 Empirical Likelihood Methods 336
A.3.1 Lagrange Multiplier: Univariate Case 337
A.3.2 Lagrange Multiplier: Vector Case 337
A.3.3 Lagrange Multiplier: Stratified Sampling 338
A.3.4 Empirical Likelihood Ratio Confidence Intervals 339

A.4 Survey Weighted Regression Analysis 340
A.4.1 Linear Regression Analysis 340
A.4.2 Logistic Regression Analysis 340

References 343

Author Index 357

Subject Index 361

Acronyms

ALB	Average lower bound
BRR	Balanced repeated replication
CAPI	Computer-aided personal interviewing
CATI	Computer-assisted telephone interviewing
CCA	Complete-case analysis
CLSA	Canadian Longitudinal Study on Aging
CP	Coverage probability
CPUE	Catch per unit effort
DF	Dual frame
EL	Empirical likelihood
ET	Exponential tilting
FPI	Fishery Products International
GD	Generalized difference
GEE	Generalized estimating equation
GLM	Generalized linear model
GR	Generalized ratio
GREG	Generalized regression
HH	Hansen–Hurwitz
HT	Horvitz–Thompson
IID	Independent and identically distributed
ITC	International Tobacco Control
IPW	Inverse probability weighting
LASSO	Least absolute shrinkage and selection operator
LE	Lower tail error rate
MAR	Missing at random
MC	Model calibration
MCAR	Missing completely at random
MCMC	Markov chain Monte Carlo
MNAR	Missing not at random
NNI	Nearest neighbor imputation
PEL	Pseudo empirical likelihood

PPS	Probability proportional to size
PSA	Propensity score adjusted
PSU	Primary sampling unit
RDD	Random digit dialing
RDS	Respondent-driven sampling
REG	Regression
RHC	Rao–Hartley–Cochran
SCAD	Smoothly clipped absolute deviation
SEL	Sample empirical likelihood
SRSWOR	Simple random sampling without replacement
SRSWR	Simple random sampling with replacement
SSS	Simple systematic sampling
SSU	Secondary sampling unit
UE	Upper tail error rate

Part I
Basic Concepts and Methods in Survey Sampling

Chapter 1
Basic Concepts in Survey Sampling

This chapter provides brief descriptions of basic concepts and some commonly used terminology in survey sampling. It also gives a glimpse of the notation system to be used for the rest of the book.

1.1 Survey Populations

Statistics is the science of how to collect and analyze data, and draw statements and conclusions about unknown populations. The term *population* usually refers to a real or hypothetical set of units with characteristics and attributes which can be modeled by random variables and their respective probability distributions.

A hypothetical population is *infinite* in nature and can never be known exactly. For instance, the population of all outcomes when a coin is tossed repeatedly may be considered as an infinite set of outcomes labeled “H” (head) or “T” (tail) under repeated trials of flipping the coin. The trial can be modeled by a Bernoulli random variable X taking value 1 when the outcome is head and 0 otherwise. However, the probability of obtaining a head from flipping a particular coin, $p = P(X = 1)$, may never be known exactly. No coins are perfectly symmetric and it would be very difficult to tell whether $p = 0.5$ or, for instance, $p = 0.5001$. Moreover, the claim that the population of all outcomes from flipping a coin can be modeled by a Bernoulli random variable or a Bernoulli distribution is simply an assumption. One could easily argue that there are possible outcomes other than head or tail from flipping a coin: the coin could rest on its side and be neither head nor tail, or the coin could roll onto the ground and fall into a crack.

Survey sampling deals with real world populations which can be viewed as a *finite* set of labeled units. A survey population can therefore be represented by the set of N labels:

C. Wu, M. E. Thompson, *Sampling Theory and Practice*, ICSA Book Series in Statistics, https://doi.org/10.1007/978-3-030-44246-0_1

$$\mathbf{U} = \{1, 2, \cdots, N\},$$

where $\mathbf{U}$ refers to the population, or equivalently, the *universe*, and N is the population size. Each label i ($i = 1, 2, \cdots, N$) represents a unique individual unit in the population.

Many surveys are conducted over human populations; these are the so-called *population surveys*. Depending on the objectives of the study, the survey population is often a subset of a general population. Here are some simple examples, each determined at a specified point in time:

- The population of all adult Canadians (age 18+).
- The population of all adult Canadians who are regular smokers.
- The population of all full-time college students in Ontario.
- The population of all children aged 6–12 (inclusive) who attend public schools in the Kitchener-Waterloo area.

Surveys can be carried out for fish and animal populations as well. Other important categories of surveys include

- *Business surveys*: the population consists of certain types of business establishment, such as all hi-tech companies in Ontario.
- *Agricultural surveys*: the population is related to farms, farmers, specific agricultural products or related operations.
- *Natural resource surveys*: the population is usually a collection of land sections or the complete inventory of a particular natural resource, such as all forests in British Columbia.

1.1.1 Eligibility Criteria for Survey Populations

Eligibility criteria for the inclusion or exclusion of individuals or units are used to define the survey population. However, it is much easier to ascertain someone who is an adult (18+) than to identify someone who is a regular adult smoker. For the International Tobacco Control Policy Evaluation Project (the ITC Project) China Survey (Wu et al. 2010), a regular adult smoker is defined as

> An adult who has smoked more than 100 cigarettes in his/her lifetime and currently smokes at least once a week.

The term *farm* seems to be clear in all general readings but the exact definition for surveys can be quite difficult. The National Agricultural Statistical Service (NASS) in the United States has changed the definition several times over the history of agricultural data collection. At one time it defined *farm* as

> A unit which grows and sells agricultural products with a revenue of at least $200.00 or a unit which owns at least 12 horses.

By this definition, the former Governor General of Canada and former President of University of Waterloo David Johnston might have been classified as a farmer during his time at Waterloo, since he was said to have owned 12 horses at his residence in Heidelberg, Ontario.

In survey practice, a short screening questionnaire is often used first to establish the eligibility of a unit for the survey before any formal survey interviews or any measures of the unit being taken.

1.1.2 Three Versions of Survey Populations

In an ideal world for surveys, one would like to have a uniquely defined and accessible population throughout the survey process. In reality, survey samplers have to deal with important practical issues such as *incomplete sampling frame* and *nonresponse*. The term *sampling frame* will be defined and elaborated further in the next section. Because of frame imperfections and nonresponse, most surveys could involve three conceptually different populations.

(a) *The target population*: The set of all units covered by the main objective of the study.
(b) *The frame population*: The set of all units covered by the sampling frame.
(c) *The sampled population*: The population represented by the survey sample. Under probability sampling, the sampled population is the set of all units which have a non-zero probability to be selected in the sample.

The sampled population is sometimes also called *the study population*. The target population and the frame population are not the same if the sampling frame is not complete, i.e., certain units in the target population are not listed on the sampling frame. An incomplete sampling frame implies *coverage error*, one of the major sources of survey error. The sampled population is not the same as the frame population in the presence of nonresponse. How to deal with nonresponse is a major topic for both theory and practice in survey sampling.

We will present an example in Sect. 1.3 of the three versions of survey populations.

1.2 Survey Samples

A survey sample, denoted by $\mathbf{S}$, is a subset of the survey population $\mathbf{U}$:

$$\mathbf{S} = \{i_1, i_2, \cdots, i_n\} \subseteq \mathbf{U},$$

where n is the sample size and i_1, i_2, $\cdots$, i_n are the distinct labels for the n units in the sample. With some misuse of notation the sample may simply be denoted as $\mathbf{S} = \{1, 2, \cdots, n\}$ if it does not cause any confusion. It is apparent with the latter notation that the unit "1" in the sample **S** is not necessarily the unit "1" in the population **U**.

There are two general approaches for selecting a survey sample from the survey population: *non-probability sampling* and *probability sampling*. Modern survey sampling theory and methods have been primarily developed for probability sampling. Non-probability sampling methods were widely used in the early days of conducting surveys but had a diminished role since the 1950s when probability sampling methods became dominant in survey practice. However, there has been increased use of non-probability survey samples in the twenty-first century as a time-efficient and cost-effective data source.

1.2.1 Non-probability Survey Samples

Some commonly used non-probability sampling methods include:

1. *Restricted sampling*: The sample is restricted to certain parts of the population which are readily accessible.
2. *Quota sampling*: The sample is obtained by a number of interviewers, each of whom is required to sample certain numbers of units with certain types or characteristics. How to select the units is completely left in the hands of the interviewers.
3. *Judgement or purposive sampling*: The sample is selected based on what the sampler believes to be "typical" or "most representative" of the population.
4. *Sample of convenience*: The sample is taken from those who are easy to reach.
5. *Sample of volunteers*: The sample consists of those who volunteer to participate.

The most popular non-probability sampling method in modern time is the so-called *Opt-in Panel Surveys*, where members of the panel signed up to take surveys, usually in order to earn cash or rewards. Non-probability sampling is used in practice due to various reasons. When the survey sampler opts not to use a probability sampling method, it is often because that such a plan is too time-consuming or too expensive or even not feasible at all. The most challenging task for non-probability survey samples is on data analysis. The validity of statistical inferences based on non-probability survey samples typically replies on model assumptions and the availability of additional information on the survey population. The last chapter of the book contains an overview of statistical inferences with non-probability survey samples.

This book mainly focuses on probability sampling methods, which will be formally introduced in Sect. 1.5. For finite populations, it is sometimes possible to carry out a complete enumeration of the entire population, i.e., to conduct a *census*, and various population quantities can be determined exactly. Why do we need survey sampling?

1.2.2 Justifications for Using Survey Samples

There are three main justifications for using a survey sample instead of a census.

- *Cost*: Survey samples can provide sufficient and reliable information about the survey population at far less cost. With a fixed and limited budget, it is usually impossible to conduct a census.
- *Time*: Survey samples can be collected relatively quickly, and results can be published or made available in a timely fashion. There is virtually no value, for instance, in determining the current unemployment rate exactly through a census if the result would not be available until the census is completed many months or even years later.
- *Accuracy*: Estimates based on a well-designed and well-executed survey sample are often more accurate (in terms of closeness to the true value) than results based on a loosely conducted census. This may be a little surprising. With large populations, a census requires a large administrative organization and involves many persons in the field for data collection. Inaccurate or biased measurements, recording mistakes, and other types of errors can easily be injected into a census. On the contrary, survey sample data can be collected by well trained personnel with high standards of data quality in place. In addition, with suitable probability sampling methods and large enough sample sizes, statements and conclusions can be made with any desired level of accuracy.

Survey sampling has been widely used by social and behavioural scientists as well as medical and health researchers as an important tool for collecting critical information on human populations. In addition to information on basic demographic variables such as age and gender, the research problems often involve sophisticated measures of social, behavioural and psychological indicators and measurements taken on blood, urine or other medical and biological samples. Such measurements can only be taken by trained interviewers or professionals. The amount of information to be collected can also be extremely large, with tens and sometimes hundreds of measurements for each respondent. Survey sampling is the only feasible approach for those types of studies.

1.3 Population Structures and Sampling Frames

Survey populations often possess geographic or administrative structures, which play an important role in survey design and data collection. Two primary types of population structures are *stratification* and *clustering*.

1.3.1 Stratification

The population $\mathbf{U}$ has a stratified structure if it is divided into H non-overlapping subpopulations:

$$\mathbf{U} = \mathbf{U}_1 \cup \mathbf{U}_2 \cup \cdots \cup \mathbf{U}_\mathrm{H} ,$$

where the subpopulation $\mathbf{U}_h$ is called stratum h, with stratum population size N_h, $h = 1, 2, \cdots, H$. It follows that $N = \sum_{h=1}^{\mathrm{H}} N_h$.

The general Canadian population may be stratified by the ten provinces and three territories. The whole country can also be divided into more refined and smaller areas, such as *federal electoral districts* or *health regions*. For the student population of an elementary school (grade 1 to grade 6), the population may be stratified by grades.

With a particular survey population, stratification may also be created in an artificial way, not necessarily confined to the geographic and administrative structures of the population. When the sampler has the freedom to specify the strata, stratification may be used for more efficient sampling designs or for a more balanced survey sample. For instance, human populations can be divided into subpopulations by gender and age groups. Such a stratification can be efficient for many studies but is not formed based on the physical or administrative structure of the population.

1.3.2 Clustering

Certain population units may be associated with each other or belong to a particular group. If the survey population can be divided into groups, called *clusters*, such that every unit in the population belongs to one and only one group, we say the population is clustered.

A city population may be viewed as clustered, where each residential block forms a cluster. For a rural population, a village or township could represent a cluster. The elementary and high school student population in Kitchener-Waterloo is also clustered, where each school is viewed as a cluster.

The two terms *stratum* (i.e., subpopulation) and *cluster* (i.e., group) both refer to a subset of the survey population, and the definitions seem to be arbitrary. The conceptual difference between *stratification* and *clustering*, however, is related to how the survey sample is selected.

- Under stratified sampling, sample data are collected from every stratum.
- Under cluster sampling, only a portion of the clusters has members in the final sample.

For example, the student population in Kitchener-Waterloo is stratified, with schools as strata, if sample data are collected from every school. It is a clustered population with schools as clusters if a random sample of schools is selected first and data are collected from those selected schools.

1.3.3 Sampling Frames

Individual units of the survey population are also called *observational units*; in principle, values and measures of population characteristics can be obtained directly for each unit. *Sampling units* refer to units used for selecting the survey sample. Depending on the sampling methods to be used, sampling units could be the individual units or clusters of the population, as described below. Lists of sampling units are called *sampling frames*.

Consider an un-stratified survey population $\mathbf{U} = \{1, 2, \cdots, N\}$. There are two commonly used sampling frames:

(a) A complete list of all N individual units. It could be the list of addresses (or telephone numbers or email addresses) of the N units for the survey population.
(b) A list of K clusters. The clusters are non-overlapping and together they cover the entire survey population.

Frame (a) can be used to select a sample by *simple random sampling without replacement*; see Sect. 2.1 for further details. In this case sampling units coincide with observational units, and both are the individual units in the population. Frame (b) is used for cluster sampling, and the clusters on the list are the primary sampling units (psu). To obtain the final survey sample, one would need secondary sampling frames within each selected cluster; see Chap. 3 for further details.

For stratified survey populations, sampling frames are constructed within each stratum. The stratum-specific sampling frame is either a complete list of all individual units or a list of clusters within the stratum.

Example 1.1 An education worker wanted to find out the average number of hours each week (of a certain month and year) spent on watching television by 4 and 5 year old children in the Region of Waterloo. She conducted a survey using the list of 123 pre-school kindergartens administered by the Waterloo Region District School Board. She first randomly selected ten kindergartens from the list. Within each selected kindergarten, she was able to obtain a complete list of all 4 and 5 year old children, with contact information for their parents/guardians. She then randomly selected 50 children from the list and mailed the survey questionnaire to their parents/guardians. The sample data were compiled from those who completed and returned the questionnaires.

- *The target population*: All 4 and 5 year old children in the Region of Waterloo at the time of the survey. This is defined by the overall objective of the study.
- *Sampling frames*: Two-stage cluster sampling methods were used (see Sect. 3.4 for further details). The first stage sampling frame is the list of 123 kindergartens

administered by the school board. The second stage sampling frames are the complete lists of all 4 and 5 year old children for the ten selected kindergartens.

- *Sampling units and observational units*: The first stage sampling units are the kindergartens; the second stage sampling units are the individual children (or equivalently, their parents); observational units are individual children.
- *The frame population*: All 4 and 5 year old children who attend one of the 123 kindergartens in the Region of Waterloo. It is apparent that children who are home-schooled are not covered by the frame population. Thus, as is frequently the case, the frame population is not the same as the target population.
- *The sampled population*: All 4 and 5 year old children who attend one of the 123 kindergartens in the Region of Waterloo and whose parents/guardians would complete and return the survey questionnaire if the child was selected for the survey.

The sampled population would be the same as the frame population if all parents would be willing to participate in the survey. In practice, it is very common to have nonresponse, and as a result the sampled population becomes a subset of the frame population. ◇

1.4 Descriptive Population Parameters

For most of the book, we assume that the target population and the frame population are identical. They are generally referred to as the survey population, unless indicated otherwise. In other words, we assume that the sampling frames are complete and cover all units in the target population.

Let $\mathbf{U} = \{1, 2, \cdots, N\}$ be the (un-stratified) survey population. Let y_i be the value of the *study variable* y and $\mathbf{x}_i$ be the value of the vector $\mathbf{x}$ of *auxiliary variables* attached to the individual unit i, $i = 1, 2, \cdots, N$. Those values may change over time. The definitions given below consider a fixed time point for the survey population. The population totals are defined as

$$T_y = \sum_{i=1}^{N} y_i \quad \text{and} \quad T_{\mathbf{x}} = \sum_{i=1}^{N} \mathbf{x}_i \, .$$

The population means are defined as

$$\mu_y = \frac{1}{N} \sum_{i=1}^{N} y_i \quad \text{and} \quad \mu_{\mathbf{x}} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{x}_i \, .$$

It follows that $\mu_y = T_y/N$ and $\mu_{\mathbf{x}} = T_{\mathbf{x}}/N$. The population variance of the study variable y is defined as

$$\sigma_y^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(y_i - \mu_y\right)^2 .$$

An important special case is when y is an indicator variable. Let

$$y_i = \begin{cases} 1 & \text{if unit } i \text{ has attribute "}A\text{"}, \\ 0 & \text{otherwise} . \end{cases}$$

The indicator variable y could represent any particular attribute, such as gender (male or female), educational level (with or without a college degree), income (annual income above certain level or not), opinion and attitude (yes or no; like or dislike; support or not support), etc. Let M be the total number of units in the population with attribute "A". We have

$$\mu_y = \frac{1}{N}\sum_{i=1}^{N} y_i = \frac{M}{N} = P ,$$

where $P = M/N$ is the population proportion of units with attribute "A". We also have

$$\sigma_y^2 = \frac{N}{N-1} P(1-P) .$$

When N is large, we have $\sigma_y^2 \doteq P(1-P)$.

The sampling theory presented in Chaps. 2–4 is primarily focused on the estimation of μ_y or T_y. It is shown in later chapters that the theory can be extended to statistical inference for more complex population parameters.

1.5 Probability Sampling and Design-Based Inference

Probability sampling is the foundation for modern survey sampling theory and practice. It was first laid out in the seminal paper of Neyman (1934). Under probability sampling, the selection of units for the survey sample is guided by a probability measure over all possible samples. The probability measure effectively plays the role of randomization as we see in other areas of modern statistics. It removes potential biases associated with subjective and other non-probability sampling methods. More importantly, probability sampling makes it possible to have rigorous statements about the unknown population with margins of error controlled through samples of suitable sizes.

1.5.1 Probability Sampling Designs

A *probability sampling design* is a probability measure over all possible candidate survey samples. Let

$$\Omega = \{\mathbf{S} \mid \mathbf{S} \subseteq \mathbf{U}\}$$

be the set of all possible subsets of the survey population $\mathbf{U}$. Let $\mathscr{P}$ be a probability measure over Ω such that

$$\mathscr{P}(\mathbf{S}) \geq 0 \quad \text{for any} \quad \mathbf{S} \in \Omega \qquad \text{and} \qquad \sum_{\mathbf{S}:\, \mathbf{S} \in \Omega} \mathscr{P}(\mathbf{S}) = 1\,.$$

A probability sample $\mathbf{S}$ can be selected based on the probability design, $\mathscr{P}$.

Example 1.2 Consider an over-simplified situation where the population size is $N = 3$ and $\mathbf{U} = \{1, 2, 3\}$. There are seven possible candidate samples:

$$\mathbf{S}_1 = \{1\},\ \mathbf{S}_2 = \{2\},\ \mathbf{S}_3 = \{3\},\ \mathbf{S}_4 = \{1, 2\},\ \mathbf{S}_5 = \{1, 3\},\ \mathbf{S}_6 = \{2, 3\},$$

$$\mathbf{S}_7 = \{1, 2, 3\}\,.$$

Note that $\mathbf{S}_7 = \mathbf{U}$, which corresponds to a census. Here are two sampling designs:

a. $\mathscr{P}(\mathbf{S}_k) = 1/6, k = 1, 2, \cdots, 6$ and $\mathscr{P}(\mathbf{S}_7) = 0$.
b. $\mathscr{P}(\mathbf{S}_k) = 1/3, k = 4, 5, 6$ and $\mathscr{P}(\mathbf{S}_k) = 0, k = 1, 2, 3, 7$.

Both sampling designs eliminate $\mathbf{S}_7 = \mathbf{U}$ as a possible sample. To select a sample under design b, for instance, we first generate a random number R from the uniform distribution over $[0, 1]$. If $0 \leq R \leq 1/3$, $\mathbf{S}_4$ is selected; if $1/3 < R \leq 2/3$, $\mathbf{S}_5$ is selected; and if $2/3 < R \leq 1$, $\mathbf{S}_6$ is selected. See Problem 1.3 for further detail on how to use random numbers generated from the uniform distribution to select a sample under general situations. ◇

The second sampling design in the above example gives non-zero probability only to candidate samples with size $n = 2$. Let $|\mathbf{S}|$ denote the cardinality of $\mathbf{S}$. A sampling design $\mathscr{P}$ has a fixed sample size n if $\mathscr{P}(\mathbf{S}) = 0$ for any $\mathbf{S}$ such that $|\mathbf{S}| \neq n$. In other words, the design $\mathscr{P}$ is a probability measure over the set

$$\Omega_n = \{\mathbf{S} \mid \mathbf{S} \subseteq \mathbf{U} \ \text{ and } \ |\mathbf{S}| = n\}\,.$$

For a population with size N, there is a total of $\binom{N}{n}$ candidate samples in Ω_n with sample size n. Even with moderate values of N and n, the total number would be excessively large to make the listing of all candidate samples extremely difficult or impossible. The probability measure $\mathscr{P}$ therefore has little practical use in terms of selecting the survey sample. Instead, probability survey samples are selected through specially designed procedures which draw units one at a time from the sampling frames. See Chaps. 2–4 for commonly used probability sampling methods.

1.5.2 Design-Based Inference in Survey Sampling

There are different approaches to inference when survey sample data are used. The so-called *design-based inference* assumes the following framework:

- The survey population $\mathbf{U} = \{1, 2, \cdots, N\}$ is viewed as fixed.
- The values y_i and $\mathbf{x}_i$ attached to unit i and the population parameters such as T_y and μ_y are also viewed as fixed.
- Randomization is induced by the probability sampling design for the selection of the survey sample.

Under the design-based framework for inference, probability statements and frequentist interpretations are under repeated sampling selections of survey samples. Probability sampling methods make it possible to use both classic and modern statistical tools for analyzing survey data, and to bring finite population surveys into the mainstream statistical sciences. Fundamental statistical concepts, such as unbiased estimators, variance estimation and confidence intervals become cornerstones of design-based inference in survey sampling.

1.5.3 Principal Steps in Survey Sampling

Survey sampling is used to collect information from populations as small as a group of 25 people in a sports bar or as large as the entire country of Canada. The objectives of a survey could be as simple as whether one likes a particular brand of chocolate or a complex study on aging with measures on hundreds of variables. However, there are a number of steps which are shared by all surveys.

1. A clear statement of the objectives of the survey.
2. Determination of the population to be sampled.
3. Determination of the relevant data to be collected.
4. Determination of the required accuracy of estimates.
5. Construction of sampling frames.
6. Choice of the sampling method.
7. Organization of the field work for data collection.
8. Plans for handling nonresponse.
9. Production of the survey dataset.
10. Summaries and analyses of the survey data.
11. Reports or publications on the study.

Steps 1 and 2 define the target population. Step 3 specifies the population quantities to be estimated and how the measures are to be taken. Some measures can be obtained through questionnaires but some others require physical procedures. Step 4 involves planning of sample sizes; Steps 5 and 6 are related to each other and need to be considered simultaneously; Step 7 involves the building of the survey team and the training of interviewers, data quality control and data entry.

Most surveys are conducted by subject area researchers who design the questionnaires or the measurement procedures. One important aspect in the design and analysis of surveys, which does not appear in above listed steps, is the overall budget for the study. Financial constraints often dictate the methods to be used and the size of the sample to be collected.

1.6 Problems

1.1 For each of the following surveys, describe briefly the *target population*, the *sampling frame* and the *frame population*, and the *sampled population* (or the *study population*). Discuss possible problems/issues with the sampling frames and the survey data in terms of coverage error and nonresponse bias.

(a) A survey conducted by the Dean of Mathematics at University of Waterloo (UW) indicates that about 25% of UW Computer Science graduates went to positions in the United States. Data were collected through questionnaires E-mailed to graduates from the past 5 years.

(b) A survey was conducted to find out the percentage of beer consumers in the Region of Waterloo who regularly drink the local brand Waterloo Dark. Data were collected through telephone interviews and phone numbers were selected from the published regional phone directory.

(c) A pilot survey for The Canadian Longitudinal Study on Aging (CLSA) was conducted in the province of Ontario. The survey intended to cover the general population of the province with age 45–80 (inclusive). Survey questionnaires were sent to selected individuals through regular mail. Individuals and their mailing addresses were selected and obtained from the Provincial Health Records.

1.2 Let $\pi_i = P(i \in \mathbf{S})$ be the probability that unit i is included in the sample. Consider Example 1.2. discussed in Sect. 1.5. For each of the two probability sampling designs, calculate the inclusion probabilities π_1, π_2 and π_3.

1.3 Suppose that the survey population has N elements: $\mathbf{U} = \{1, 2, \cdots, N\}$, and information on an auxiliary variable x is available for the entire population, i.e., values $x_1, x_2, \cdots, x_N$ are known. Assume that $x_i > 0$ for all i. Let n be the desired fixed sample size. The so-called *sampling proportional to the aggregated size design* is defined as follows: For any $\mathbf{S} \subseteq U$, $\mathscr{P}(\mathbf{S}) = 0$ if $|S| \neq n$ and $\mathscr{P}(\mathbf{S}) = C\sum_{i\in\mathbf{S}} x_i$ if $|\mathbf{S}| = n$. Here C is a normalization constant.

(a) For the special case of $N = 3$ and $n = 2$, find the constant C and compute the inclusion probabilities $\pi_i = P(i \in \mathbf{S})$ for $i = 1, 2, 3$.

(b) Find the constant C and compute the inclusion probabilities $\pi_i = P(i \in \mathbf{S})$ for the general case of arbitrary N and n, where $n < N$.

1.4 (Discrete Random Number Generator for Sample Selection)

(a) Let X be a discrete random variable with probability function $p_i = P(X = x_i)$, $i = 1, 2, \cdots, m$. Let $b_0 = 0$; let $b_j = \sum_{i=1}^{j} p_i$, $j = 1, 2, \cdots, m$, with $b_m = 1$. Generate R from $U[0, 1]$, the discrete uniform distribution over [0, 1]. Let $X^* = x_j$ if $b_{j-1} < R \le b_j$. Show that X^* has the same distribution as X.

(b) Discuss in detail how the random number generator described in (a) can be used to select a survey sample based on the given probability sampling design, $\mathscr{P}$.

(c) Discuss practical issues in using the method for selecting a probability sample from the survey population.

Chapter 2
Simple Single-Stage Sampling Methods

Single-stage probability sampling methods are used when a complete list of all population units is available and the list is used as the sampling frame for the probability selection of units for the survey sample. The *sampling units* are the same as the individual population units. The term *single-stage* distinguishes this kind of sampling from *two-stage* or *multi-stage* sampling methods to be discussed in Chap. 3, where the first stage sampling units are clusters.

There are two commonly used sampling frames for single-stage sampling for human populations: (i) a *List Frame*, which contains a complete list of addresses for all units; and (ii) a *Telephone Frame*, which contains a complete list of telephone numbers for all units. A third sampling frame, an *Email Frame*, which contains a complete list of email addresses for all units, has also become available for certain populations. For surveys of non-human populations, a list frame for all units in the population is required if one wishes to use a single-stage sampling method.

2.1 Simple Random Sampling Without Replacement

One of the simplest probability sampling designs is to select a sample of fixed size n with equal probability. The total number of candidate samples is $\binom{N}{n}$ for a population of size N, and the sampling design is specified by the probability measure given by

$$\mathscr{P}(\mathbf{S}) = \begin{cases} 1/\binom{N}{n} & \text{if } \ |\mathbf{S}| = n\,, \\ 0 & \text{otherwise}\,. \end{cases} \tag{2.1}$$

It would be a simple task to select a sample based on $\mathscr{P}$ if one could make a list of all $\binom{N}{n}$ candidate samples. Unfortunately, such a list is practically impossible to create when both N and n are moderately large.

C. Wu, M. E. Thompson, *Sampling Theory and Practice*, ICSA Book Series in Statistics, https://doi.org/10.1007/978-3-030-44246-0_2

In practice, survey samples are selected through a draw-by-draw method, the so-called *sampling scheme* or *sampling procedure*, which selects units from the sampling frame one at a time until the desired sample is taken. The types of available sample frames often dictate the types of sampling methods to be used; specific sampling methods require specific sampling frames.

The following sampling procedure with prespecified N and n is called *Simple Random Sampling Without Replacement* (SRSWOR). We assume that a complete list of all population units has already been created and can be used as the sampling frame.

1. Select the first unit from the N units on the sampling frame with equal probabilities $1/N$; denote the selected unit as i_1;
2. Select the second unit from the remaining $N-1$ units on the sampling frame with equal probabilities $1/(N-1)$; denote the selected unit as i_2;
3. Continue the process and select the nth unit from the remaining $N-n+1$ units on the sampling frame with equal probabilities $1/(N-n+1)$; denote the selected unit as i_n.

Let $\mathbf{S} = \{i_1, i_2, \cdots, i_n\}$ be the final set of n selected units. We have the following basic result on SRSWOR. The proof of the theorem is left as an exercise.

Theorem 2.1 *Under simple random sampling without replacement, the selected sample satisfies the probability measure given by (2.1), i.e.,* $\mathscr{P}(\mathbf{S}) = 1/\binom{N}{n}$ *if* $|\mathbf{S}| = n$, *and* $\mathscr{P}(\mathbf{S}) = 0$ *otherwise.*

Let $\{(i, y_i), i \in \mathbf{S}\}$ be the survey data, consisting of the labels i of the units sampled, and the associated values of y . Let the sample mean and sample variance be defined respectively as

$$\bar{y} = \frac{1}{n}\sum_{i\in\mathbf{S}} y_i \quad \text{and} \quad s_y^2 = \frac{1}{n-1}\sum_{i\in\mathbf{S}} \left(y_i - \bar{y}\right)^2 .$$

It is preferred to use $\sum_{i\in\mathbf{S}} y_i$ rather than $\sum_{i=1}^n y_i$ since the latter expression involves $\{y_1, y_2, \cdots, y_n\}$, which may be neither the ordered sequence from the draw-by-draw sampling selections nor the first n values of $\{y_1, y_2, \cdots, y_N\}$ for the survey population. For simplicity of notation, we will use $\{y_i, i \in \mathbf{S}\}$ or $\{(y_i, \mathbf{x}_i), i \in \mathbf{S}\}$ to denote the survey dataset and assume that the labels i are included as a separate ID column in the data file.

There are two major aspects of survey sampling: (i) sampling designs for selecting the survey sample; and (ii) statistical analysis of survey data. For (i), SRSWOR is the fundamental probability sampling method and is viewed as the baseline sampling design. More complex sampling methods are often assessed against SRSWOR in terms of practical constraints and efficiency comparisons . For (ii), the inference tools are first built for the estimation of the population mean μ_y or the population total T_y and then extended to cover more sophisticated problems. The two aspects, sampling design and estimation method, form what is termed as the *sampling strategy* (Thompson 1997, Sect. 2.4; Rao 2005, Sect. 3.1), and they should not be considered separately.

Selecting the sample $\mathbf{S}$ by SRSWOR and estimating the population mean μ_y by the sample mean $\bar{y}$ is one of the basic sampling strategies. We have the following fundamental results under the design-based framework.

Theorem 2.2 *Under simple random sampling without replacement:*

(a) The sample mean $\bar{y}$ is a design-unbiased estimator for the population mean μ_y, i.e.,

$$E(\bar{y}) = \mu_y \, .$$

(b) The design-based variance of $\bar{y}$ is given by

$$V(\bar{y}) = \Big(1 - \frac{n}{N}\Big)\frac{\sigma_y^2}{n} \, ,$$

where σ_y^2 is the population variance.

(c) An unbiased variance estimator for $\bar{y}$ is given by

$$v(\bar{y}) = \Big(1 - \frac{n}{N}\Big)\frac{s_y^2}{n} \, ,$$

which satisfies $E\big\{v(\bar{y})\big\} = V(\bar{y})$.

The expectation $E(\cdot)$ and variance $V(\cdot)$ are with respect to the probability sampling design $\mathscr{P}$ specified by (2.1).

The factor $1 - n/N$ appearing in $V(\bar{y})$ and $v(\bar{y})$ is called the *finite population correction* (fpc). Other than this factor, the results of the theorem look similar to what we see in conventional statistics with iid data. However, theoretical arguments under the design-based framework in survey sampling are quite different. For instance, the conventional equation $E\big(\sum_{i\in\mathbf{S}} y_i\big) = \sum_{i\in\mathbf{S}} E(y_i)$ does not make any sense under the probability sampling design, because the set $\mathbf{S}$ itself is random while y_i is a fixed number for any given i.

Proof We now present three different methods to prove part (a) of the theorem. Proofs for parts (b) and (c) are outlined in Problem 2.1 at the end of the chapter.

Method 1 Consider the sampling design specified by the probability measure $\mathscr{P}$. Noting that the population values $\{y_1, y_2, \cdots, y_N\}$ are fixed, we can view $\bar{y} = n^{-1}\sum_{i\in\mathbf{S}} y_i$ as a function of $\mathbf{S}$, i.e., $\bar{y} = \bar{y}(\mathbf{S})$. It follows that, under the probability measure $\mathscr{P}$ over Ω, we have

$$\begin{aligned}
E\big\{\bar{y}(\mathbf{S})\big\} &= \sum_{\mathbf{S}:\, \mathbf{S}\in\Omega} \bar{y}(\mathbf{S})\mathscr{P}(\mathbf{S}) \\
&= \binom{N}{n}^{-1} \sum_{\mathbf{S}:\, |\mathbf{S}|=n} \frac{1}{n}\sum_{i\in\mathbf{S}} y_i \\
&= \binom{N}{n}^{-1} \frac{1}{n}\sum_{i=1}^{N} \sum_{\mathbf{S}:\, i\in\mathbf{S}} y_i \\
&= \binom{N}{n}^{-1} \frac{1}{n}\sum_{i=1}^{N} \binom{N-1}{n-1} y_i \\
&= \frac{1}{N}\sum_{i=1}^{N} y_i = \mu_y \,.
\end{aligned}$$

Method 2 Consider the ordered sequence $(i_1, i_2, \cdots, i_n)$ from the draw-by-draw selections of the n units in the sample. Let $Z_k = y_{i_k}$ be the value of y from the kth draw. It follows that Z_k is a random variable with probability function given by

$$P(Z_k = y_i) = \frac{1}{N}\,, \quad i = 1, 2, \cdots, N \tag{2.2}$$

for any k $(k = 1, 2, \cdots, n)$. This leads to $E(Z_k) = \mu_y, k = 1, 2, \cdots, n$. We further have

$$E(\bar{y}) = E\left(\frac{1}{n}\sum_{i\in\mathbf{S}} y_i\right) = E\left(\frac{1}{n}\sum_{k=1}^{n} Z_k\right) = \frac{1}{n}\sum_{k=1}^{n} E(Z_k) = \mu_y \,.$$

Method 3 Use sample indicator variables. This is a general technique to be used extensively in Chap. 4 for handling unequal probability sampling designs. Let

$$A_i = \begin{cases} 1 & \text{if } i \in \mathbf{S}\,, \\ 0 & \text{if } i \notin \mathbf{S}\,, \end{cases} \qquad i = 1, 2, \cdots, N\,.$$

Note that A_i is a Bernoulli random variable under the sampling design and is defined for every i in the population, with $E(A_i) = P(i \in \mathbf{S}) = n/N$ and $V(A_i) = (n/N)(1 - n/N)$. We have

$$E(\bar{y}) = E\left(\frac{1}{n}\sum_{i\in \mathbf{S}} y_i\right) = E\left(\frac{1}{n}\sum_{i=1}^{N} A_i y_i\right) = \frac{1}{n}\sum_{i=1}^{N} y_i E(A_i) = \mu_y \,.$$

□

In many surveys the response y is a binary variable indicating a particular population attribute. We have $\mu_y = P = M/N$ as the population proportion of units with the attribute. Similarly, we have $\bar{y} = p = m/n$, where m is the number of units in the sample $\mathbf{S}$ with the attribute and p is the sample proportion of units with the attribute. It can be shown that $s_y^2 = (n/(n-1))p(1-p) \doteq p(1-p)$ if n is large. Under SRSWOR, we have

$$E(p) = P\,, \quad V(p) \doteq \left(1 - \frac{n}{N}\right)\frac{P(1-P)}{n} \quad \text{and} \quad v(p) \doteq \left(1 - \frac{n}{N}\right)\frac{p(1-p)}{n}\,,$$

where the approximation amounts to using $N/(N-1) \doteq 1$ and $n/(n-1) \doteq 1$.

2.2 Simple Random Sampling with Replacement

With-replacement sampling methods are of interest for two reasons. In certain practical situations it is sometimes impossible to remove a unit from the sampling frame, which makes with-replacement sampling selection of units unavoidable. More importantly, it is often of great interest to compare without-replacement sampling methods to with-replacement sampling methods, since the latter are easier to handle in terms of theoretical development.

The following sampling procedure with prespecified N and n is called *Simple Random Sampling With Replacement* (SRSWR). The required sampling frame is a complete list of all N units in the population.

1. Select the first unit from the N units on the sampling frame with equal probabilities $1/N$; denote the selected unit as i_1;
2. Select the second unit from the N units on the sampling frame with equal probabilities $1/N$; denote the selected unit as i_2;
3. Continue the process and select the nth unit from the N units on the sampling frame with equal probabilities $1/N$; denote the selected unit as i_n.

It is apparent that under SRSWR some units could be selected more than once. There are two possible ways to handle the sample from SRSWR: (i) remove duplicated units from the sample; and (ii) keep all selected units, including duplicated ones.

Let $\mathbf{S}$ be the set of distinct units selected by SRSWR; let $m = |\mathbf{S}|$ be the number of distinct units, i.e., the sample size of $\mathbf{S}$; let $\{y_i,\ i \in \mathbf{S}\}$ be the survey data. Let

$$\bar{y}_m = \frac{1}{m}\sum_{i\in\mathbf{S}} y_i$$

be the sample mean. It is shown in Problem 2.2 that $\bar{y}_m$ is an unbiased estimator of μ_y but it is less efficient than $\bar{y}$ from SRSWOR. Note that m is a random number satisfying $1 \le m \le n$.

Suppose that we keep all n selected units $(i_1, i_2, \cdots, i_n)$, including duplicated ones. Let $Z_k = y_{i_k}$ be the y value from the kth selection, $k = 1, 2, \cdots, n$. Under SRSWR the Z_k's are independent of each other and follow the same probability distribution given by (2.2). Let

$$\bar{Z} = \frac{1}{n}\sum_{k=1}^{n} Z_k$$

be the sample mean. It can be shown (Problem 2.3) that

$$E\big(\bar{Z}\big) = \mu_y \quad \text{and} \quad V\big(\bar{Z}\big) = \Big(1 - \frac{1}{N}\Big)\frac{\sigma_y^2}{n}\,. \tag{2.3}$$

For nontrivial cases where $n \ge 2$ and $\sigma_y^2 > 0$, we have $V(\bar{Z}) > V(\bar{y})$, where $V(\bar{y})$ is the variance of the sample mean from SRSWOR. In other words, SRSWR is less efficient than SRSWOR. However, the difference between the two sampling procedures becomes negligible when N is large and the sampling fraction n/N is small. Under such scenarios the probability that a unit is selected more than once under SRSWR becomes very small. If $m = n$, i.e., all selected units are distinct, the resulting sample is equivalent to a sample selected by SRSWOR. More formally, we have

$$\mathscr{P}\big(\mathbf{S} \mid m = n\big) = \binom{N}{n}^{-1},$$

since all candidate samples with $m = n$ are equally likely. We also have $V(\bar{Z})/V(\bar{y}) \to 1$ as $N \to \infty$ and $n/N \to 0$. See Sect. 2.4 for further detail on frameworks which allow $N \to \infty$ or $n \to \infty$.

2.3 Simple Systematic Sampling

The selection of units for the survey sample is typically done on paper. The process of collecting survey data often requires traveling and physically locating the selected units. Sampling procedures such as SRSWOR do not take into account the physical settings of the population units at the design stage of the survey, and can create unnecessary burdens for field workers in identifying selected units for the sample.

Systematic sampling methods have been used in agricultural surveys where selecting equally spaced units in the farm fields brings tremendous convenience for the survey team to make measurements on sampled units. The methods can also be used for household surveys where the population units may be arranged in a clearly defined sequence using maps for residential areas.

Suppose that the N population units have been arranged in a sequence, labelled as $1, 2, \cdots, N$. Let n be the desired sample size. Assume that $N = nK$ where K is an integer. The *simple systematic sampling* (SSS) method selects the n sampled units as follows:

1. Select the first unit, denoted as k, from the first K units (i.e., $1, 2, \cdots, K$) with equal probability $1/K$;
2. The remaining $n - 1$ units for the sample are automatically determined as $k + K$, $k + 2K, \cdots, k + (n - 1)K$.

Under SSS, there are only K candidate samples, $\mathbf{S}_k = \{k, k + K, \cdots, k + (n - 1)K\}$, $k = 1, 2, \cdots, K$, and each sample is completely determined by the initial unit, k. The sampling design is given by $\mathscr{P}(\mathbf{S}_k) = 1/K$, $k = 1, 2, \cdots, K$ and $\mathscr{P}(\mathbf{S}) = 0$ otherwise. Let $\bar{y}$ be the sample mean for the final selected sample.

Theorem 2.3 *Under simple systematic sampling, we have*

$$E(\bar{y}) = \mu_y \quad \textit{and} \quad V(\bar{y}) = \frac{1}{K}\sum_{k=1}^{K}\left(\bar{y}_k - \mu_y\right)^2,$$

where $\bar{y}_k = n^{-1}\sum_{i\in\mathbf{S}_k} y_i$ *is the sample mean for the kth candidate sample* $\mathbf{S}_k$.

Proof Noting that there are only K possible candidate samples which are determined by the K initially selected units, the sample mean $\bar{y}$ can be viewed as a random number selected from K possible values $\bar{y}_1$, $\bar{y}_2$, $\cdots$, $\bar{y}_K$ with equal probability $1/K$. Consequently,

$$E(\bar{y}) = \frac{1}{K}\sum_{k=1}^{K}\bar{y}_k = \frac{1}{nK}\sum_{k=1}^{K}\sum_{j=1}^{n} y_{k+(j-1)K} = \frac{1}{N}\sum_{i=1}^{N} y_i = \mu_y$$

and

$$V(\bar{y}) = E\left\{\left(\bar{y} - \mu_y\right)^2\right\} = \frac{1}{K}\sum_{k=1}^{K}\left(\bar{y}_k - \mu_y\right)^2.$$

□

One of the major difficulties with SSS concerns variance estimation. In this design the sample mean $\bar{y}$ is technically a single value randomly selected from a "population" with K possible values. It is common knowledge in statistics that a single data point does not capture the variation in the population and an unbiased variance estimator cannot be constructed. If there is sufficient evidence to suggest that the population sequence does not show any particular pattern, then certain approximate variance estimators may be used (Problem 2.4).

Another major difficulty with SSS arises when the population sequence possesses certain periodic patterns, leading to potentially serious bias for estimation. Since the n selected units are equally spaced, there is a chance that values of the response variable y for the sampled units may all take extreme values if the spacing for the sampled units matches the period of the cycles.

When the population size N is not a perfect multiple of the sample size n, a modified SSS method is to view the N population units as arranged round a circle and let k be the integer nearest to N/n (Cochran 1977, page 206; Murthy 1967, page 139). Select an initial number between 1 and N with equal probability and take every kth unit thereafter, going round the circle until n units have been selected for the final sample. Bellhouse (1988) provides an excellent overview with relevant references on systematic sampling.

2.4 Central Limit Theorems and Confidence Intervals

Construction of confidence intervals for the population mean μ_y replies on the sampling distribution of the point estimator $\hat{\mu}_y$. Under non-survey contexts, confidence intervals may be constructed based on the asymptotic distribution of the Wald statistic $\big(\hat{\mu}_y - \mu_y\big)/\{v(\hat{\mu}_y)\}^{1/2}$. The theoretical arguments involve limiting processes with the sample size $n \to \infty$. Under the design-based inferences for survey sampling, the survey population is finite and is viewed as fixed. It is logically impossible to let $n \to \infty$ while holding N fixed. We need a suitable framework to allow asymptotic development for finite populations.

A commonly used setting for asymptotic theory in survey sampling is as follows. Assume there is a sequence of finite populations, $\mathbf{U}_\nu$, indexed by ν, with population size N_ν, $\nu = 1, 2, \cdots$. The actual survey population is one of them. The population sizes $N_\nu \to \infty$ as $\nu \to \infty$. For a particular population $\mathbf{U}_\nu$, a sample of size n_ν is taken from the population with the chosen sampling design. The sample sizes also satisfy $n_\nu \to \infty$ as $\nu \to \infty$. For most asymptotic developments, the index ν is suppressed for notational simplicity, but all limiting processes are understood as governed by $\nu \to \infty$. See Fuller (2009, Sect. 1.3) for further discussion.

Madow (1948) was the first to discuss limiting distributions under finite population sampling. The following Central Limit Theorem is the first major asymptotic result in survey sampling and was established by Hájek (1960). The same topic was also discussed by Erdös and Rényi (1959) who proved asymptotic normality of the standardized mean.

Theorem 2.4 *Suppose that the sampling fraction* $n/N \to f \in (0,1)$ *as* $n \to \infty$. *Suppose also that the population values of the response variable* y *satisfy*

$$\lim_{N\to\infty} \frac{\max_{1\le i\le N}(y_i-\mu_y)^2}{\sum_{i=1}^{N}(y_i-\mu_y)^2} = 0\,. \tag{2.4}$$

Then under SRSWOR, the Wald-type statistic

$$\frac{\bar{y}-\mu_y}{\sqrt{v(\bar{y})}} \quad \text{converges in distribution to} \quad N(0,1)$$

as $n \to \infty$, *where* $v(\bar{y}) = (1-n/N)s_y^2/n$ *is the estimated variance of* $\bar{y}$.

Equation (2.4) is also called Noether's condition. It states that no single unit dominates the total variation of the population. A consequence is that an approximate $(1-\alpha)$-level confidence interval for the population mean μ_y can be constructed as

$$\left(\bar{y} - Z_{\alpha/2}\sqrt{v(\bar{y})}\,, \quad \bar{y} + Z_{\alpha/2}\sqrt{v(\bar{y})}\right), \tag{2.5}$$

where $Z_{\alpha/2}$ is the upper $\alpha/2$ quantile from the standard normal distribution. The interval (2.5) performs well for most real world applications when $n \ge 50$.

Central limit theorems in survey sampling under more complex sampling designs are difficult to establish. There are no universal results which apply to all sampling designs. Survey researchers have addressed some specific cases and established central limit theorems under certain regularity conditions. Examples include Erdös and Rényi (1959) and Hájek (1960) on simple random sampling with or without replacement (Theorem 2.4), Hájek (1964) on rejective sampling with unequal probabilities, Krewski and Rao (1981) on stratified unequal probability sampling with replacement, Bickel and Freedman (1984) on stratified simple random sampling without replacement, and Chen and Rao (2007) on two-phase sampling. Sen (1988) and Fuller (2009, Sect. 1.3) contain reviews on related literature. It is a research topic which requires further exploration by theoretical survey statisticians, especially for survey designs involving combinations of stratification, clustering and multistage sampling with unequal probabilities.

2.5 Sample Size Calculation

It is always one of the first questions to be asked at the survey planning stage: How large should the sample size n be? The answer depends on three factors:

(i) The total budget for the survey; (ii) The cost for surveying one unit and taking all required measurements; and (iii) The accuracy required for the main statistical inferences from the survey data. In this section we address the sample size calculation problem based only on the required accuracy, assuming that the sample is going to be selected by SRSWOR. The required sample size will need to be adjusted if more complex sampling designs are to be used. We focus on the accuracy for estimating the population mean μ_y.

2.5.1 Accuracy Specified by the Absolute Tolerable Error

It may not be possible to control the estimation error with certainty. Instead, one could specify the margin of error, e, and require that

$$P\big(|\bar{y} - \mu_y| \geq e\big) \leq \alpha \tag{2.6}$$

for a given $\alpha \in (0, 1)$. Treating $(\bar{y} - \mu_y)/\{V(\bar{y})\}^{1/2}$ as approximately distributed as $N(0, 1)$, where $V(\bar{y}) = (1 - n/N)\sigma_y^2/n$, we obtain the minimum required sample size n as

$$n \doteq \frac{Z_{\alpha/2}^2 \sigma_y^2}{e^2 + Z_{\alpha/2}^2 \sigma_y^2/N} = \frac{n_0}{1 + n_0/N}\,, \tag{2.7}$$

where $n_0 = Z_{\alpha/2}^2 \sigma_y^2/e^2$, which is the sample size required under SRSWR.

The sample size formula given by (2.7) requires information on N and σ_y^2. It should be noted that sample size calculation occurs at the survey planning stage, and the actual survey data are not available at this time. There are two commonly used strategies for obtaining the required information for the purpose of sample size calculation.

- **Historical data**: It is often the case that census data or surveys conducted in the past over the same population can be accessed to obtain rough estimates for the unknown population quantities.
- **Pilot survey**: If no information is available, it is possible to spend a small fraction of the total budget to conduct a pilot survey with a small sample size. Data from the pilot survey can then be used to estimate the population parameters.

A larger sample would seem to be more desirable from a statistical point of view. However, the real motivation behind sample size calculation at the planning stage is to avoid the mistake that at the end of the survey project the chosen sample size tends out to be too small to have the desired level of accuracy for inferences. Estimates for the unknown population quantities obtained from historical data or a small pilot survey may not be very accurate but they are typical good enough for planning purposes.

Example 2.1 Suppose that the goal is to estimate the population proportion $P = M/N$ using a survey sample to be selected by SRSWOR. Using the sample proportion $p = m/n$ to estimate P, a commonly used absolute tolerable error is $e = 3\%$ and the α is set at 0.05. In other words, the estimation accuracy is specified as

$$P(|p - P| \le 0.03) \ge 0.95\,. \tag{2.8}$$

Noting that $0.95 = 19/20$, the probability statement given in (2.8) is often quoted in media reports as "*The result is accurate within three percentage points, 19 times out of 20*".

For the population proportion P, we have $\sigma_y^2 \doteq P(1 - P) \le 1/4$. It follows that, for any population size N and with the choices $e = 0.03$ and $\alpha = 0.05$, we have

$$n \le n_0 = Z_{\alpha/2}^2 \sigma_y^2 / e^2 \le 1067\,.$$

The value $n = 1067$ is the most conservative sample size required to achieve (2.8) in estimation of an unknown population proportion regardless of the actual value of P and the population size N. ◇

2.5.2 Accuracy Specified by the Relative Tolerable Error

The specification of an absolute tolerable error e in (2.6) for the estimation accuracy can be affected dramatically by the use of different measurement scales such as meter or centimeter for the survey data. Assuming that $\mu_y \ne 0$, the issue of scales can be avoided by using a relative tolerable error and replacing (2.6) by

$$P\left(\frac{|\bar{y} - \mu_y|}{|\mu_y|} \ge e\right) \le \alpha\,. \tag{2.9}$$

Regardless of the scale used in measurements, the choice $e = 1\%$ would have the clear meaning of controlling the relative error at the one percent level. Noting that (2.9) can be rewritten as $P(|\bar{y} - \mu_y| \ge e^*) \le \alpha$, where $e^* = e|\mu_y|$, we obtain the minimum required sample size n specified by

$$n = \frac{Z_{\alpha/2}^2 \sigma_y^2}{(e^*)^2 + Z_{\alpha/2}^2 \sigma_y^2 / N} \doteq \frac{n_0}{1 + n_0/N} \tag{2.10}$$

where $n_0 = Z_{\alpha/2}^2 (CV_y)^2 / e^2$ and $CV_y = \sigma_y / \mu_y$ is the *coefficient of variation* for the y variable.

We once again need information on N and σ_y^2 as well as μ_y in order to calculate the required sample size n. In addition to using available historical data or data from a pilot survey, there is a third option for the current case. Suppose that $y_i \propto x_i$, where x_i is the value of the auxiliary variable x. It can be shown that $CV_y = CV_x$. In other words, if information on an auxiliary variable x is available and if y is roughly proportional to x, we can use CV_x for the purpose of sample size calculation.

The criterion specified by (2.9) is not suitable when μ_y is close to 0. One of such situations is the estimation of the population proportion $P = M/N$ when the value of P is very small. It can be shown that the value of n given by (2.10) will approach to N as P goes to zero. On the other hand, the absolute tolerable error specified by (2.6) is also difficult to use in this case. For instance, if $P = 0.001$, it would be difficult to decide what value of e is appropriate for the accuracy in estimating P. Alternative sampling methods have been proposed; see Problem 2.6 for further details.

2.6 Problems

2.1 (Proofs for Parts (b) and (c) of Theorem 2.2)

1. Use Method 2 for the proof of part (a) to prove part (b). **Hint**: Find $V(Z_k)$ and $Cov(Z_k, Z_j)$ for $k < j$ first.
2. Use Method 3 for the proof of part (a) to prove part (b). **Hint**: Find $Cov(A_i, A_j)$ for $i < j$ first.
3. Show that $E\big(n^{-1}\sum_{\in \mathbf{S}} y_i^2\big) = N^{-1}\sum_{i=1}^{N} y_i^2$.
4. Prove part (c) of the theorem by showing that $E(s_y^2) = \sigma_y^2$, i.e., the sample variance s_y^2 is an unbiased estimator for the population variance σ_y^2 under SRSWOR. **Hint**: Rewrite the sample variance as
$$s_y^2 = \frac{n}{n-1}\left[\frac{1}{n}\sum_{i\in\mathbf{S}} y_i^2 - (\bar{y})^2\right]$$
and use the result from (3) and $E\{(\bar{y})^2\} = \{E(\bar{y})\}^2 + V(\bar{y})$.
5. Show that the sample variance can be rewritten as
$$s_y^2 = \frac{n}{n-1}\left[\frac{1}{n}\sum_{i\in\mathbf{S}}(y_i - \mu_y)^2 - \left(\bar{y} - \mu_y\right)^2\right],$$
and use the result to prove that $E(s_y^2) = \sigma_y^2$. **Hint**: $E\big\{(\bar{y} - \mu_y)^2\big\} = V(\bar{y})$.
6. Show that the sample variance can alternatively be written in the so-called Laplace form
$$s_y^2 = \frac{1}{2n(n-1)}\sum_{i\in\mathbf{S}}\sum_{j\in\mathbf{S}}\left(y_i - y_j\right)^2.$$

Use this expression and the indicator variables A_i to prove that $E(s_y^2) = \sigma_y^2$. **Hint**: Consider the Laplace form for σ_y^2.

2.2 (Simple Random Sampling With Replacement (Part (i)))

(a) Argue that, conditional on a given value of m, the set of distinct units, **S**, selected by SRSWR can be viewed as a sample selected by SRSWOR with sample size m.

(b) Use the formula $E(X) = E\{E(X \mid Y)\}$ to show that $\bar{y}_m$ is a design-unbiased estimator of μ_y.

(c) Use the formula $V(X) = V\{E(X \mid Y)\} + E\{V(X \mid Y)\}$ to show that

$$V(\bar{y}_m) = \left[E\left(\frac{1}{m}\right) - \frac{1}{N}\right]\sigma_y^2 .$$

(d) Assume that $n \geq 2$ and $\sigma_y^2 > 0$. Argue that $V(\bar{y}_m) > V(\bar{y})$, where $\bar{y}$ is the sample mean and the sample of size n is selected by SRSWOR.

(e) Show that an unbiased variance estimator for $\bar{y}_m$ is given by

$$v\big(\bar{y}_m\big) = \left(\frac{1}{m} - \frac{1}{N}\right)\frac{1}{m-1}\sum_{i \in \mathbf{S}^*}\big(y_i - \bar{y}_m\big)^2 .$$

2.3 (Simple Random Sampling With Replacement (Part (ii)))

(a) Show that $\bar{Z}$ is an unbiased estimator of μ_y with $V(\bar{Z})$ given in (2.3). **Hint**: Find $V(Z_k)$ first.

(b) Let $\bar{y}_m$ and $\bar{Z}$ be the two versions of the sample mean under SRSWR and assume that $n \geq 2$ and $\sigma_y^2 > 0$. Show that

$$V(\bar{y}_m) > V(\bar{y}) \quad \text{and} \quad V(\bar{Z}) > V(\bar{y}) ,$$

where $\bar{y}$ is the sample mean, with the sample of size n selected by SRSWOR.

(c) Let γ_i be the number of times that unit i is selected under SRSWR, $i = 1, 2, \cdots, N$. We can rewrite $\bar{Z}$ as

$$\bar{Z} = \frac{1}{n}\sum_{k=1}^{n} Z_k = \frac{1}{n}\sum_{i=1}^{N}\gamma_i y_i .$$

Use this expression to provide an alternative proof for (2.3). **Hint**: $(\gamma_1, \gamma_2, \cdots, \gamma_N)$ follows the Multinomial$(n; p_1, p_2, \cdots, p_N)$ distribution, with $p_1 = \cdots = p_N = 1/N$.

2.4 (Randomized Systematic Sampling) Suppose that $N = nK$ and a complete list of the N units is available as the sampling frame. The *randomized systematic sampling* method selects the sample of size n as follows:

1. Randomize the order of the sequence for the N population units; denote the randomized sequence as $1, 2, \cdots, N$;
2. Select the sample of n units using the simple systematic sampling method with the given randomized population sequence.

Let **S** be the set of n units selected by the above sampling procedure.

(a) Show that the randomized systematic sampling method is equivalent to SRSWOR, i.e., the sampling design satisfies (2.1).
(b) Provide a variance estimator for $\bar{y}$ under simple systematic sampling if it can be assumed that the population sequence does not show any particular pattern. Explain what are the possible patterns of concern.

2.5 (Sample Selection Using the R Software) The R function `sample()` is a convenient tool for drawing survey samples using SRSWOR or SRSWR. The N population units are labelled as $1, 2, \cdots, N$. With pre-specified values of N and n, the R command `sample(N,n)` returns a set of n units drawn from the population by SRSWOR. The command `sample(N,n, replace=TRUE)` returns n units drawn from the population by SRSWR.

(a) (i) Create an arbitrary survey population with the values of the response variable, i.e., $\{y_1, y_2, \cdots, y_N\}$ specified; (ii) Compute the population values μ_y and σ_y^2; (iii) Choose a sample size n; (iv) Take a sample **S** using SRSWOR; (v) Calculate the sample mean $\bar{y}$ and the sample variance s_y^2 using the R functions `mean()` and `var()`; (vi) Repeat steps (iv) and (v) 1000 times, and compute the average of the $\bar{y}$'s and of the s_y^2's from the 1000 simulated samples; (vii) Compare the values of the averages from (vi) to the population values μ_y and σ_y^2.
(b) (i) Take a sample using SRSWR with $N = 10$ and $n = 5$; repeat the process a number of times to see how often the sample contains duplicated units; (ii) Re-do (i) with $N = 50$ and $n = 5$; (iii) Re-do (i) with $N = 100$ and $n = 5$.
(c) Describe how to use the R function `sample()` to take a simple systematic sample.

2.6 (Inverse Sampling for Rare Items) In estimating the population proportion of rare items where $P = M/N$ is small, the following *inverse sampling* procedure can be used. It was first described by Haldane (1945). Let m be a pre-specified integer. We keep sampling, one unit at a time, from the population using SRSWOR (without a pre-determined n) until m rare items are found. Let n be the total number of units sampled when the procedure stops. It is apparent that n is the actual sample size and is a random number. Note that there is a total of M rare items among the N items (units) in the population.

(a) Show that $p = (m - 1)/(n - 1)$ is an unbiased estimator of $P = M/N$ under inverse sampling.
(b) Find $E(n)$, the expected sample size under inverse sampling with the given m.
(c) Show that, by ignoring terms of the order of $1/N$,

$$V(p) \doteq \frac{mP^2(1-P)}{(m-1)^2}\,, \quad \text{where} \quad p = \frac{m-1}{n-1}\,.$$

Hint: For the given m, the sample size n follows a negative hypergeometric distribution with the probability function given by

$$f(k) = P(n=k) = \frac{\binom{M}{m-1}\binom{N-M}{k-m}}{\binom{N}{k-1}} \cdot \frac{M-m+1}{N-k+1}\,, \quad k = m, m{+}1, \cdots, N{-}M{+}m\,.$$

Part (c) is a difficult question. Some details can be found in Govindarajulu (1999).

2.7 (Applications for Sample Size Calculations) Suppose that there are $N = 7000$ public schools in Canada. The goal is to estimate $P = M/N$, where M is the number of schools which have Wireless Internet Access in all school buildings.

(a) Find the required sample size n such that the estimate is accurate within two percentage points, 19 times out of 20 under SRSWOR, assuming that prior information on P is not available.

(b) Re-do (a) assuming $P = 5\%$ for the purpose of sample size calculations.

(c) If inverse sampling is going to be used with the pre-chosen $m = 10$, what is the expected number of schools we have to sample? Assume that $P = 5\%$ for the purpose of planning the survey.

Chapter 3
Stratified Sampling and Cluster Sampling

Complex survey designs involve at least one of the three features: (i) stratification; (ii) clustering; and (iii) unequal probability selection of units. In this chapter we provide some basic results on stratified sampling and cluster sampling. In Sect. 3.5 we provide a brief discussion on stratified two-stage cluster sampling, which reveals the notational complexities for complex surveys. General unequal probability sampling methods will be discussed in the next chapter.

3.1 Stratified Simple Random Sampling

Suppose that the survey population is divided into H non-overlapping strata: $\mathbf{U} = \mathbf{U}_1 \cup \cdots \cup \mathbf{U}_\mathrm{H}$, with corresponding break-down of population size as $N = \sum_{h=1}^{H} N_h$, where N_h is the size of stratum h. For any *stratified sampling* designs, there are two basic features:

- A sample $\mathbf{S}_h$ is taken from stratum h using a chosen sampling design, and this is done for every stratum.
- The H stratum samples $\mathbf{S}_h$, $h = 1, 2, \cdots, H$ are selected independent of each other.

Under stratified sampling, each stratum is viewed as an independent population in terms of survey design and sample selection. Let n_h be the size of the stratum sample $\mathbf{S}_h$. The overall sample size is given by $n = \sum_{h=1}^{H} n_h$. Let $\mathbf{S} = \mathbf{S}_1 \cup \cdots \cup \mathbf{S}_\mathrm{H}$ be the combined stratified sample.

The sampling designs used for selecting $\mathbf{S}_h$ could be different for different strata. The choice of the design depends partially on the availability of sampling frames in each stratum. Suppose that a complete list of N_h units is available and can be used as the sampling frame for stratum h, $h = 1, 2, \cdots, H$. The design is called *stratified simple random sampling* if $\mathbf{S}_h$ is selected by SRSWOR.

C. Wu, M. E. Thompson, *Sampling Theory and Practice*, ICSA Book Series in Statistics, https://doi.org/10.1007/978-3-030-44246-0_3

Estimation of the population mean μ_y under stratified simple random sampling is one of the fundamental pieces for more complex situations. We first introduce the notation for the population parameters under stratified sampling.

3.1.1 Population Parameters

Let y_{hi} be the value of the study variable y for unit i in stratum h, $i = 1, 2, \cdots, N_h$, $h = 1, 2, \cdots, H$. The population mean and the population total for stratum h are given by

$$\mu_{yh} = \frac{1}{N_h}\sum_{i=1}^{N_h} y_{hi} \quad \text{and} \quad T_{yh} = \sum_{i=1}^{N_h} y_{hi} \, .$$

It follows that $T_{yh} = N_h\mu_{yh}$. Let $W_h = N_h/N$ be the *stratum weight*, which is the relative size of the stratum within the overall population. We have

$$\sum_{h=1}^{H} N_h = N \quad \text{and} \quad \sum_{h=1}^{H} W_h = 1 \, .$$

The overall population mean and the population total are given by

$$\mu_y = \frac{1}{N}\sum_{h=1}^{H}\sum_{i=1}^{N_h} y_{hi} \quad \text{and} \quad T_y = \sum_{h=1}^{H}\sum_{i=1}^{N_h} y_{hi} \, .$$

We have the following basic relations among the population parameters:

$$\mu_y = \sum_{h=1}^{H} W_h\mu_{yh} \quad \text{and} \quad T_y = \sum_{h=1}^{H} T_{yh} = \sum_{h=1}^{H} N_h\mu_{yh} \, . \tag{3.1}$$

The stratum population variances are given by

$$\sigma_{yh}^2 = \frac{1}{N_h - 1}\sum_{i=1}^{N_h}\left(y_{hi} - \mu_{yh}\right)^2 , \quad h = 1, 2, \cdots, H \, .$$

The overall population variance is given by

$$\sigma_y^2 = \frac{1}{N-1}\sum_{h=1}^{H}\sum_{i=1}^{N_h}\left(y_{hi} - \mu_y\right)^2 .$$

Treating $(N_h - 1)/(N - 1) \doteq W_h$ and $N_h/(N - 1) \doteq W_h$, we have the following decomposition of the overall population variance:

$$\sigma_y^2 \doteq \sum_{h=1}^{H} W_h \sigma_{yh}^2 + \sum_{h=1}^{H} W_h \big(\mu_{yh} - \mu_y\big)^2 . \tag{3.2}$$

Equation (3.2) is similar to the variance decomposition in the analysis of variance (ANOVA). The "Total Population Variation" σ_y^2 can be decomposed into "Within Strata Variation" $\sum_{h=1}^{H} W_h \sigma_{yh}^2$ and "Between Strata Variation" $\sum_{h=1}^{H} W_h \big(\mu_{yh} - \mu_y\big)^2$. This result will play a crucial role in assessing the efficiency of stratified simple random sampling.

3.1.2 Sample Data

The sample data under stratified sampling are given by $\{y_{hi}, i \in \mathbf{S}_h, h = 1, 2, \cdots, H\}$. The stratum sample mean and the stratum sample variance are defined as

$$\bar{y}_h = \frac{1}{n_h} \sum_{i \in \mathbf{S}_h} y_{hi} \quad \text{and} \quad s_{yh}^2 = \frac{1}{n_h - 1} \sum_{i \in \mathbf{S}_h} \big(y_{hi} - \bar{y}_h\big)^2 ,$$

where n_h is the stratum sample size. Under stratified simple random sampling, we have

$$E\big(\bar{y}_h\big) = \mu_{yh} \quad \text{and} \quad E\big(s_{yh}^2\big) = \sigma_{yh}^2 .$$

It would be the natural next step to give expressions for the overall sample mean $\bar{y}$ and the overall sample variance s_y^2. However, neither $\bar{y}$ nor s_y^2 is a useful statistic for inferences. See Problem 3.1 for further detail.

3.1.3 Estimation of the Overall Population Mean μ_y

Under stratified simple random sampling, the following stratified sample mean estimator is used to estimate the overall population mean μ_y:

$$\bar{y}_{st} = \sum_{h=1}^{H} W_h \bar{y}_h .$$

Note that the stratum weights $W_h = N_h/N$ are known numbers under the sampling design as part of the frame information. Properties of $\bar{y}_{st}$ are summarized in the following theorem.

Theorem 3.1 *Under stratified simple random sampling,*

(a) The stratified sample mean $\bar{y}_{st}$ is an unbiased estimator for the population mean μ_y, i.e.,

$$E(\bar{y}_{st}) = \mu_y \,.$$

(b) The design-based variance of $\bar{y}_{st}$ is given by

$$V\big(\bar{y}_{st}\big) = \sum_{h=1}^{H} W_h^2\Big(1 - \frac{n_h}{N_h}\Big)\frac{\sigma_{yh}^2}{n_h} \,.$$

(c) An unbiased variance estimator for $\bar{y}_{st}$ is given by

$$v\big(\bar{y}_{st}\big) = \sum_{h=1}^{H} W_h^2\Big(1 - \frac{n_h}{N_h}\Big)\frac{s_{yh}^2}{n_h} \,,$$

which satisfies $E\big\{v(\bar{y}_{st})\big\} = V\big(\bar{y}_{st}\big)$.

Proofs of the theorem follow directly from Theorem 2.2 on SRSWOR, equation (3.1) and the independence among the stratum sample means. An unbiased estimator for the population total T_y is given by

$$\hat{T}_y = \sum_{h=1}^{H} N_h \bar{y}_h \,,$$

where the stratum population sizes N_h are known as part of the sampling frame information.

An important special case is the estimation of the population proportion $P = M/N$ under stratified sampling. The stratified sample mean becomes the stratified sample proportion

$$p_{st} = \sum_{h=1}^{H} W_h p_h \,,$$

where $p_h = m_h/n_h$ is the sample proportion for stratum h. The stratum sample variance becomes $s_{yh}^2 = (n_h/(n_h - 1))p_h(1 - p_h)$, which further reduces to $s_{yh}^2 \doteq p_h(1 - p_h)$ for large n_h.

3.1.4 Confidence Intervals

Confidence intervals for the population mean μ_y based on the stratified sample mean can be constructed as

$$\left(\bar{y}_{st} - Z_{\alpha/2}\sqrt{v(\bar{y}_{st})}\,, \quad \bar{y}_{st} + Z_{\alpha/2}\sqrt{v(\bar{y}_{st})}\right),$$

which relies on the asymptotic normality of $(\bar{y}_{st} - \mu_y)/\{v(\bar{y}_{st})\}^{1/2}$, i.e., the central limit theorem for stratified simple random sampling. The asymptotic framework for letting $n \to \infty$ under stratified sampling has two different scenarios:

1. The number of strata, H, is bounded and the sample sizes n_h within each stratum become large;
2. The number of strata becomes large but the strata sample sizes n_h are bounded.

Krewski and Rao (1981) discussed both scenarios and established the central limit theorem for stratified simple random sampling. Bickel and Freedman (1984) also provided relevant results on the topic.

3.1.5 Justifications for Using Stratified Sampling

There are four main reasons to justify the use of stratified sampling designs.

- *Administrative convenience*. A survey at the national level can be organized more conveniently if each province surveys the allocated portion of the sample independently. In this case the provinces would be a natural choice for stratification.
- *Estimation of subpopulation parameters*. Large surveys often have multiple objectives. In addition to estimates for the entire population, estimates for certain subpopulations could also be required. For instance, a survey might be required to provide estimates for the unemployment rates (with certain precision) not only at the national level but also for some well-defined smaller regions. Those regions could be treated as strata, and sample sizes need to be determined at the stratum level.
- *Efficiency considerations*. With suitable stratification and reasonable sample size allocation, stratified sampling can provide more precise estimates and more efficient inferences. See Sect. 3.2 for further details.
- *More balanced samples*. Stratified sampling can protect from possible disproportionate samples under probability sampling. For instance, a sample of $n = 200$ students from a university selected by SRSWOR could contain disproportionally more females than males. This is not a concern from the theoretical point of view but the results might be more acceptable to the general public, and comparisons by gender will be more efficient, if the sample consists of equal numbers of female and male students. Stratification by gender makes it possible to fix the sample sizes for male and females, and thus to make the sample sizes equal.

In practice, it often occurs that the use of stratified sampling is well justified in theory but cannot be implemented because of lack of stratum membership information on the sampling frames. For instance, stratification by gender and age groups can be effective for many surveys of human populations, but separate lists of units for such groups are typically not available. However, once a unit is selected, it is very easy to identify its stratum membership, and the stratum weights W_h may also be available from other sources such as a census. Under such scenarios *post-stratification* can be used where an unstratified sample is selected first and the sample is then post-stratified to obtain more efficient estimators. See Problem 3.7 for further details.

3.2 Sample Size Allocation Under Stratified Sampling

There are two major questions associated with the use of stratified sampling: (i) How to stratify? (ii) How to allocate the sample size n_h for each stratum? The two questions are related and the answers depend on the general objectives of the survey as well as practical constraints such as the overall budget, the required accuracies of estimates and the administrative structure of the survey team. In this section, we focus on sample size allocations with the given overall sample size n or the overall budget, assuming that stratified simple random sampling is used for the survey design. The results also shed light on how to stratify for efficiency considerations and how to allocate sample sizes when more complex stratified sampling designs are to be used.

3.2.1 Proportional Allocation

Suppose that the overall sample size n has already been decided. Suppose also that there is no extra information on the stratified population except for the stratum population sizes N_h, which are part of the frame information required for stratified simple random sampling.

The *proportional allocation* method sets the stratum sample size n_h proportional to the stratum population size N_h, i.e., $n_h \propto N_h$. With the constraint $\sum_{h=1}^{H} n_h = n$ for the given n, the allocation method leads to

$$n_h = \frac{n}{N} N_h = n W_h \,, \qquad h = 1, 2, \cdots, H \,.$$

Under stratified simple random sampling, the general variance formula $V\big(\bar{y}_{st}\big)$ given in Theorem 3.1 reduces to the following form under proportional sample size allocation:

$$V_{prop}\big(\bar{y}_{st}\big) = \Big(1 - \frac{n}{N}\Big)\frac{1}{n}\sum_{h=1}^{H} W_h\sigma_{yh}^2\,, \tag{3.3}$$

where the subscript "$prop$" indicates "proportional allocation".

Let $\bar{y}$ be the sample mean, where the sample has size n and is selected by SRSWOR from the same but un-stratified population. The variance of $\bar{y}$ is given by $V(\bar{y}) = \big(1 - n/N\big)\sigma_y^2/n$. Using the variance decomposition formula given by (3.2), we have the following comparison of variances between SRSWOR and stratified simple random sampling:

$$V\big(\bar{y}\big) - V_{prop}\big(\bar{y}_{st}\big) \doteq \Big(1 - \frac{n}{N}\Big)\frac{1}{n}\sum_{h=1}^{H} W_h\big(\mu_{yh} - \mu_y\big)^2\,. \tag{3.4}$$

Noting that both $\bar{y}$ and $\bar{y}_{st}$ are unbiased estimators of μ_y under the respective sampling design, the result given by (3.4) has two important implications:

1. The stratified simple random sampling design under proportional sample size allocation always provides more efficient estimate of the population mean than SRSWOR.
2. The gain of efficiency under stratified sampling is larger when units within each stratum are more homogeneous, or equivalently, the units from different strata are more heterogeneous so that the between strata variation is large.

It also follows from (3.4) that, when the population strata behave like random groups of the population units, the stratum population means μ_{yh} would be similar to each other, resulting in $\sum_{h=1}^{H} W_h\big(\mu_{yh} - \mu_y\big)^2 \doteq 0$, and consequently $V\big(\bar{y}\big) \doteq V_{prop}\big(\bar{y}_{st}\big)$. In this case stratified simple random sampling under proportional sample size allocation is similar to SRSWOR. As a matter of fact, the two estimators are equivalent with exactly proportional allocation. Interestingly, if the approximation in (3.4) is not made, it is mathematically possible for the variance under SRSWOR to be smaller than the variance under stratified random sampling.

The second implication provides some guidance on how to stratify and use stratified sampling to obtain more efficient estimates for population parameters. The gain in efficiency is associated with the particular study variable, y. For surveys of human populations, it is often the case that the study variables are positively correlated to certain auxiliary variables such as geographical areas, gender or age. Stratification by one or combinations of several auxiliary variables can be effective in many situations. See Problem 3.5 for further details.

3.2.2 Neyman Allocation

When the overall sample size n is fixed, an optimal allocation $(n_1, n_2, \cdots, n_{\mathrm{H}})$ can be found by minimizing $V(\bar{y}_{st})$ subject to the constraint $\sum_{h=1}^{H} n_h = n$. The resulting sample size allocation method is called the *Neyman allocation* (Neyman 1934).

Theorem 3.2 *Under stratified simple random sampling, the Neyman allocation, which minimizes $V(\bar{y}_{st})$ subject to $\sum_{h=1}^{H} n_h = n$, is given by*

$$n_h = n\,\frac{W_h\sigma_{yh}}{\sum_{k=1}^{H} W_k\sigma_{yk}} = n\,\frac{N_h\sigma_{yh}}{\sum_{k=1}^{H} N_k\sigma_{yk}}\,, \quad h = 1, 2, \cdots, H\,, \tag{3.5}$$

and the corresponding minimized variance is given by

$$V_{neym}(\bar{y}_{st}) = \frac{1}{n}\Big(\sum_{h=1}^{H} W_h\sigma_{yh}\Big)^2 - \frac{1}{N}\sum_{h=1}^{H} W_h\sigma_{yh}^2\,, \tag{3.6}$$

where the subscript "neym" indicates "Neyman allocation".

Proof The Neyman allocation can be found by using the Lagrange multiplier method for the constrained minimization problem. Let

$$\begin{aligned} L(n_1, n_2, \cdots, n_{\mathrm{H}}; \lambda) &= V(\bar{y}_{st}) + \lambda\Big(\sum_{h=1}^{H} n_h - n\Big) \\ &= \sum_{h=1}^{H} W_h^2\Big(\frac{1}{n_h} - \frac{1}{N_h}\Big)\sigma_{yh}^2 + \lambda\Big(\sum_{h=1}^{H} n_h - n\Big)\,. \end{aligned}$$

Set $\partial L/\partial n_h = 0$, which leads to $n_h \propto W_h\sigma_{yh}$, $h = 1, 2, \cdots, H$. The sample size formula given by (3.5) follows from the constraint $\sum_{h=1}^{H} n_h = n$. The corresponding minimized variance $V_{neym}(\bar{y}_{st})$ can be obtained by inserting n_h given by (3.5) into the general variance formula for $V(\bar{y}_{st})$ (Problem 3.2). □

Neyman allocation requires information on stratum population variances σ_{yh}^2 and may not be practically useful. Results from Theorem 3.2, however, carry two essential messages:

- Under Neyman allocation, population strata with bigger size N_h or bigger variation (i.e., bigger σ_{yh}^2) or both should be assigned a bigger sample size n_h.
- If all strata have similar variation, i.e., similar values of σ_{yh}^2, Neyman allocation reduces to $n_h \propto W_h$, which is proportional allocation.

3.2.3 Optimal Allocation with Pre-specified Cost or Variance

If the cost for surveying a unit is different for different strata, a more complicated optimal allocation method could be developed to take into account the issue with differential costs. We consider the situation where there is a fixed "indirect" cost c_0 for surveying, and also a per unit cost, which in stratum h is c_h, the same for all units in the stratum. The cost per unit may vary from stratum to stratum. There are two versions of optimal sample size allocation for the current situation:

- The first is to allocate the stratum sample sizes $(n_1, n_2, \cdots, n_{\mathrm{H}})$ to minimize the variance $V(\bar{y}_{st})$ with a pre-specified total cost, C_0.
- The second is to allocate the sample sizes $(n_1, n_2, \cdots, n_{\mathrm{H}})$ to minimize the total cost $c_0 + \sum_{h=1}^{H} c_h n_h$ while controlling the variance $V(\bar{y}_{st})$ at a pre-specified value, V_0.

The constraint on the total cost is given by

$$C_1 = \sum_{h=1}^{H} c_h n_h \, , \tag{3.7}$$

where $C_1 = C_0 - c_0$ is the total direct cost for selecting the sample. The variance formula $V(\bar{y}_{st})$ can be re-written as

$$V\big(\bar{y}_{st}\big) = \sum_{h=1}^{H} W_h^2 \frac{\sigma_{yh}^2}{n_h} - \sum_{h=1}^{H} W_h^2 \frac{\sigma_{yh}^2}{N_h} \, .$$

Since the second term on the right hand side of the equation does not involve n_h, the variance constraint $V_0 = V(\bar{y}_{st})$ with respect to $(n_1, n_2, \cdots, n_{\mathrm{H}})$ can be written as

$$V_1 = \sum_{h=1}^{H} W_h^2 \frac{\sigma_{yh}^2}{n_h} \, , \tag{3.8}$$

where $V_1 = V_0 - \sum_{h=1}^{H} W_h^2 \sigma_{yh}^2 / N_h$. By the Cauchy-Schwarz inequality $(\sum_{i=1}^{n} a_i^2)(\sum_{i=1}^{n} b_i^2) \geq (\sum_{i=1}^{n} a_i b_i)^2$, where the equality holds if $a_i \propto b_i$, we have

$$V_1 C_1 = \Big(\sum_{h=1}^{H} W_h^2 \frac{\sigma_{yh}^2}{n_h}\Big)\Big(\sum_{h=1}^{H} c_h n_h\Big) \geq \Big(\sum_{h=1}^{H} W_h \sigma_{yh} \sqrt{c_h}\Big)^2 \, ,$$

where the equality holds if $W_h \sigma_{yh} / \sqrt{n_h} \propto \sqrt{c_h n_h}$, i.e., $n_h \propto W_h \sigma_{yh} / \sqrt{c_h}$.

The term $\big(\sum_{h=1}^{H} W_h \sigma_{yh} \sqrt{c_h}\big)^2$ does not involve n_h and therefore is the minimum value of $V_1 C_1$ with respect to $(n_1, n_2, \cdots, n_{\mathrm{H}})$. It follows that

$$n_h = n \, \frac{W_h \sigma_{yh}/\sqrt{c_h}}{\sum_{k=1}^{H} W_k \sigma_{yk}/\sqrt{c_k}} \, , \quad h = 1, 2, \cdots, H$$

gives the solution to either minimizing V_1 with C_1 fixed or minimizing C_1 with V_1 fixed, where the overall sample size n is not pre-specified but rather depends on the given C_0 or V_0 (Problem 3.3).

The optimal allocation requires information on stratum variances at the population level. Some useful practical rules for sample size allocation involving cost considerations are as follows:

- With unequal costs for different strata, the more expensive strata should be assigned smaller sample sizes.
- With equal cost for all strata, the two versions of optimal allocation both reduce to Neyman allocation, and hence the stratum sample size n_h is decided by the stratum population size N_h and the stratum variance σ_{yh}^2.
- With equal or nearly equal cost and no information on stratum variations, proportional allocation would be the natural choice for sample size allocation.

3.3 Single-Stage Cluster Sampling

In many practical situations, units of the survey population are naturally attached to groups called *clusters*, where the clusters do not overlap each other and together cover the entire survey population. More importantly, a complete list of all units in the population under such situations may not be available as the sampling frame or too expensive to be constructed. Instead, a complete list of clusters is readily available and can be used as a sampling frame.

3.3.1 Notation for Cluster Sampling

Let K be the total number of clusters in the population. Let M_i be the number of units in cluster i. The overall population size is

$$N = \sum_{i=1}^{K} M_i \, .$$

Let y_{ij} be the value of the y-variable for unit j in cluster i, $j = 1, 2, \cdots, M_i$ and $i = 1, 2, \cdots, K$. The mean and the total for the ith cluster are given by

$$\mu_i = \frac{1}{M_i} \sum_{j=1}^{M_i} y_{ij} \quad \text{and} \quad T_i = \sum_{j=1}^{M_i} y_{ij} \, , \quad i = 1, 2, \cdots, K \, .$$

The population total is given by

$$T_y = \sum_{i=1}^{K}\sum_{j=1}^{M_i} y_{ij} = \sum_{i=1}^{K} T_i = \sum_{i=1}^{K} M_i\mu_i \tag{3.9}$$

and the population mean is given by $\mu_y = T_y/N$.

Under cluster sampling, certain information may be available as part of the sampling frame:

- The total number of clusters, K, is always known as part of the first stage sampling frame.
- The number of units in the ith cluster, M_i, is also known under single-stage or two-stage cluster sampling if cluster i is selected in the first stage sample, since M_i is needed as part of the second stage sampling frame information to select and observe individual units within the cluster.
- For clusters not selected in the first stage sample, the cluster size M_i may not be available.
- The overall population size $N = \sum_{i=1}^{K} M_i$ may not be available, depending on whether all M_i are known or not.

When the population size N is unknown, estimation of T_y and μ_y needs to be handled separately.

3.3.2 Single-Stage Cluster Sampling with Clusters Selected by SRSWOR

We consider the following single-stage cluster sampling procedures:

1. Select k clusters from the list of K clusters using SRSWOR, with a pre-specified k. Let $\mathbf{S}_c$ be the set of labels for the k selected clusters.
2. For $i \in \mathbf{S}_c$, select all M_i units for the final sample.

The subscript "c" in $\mathbf{S}_c$ indicates "clusters". The total number of units in the final sample is given by $n = \sum_{i\in\mathbf{S}_c} M_i$. The sample data on the y-variable can be denoted as

$$\left\{y_{ij} : j = 1, 2, \cdots, M_i\,,\ i \in \mathbf{S}_c\right\}.$$

Under single-stage cluster sampling, there are no random selections within the selected cluster i, since all M_i units are selected and measurements on survey variables are taken. The most crucial feature of the sample data is that, for any $i \in \mathbf{S}_c$, the cluster size M_i, the cluster total T_i and the cluster mean μ_i are all available.

3.3.3 Estimation of the Population Total T_y

Using the expressions from (3.9), the population total can be re-written as

$$T_y = \sum_{i=1}^{K} T_i = K\Big(\frac{1}{K}\sum_{i=1}^{K} T_i\Big).$$

We can view T_i as a measurement on the ith cluster, and $\mu_T = K^{-1}\sum_{i=1}^{K} T_i$ can be viewed as the population mean of the measurement, i.e., of the cluster totals T_i, with K as the "population size". If the k clusters in $\mathbf{S}_c$ are selected from the list of K clusters by SRSWOR, the "sample mean" $\hat{\mu}_T = k^{-1}\sum_{i\in\mathbf{S}_c} T_i$ would be an unbiased estimator of the "population mean" μ_T. More formally, we have the following results for single-stage cluster sampling.

Theorem 3.3 *Under single-stage cluster sampling with clusters selected by SRSWOR,*

(a) An unbiased estimator for the population total T_y is given by

$$\hat{T}_y = K\Big(\frac{1}{k}\sum_{i\in\mathbf{S}_c} T_i\Big) = K\hat{\mu}_T\,,$$

where $\hat{\mu}_T = k^{-1}\sum_{i\in\mathbf{S}_c} T_i$ is the sample mean of cluster totals.

(b) The design-based variance of $\hat{T}_y$ is given by

$$V\big(\hat{T}_y\big) = K^2\Big(1-\frac{k}{K}\Big)\frac{\sigma_T^2}{k}\,,$$

where $\sigma_T^2 = (K-1)^{-1}\sum_{i=1}^{K}\big(T_i-\mu_T\big)^2$, and $\mu_T = K^{-1}\sum_{i=1}^{K} T_i$ is the population mean of cluster totals.

(c) An unbiased variance estimator for $\hat{T}_y$ is given by

$$v\big(\hat{T}_y\big) = K^2\Big(1-\frac{k}{K}\Big)\frac{s_T^2}{k}\,,$$

where $s_T^2 = (k-1)^{-1}\sum_{i\in\mathbf{S}_c}\big(T_i-\hat{\mu}_T\big)^2$ and $\hat{\mu}_T = k^{-1}\sum_{i\in\mathbf{S}_c} T_i$.

Proofs of the above results follow directly from Theorem 2.2 on SRSWOR if we treat each cluster as a unit with T_i as the value of the response variable. The "population size" is K and the "sample size" is k.

It becomes clear that the response values y_{ij} for individual units within clusters do not play any direct roles in the estimation. Only the cluster totals T_i are used directly for estimation under single-stage cluster sampling. The overall sample size $n = \sum_{i \in \mathbf{S}_c} M_i$ does not play the same role as we saw in Theorem 2.2 on SRSWOR.

3.3.4 Estimation of the Population Mean μ_y

Under single-stage cluster sampling with clusters selected by SRSWOR, the population total T_y can be unbiasedly estimated by $\hat{T}_y = K\hat{\mu}_T$, where $\hat{\mu}_T = k^{-1} \sum_{i \in \mathbf{S}_c} T_i$. When the population size N is known, an unbiased estimator of the population mean μ_y is given by $\hat{\mu}_y = \hat{T}_y/N$, with an unbiased variance estimator given by $v(\hat{\mu}_y) = v(\hat{T}_y)/N^2$. The variance estimator $v(\hat{T}_y)$ is given in Theorem 3.3.

When $N = \sum_{i=1}^K M_i$ is unknown, as is often the case in practice for cluster sampling designs, we notice that the population mean can be re-written as

$$\mu_y = \frac{1}{N} \sum_{i=1}^{K} T_i = \frac{\mu_T}{\mu_M} ,$$

where $\mu_M = K^{-1} \sum_{i=1}^K M_i$ is the average cluster size of the population. Noting that M_i is available for $i \in \mathbf{S}_c$, we can estimate μ_M by $\hat{\mu}_M = k^{-1} \sum_{i \in \mathbf{S}_c} M_i$. It follows that we can estimate μ_y by

$$\hat{\mu}_y = \frac{\hat{\mu}_T}{\hat{\mu}_M} = \frac{\sum_{i \in \mathbf{S}_c} T_i}{\sum_{i \in \mathbf{S}_c} M_i} = \frac{1}{n} \sum_{i \in \mathbf{S}_c} \sum_{j=1}^{M_i} y_{ij} , \tag{3.10}$$

where $n = \sum_{i \in \mathbf{S}_c} M_i$ is the overall sample size.

The estimator $\hat{\mu}_y$ given by (3.10) is indeed the simple sample mean for all units selected in the final sample. The design-based properties of $\hat{\mu}_y$, however, do not follow from the results on SRSWOR specified in Theorem 2.2. Under the current single-stage cluster sampling design, the overall sample size n is usually a random number and depends on the M_i's for the sampled clusters. We will provide further details on the design-based properties of $\hat{\mu}_y = \hat{\mu}_T/\hat{\mu}_M$ in Chap. 5 using the general results from ratio estimators. We will see later that even when N is known, the estimator $\hat{\mu}_y = \hat{\mu}_T/\hat{\mu}_M$ may in some cases be more efficient than the estimator $\hat{\mu}_y = \hat{T}_y/N$.

3.3.5 A Comparison Between SRSWOR and Cluster Sampling

Cluster sampling can be cost effective due to the selection of units in the same geographical area or other affiliations among the selected units, which might make it more convenient for the field workers to collect the sample data. A major price paid for the low cost and practical convenience is the loss of efficiency in estimation under certain situations.

Consider a simple scenario where all clusters have the same size, i.e., $M_i = M$ for all i. The population size in this case is $N = KM$ and the final sample size is $n = kM$. Under single-stage cluster sampling with clusters selected by SRSWOR, the estimator for the population mean μ_y reduces to

$$\hat{\mu}_y = \frac{1}{M}\frac{1}{k}\sum_{i\in\mathbf{S}_c} T_i = \frac{1}{M}\hat{\mu}_T\,,$$

with design-based variance given by

$$V\big(\hat{\mu}_y\big) = \frac{1}{M^2}\Big(1-\frac{k}{K}\Big)\frac{\sigma_T^2}{k}\,.$$

Using the result $M^{-1}\sigma_T^2 \doteq \sigma_y^2\{1+(M-1)\rho\}$ from Problem 3.7, we have

$$V\big(\hat{\mu}_y\big) \doteq \Big(1-\frac{n}{N}\Big)\frac{\sigma_y^2}{n}\big[1+(M-1)\rho\big]\,,$$

where $N = KM, n = kM$ and ρ is the so-called *intra-cluster correlation coefficient* (see Problem 3.8).

Suppose that $\bar{y}$ is the sample mean, where the sample of size $n = kM$ is selected from the same but unclustered population by SRSWOR. It follows immediately that $V\big(\hat{\mu}_y\big) < V(\bar{y})$ if and only if $\rho < 0$, assuming $M \geq 2$.

- It is very common in survey practice that units within the same cluster are positively correlated, i.e., $\rho > 0$, and consequently single-stage cluster sampling is less efficient than SRSWOR.
- For situations where $\rho < 0$, cluster sampling can be more efficient. For household surveys we can treat couples (husband and wife; common-law partners; etc.) as clusters with size $M_i = 2$. The values of the response variable such as income for the two bodies in the cluster might be negatively correlated to each other. In this case cluster sampling would be more desirable than SRSWOR.
- When $\rho = 0$, the clusters behave like random groups of units from the population. Under such scenarios single-stage cluster sampling will result in a final sample which is similar to the one selected by non-cluster sampling methods.

3.4 Two-Stage Cluster Sampling

When the cluster sizes M_i, $i = 1, 2, \cdots, K$ are not small, measuring all units in selected clusters may not be desirable. Two-stage cluster sampling designs are often used in such scenarios. For each selected cluster, a second-stage sample of units is selected within the cluster. The first-stage clusters are called the *primary sampling unit* (*psu*); the units within clusters are called the *secondary sampling unit* (*ssu*).

3.4.1 Two-Stage Cluster Sampling with SRSWOR at Both Stages

We consider the following two-stage cluster sampling procedures:

1. Select k clusters from the list of K clusters using SRSWOR, with a pre-specified k. Let $\mathbf{S}_c$ be the set of labels for the k selected clusters.
2. For $i \in \mathbf{S}_c$ and a pre-specified m_i, select a second-stage sample $\mathbf{S}_i$ of m_i units from the list of M_i units in cluster i using SRSWOR; the processes are carried out independently for different clusters.

The final sample size is $n = \sum_{i \in \mathbf{S}_c} m_i$. The sample data for the y-variable are denoted as

$$\left\{ y_{ij} : \ j \in \mathbf{S}_i \, , \ i \in \mathbf{S}_c \right\} .$$

Under two-stage cluster sampling, neither the cluster total T_i nor the cluster mean μ_i is available even if cluster i is selected in the first stage sample $\mathbf{S}_c$. The following sampling frames are required for the method:

- A complete list of the K clusters in the population.
- A complete list of M_i units in cluster i, for all clusters selected in the first stage (i.e., $i \in \mathbf{S}_c$).

Those are the same types of sampling frames as required for single-stage cluster sampling.

The second-stage sample mean and sample variance of the data collected from cluster i are computed as

$$\bar{y}_i = \frac{1}{m_i} \sum_{j \in \mathbf{S}_i} y_{ij} \quad \text{and} \quad s_i^2 = \frac{1}{m_i - 1} \sum_{j \in \mathbf{S}_i} \left(y_{ij} - \bar{y}_i \right)^2 .$$

Since $\mathbf{S}_i$ is a sample of size m_i selected by SRSWOR from the cluster of M_i units, it follows from Theorem 2.2 that

$$E(\bar{y}_i) = \mu_i \quad \text{and} \quad V(\bar{y}_i) = \left(1 - \frac{m_i}{M_i}\right) \frac{\sigma_i^2}{m_i} ,$$

where $\mu_i = M_i^{-1} \sum_{j=1}^{M_i} y_{ij}$ is the cluster mean and $\sigma_i^2 = (M_i - 1)^{-1} \sum_{j=1}^{M_i} (y_{ij} - \mu_i)^2$ is the cluster variance. In addition, the second-stage sample variance s_i^2 is an unbiased estimator of the cluster variance σ_i^2 under SRSWOR, i.e., $E(s_i^2) = \sigma_i^2$.

As part of the sampling frame information, the total number of clusters, K, is known. The cluster size M_i is also known for $i \in \mathbf{S}_c$ but may not be available otherwise. The overall population size $N = \sum_{i=1}^K M_i$ may not be available, depending on the situation for the M_i's.

3.4.2 Estimation of the Population Total T_y

The cluster total $T_i = M_i \mu_i$ can be unbiasedly estimated by $\hat{T}_i = M_i \bar{y}_i$ for $i \in \mathbf{S}_c$. Following the notation in Sect. 3.3, we have $T_y = K\mu_T$ where $\mu_T = K^{-1}\sum_{i=1}^K T_i$ is the population mean of cluster totals. The key is to construct an estimator of μ_T using the two-stage sample data.

Theorem 3.4 *Under two-stage cluster sampling with SRSWOR at both stages,*

(a) An unbiased estimator for the population mean of cluster totals, μ_T, is given by

$$\tilde{\mu}_T = \frac{1}{k}\sum_{i \in \mathbf{S}_c} \hat{T}_i = \frac{1}{k}\sum_{i \in \mathbf{S}_c} M_i \bar{y}_i \,.$$

(b) The design-based variance of $\tilde{\mu}_T$ is given by

$$V\big(\tilde{\mu}_T\big) = \Big(1 - \frac{k}{K}\Big)\frac{\sigma_T^2}{k} + \frac{1}{k}\,\frac{1}{K}\sum_{i=1}^K M_i^2 \Big(1 - \frac{m_i}{M_i}\Big)\frac{\sigma_i^2}{m_i}\,,$$

where $\sigma_T^2 = (K-1)^{-1}\sum_{i=1}^K \big(T_i - \mu_T\big)^2$ and σ_i^2 is the cluster variance.

(c) An unbiased variance estimator for $\tilde{\mu}_T$ is given by

$$v\big(\tilde{\mu}_T\big) = \Big(1 - \frac{k}{K}\Big)\frac{\hat{\sigma}_T^2}{k} + \frac{1}{K}\,\frac{1}{k}\sum_{i \in \mathbf{S}_c} M_i^2 \Big(1 - \frac{m_i}{M_i}\Big)\frac{s_i^2}{m_i}\,,$$

where $\hat{\sigma}_T^2 = (k-1)^{-1}\sum_{i \in \mathbf{S}_c} \big(\hat{T}_i - \tilde{\mu}_T\big)^2$ and s_i^2 is the cluster sample variance.

Proof Let E_1 and V_1 be the expectation and variance under the first stage sampling design. Let E_2 and V_2 be the conditional expectation and variance under the second stage sampling design given the first stage sample, $\mathbf{S}_c$. Part (a) follows from

$$E\big(\tilde{\mu}_T\big) = E_1\Big\{\frac{1}{k}\sum_{i\in\mathbf{S}_c} M_i E_2(\bar{y}_i)\Big\} = E_1\Big(\frac{1}{k}\sum_{i\in\mathbf{S}_c} T_i\Big) = \mu_T\,.$$

Part (b) of the theorem follows from

$$\begin{aligned} V\big(\tilde{\mu}_T\big) &= V_1\big\{E_2\big(\tilde{\mu}_T\big)\big\} + E_1\big\{V_2\big(\tilde{\mu}_T\big)\big\} \\ &= V_1\Big(\frac{1}{k}\sum_{i\in\mathbf{S}_c} T_i\Big) + E_1\Big\{\frac{1}{k^2}\sum_{i\in\mathbf{S}_c} M_i^2 V_2(\bar{y}_i)\Big\} \\ &= \Big(1-\frac{k}{K}\Big)\frac{\sigma_T^2}{k} + \frac{1}{k}\,E_1\Big\{\frac{1}{k}\sum_{i\in\mathbf{S}_c} M_i^2\Big(1-\frac{m_i}{M_i}\Big)\frac{\sigma_i^2}{m_i}\Big\} \\ &= \Big(1-\frac{k}{K}\Big)\frac{\sigma_T^2}{k} + \frac{1}{k}\,\frac{1}{K}\sum_{i=1}^{K} M_i^2\Big(1-\frac{m_i}{M_i}\Big)\frac{\sigma_i^2}{m_i}\,, \end{aligned}$$

where $\sigma_T^2 = (K-1)^{-1}\sum_{i=1}^{K}\big(T_i-\mu_T\big)^2$ and $\sigma_i^2 = (M_i-1)^{-1}\sum_{j=1}^{M_i}(y_{ij}-\mu_i)^2$ is the cluster variance.

To prove part (c), we first let $W = K^{-1}\sum_{i=1}^{K} M_i^2(1-m_i/M_i)\sigma_i^2/m_i$. We can then rewrite $V\big(\tilde{\mu}_T\big)$ as

$$V\big(\tilde{\mu}_T\big) = \Big(\frac{1}{k}-\frac{1}{K}\Big)\sigma_T^2 + \frac{1}{k}W\,.$$

It is straightforward to show that $\hat{W} = k^{-1}\sum_{i\in\mathbf{S}_c} M_i^2\big(1-m_i/M_i\big)s_i^2/m_i$ is an unbiased estimator for W, i.e., $E(\hat{W}) = W$. Let $\hat{\sigma}_T^2 = (k-1)^{-1}\sum_{i\in\mathbf{S}_c}\big(\hat{T}_i-\tilde{\mu}_T\big)^2$. It is shown in Problem 3.11 that $E(\hat{\sigma}_T^2) = \sigma_T^2 + W$. If we let

$$v\big(\tilde{\mu}_T\big) = \Big(\frac{1}{k}-\frac{1}{K}\Big)\hat{\sigma}_T^2 + \frac{1}{K}\hat{W}\,,$$

it follows immediately that $E\big\{v\big(\tilde{\mu}_T\big)\big\} = V\big(\tilde{\mu}_T\big)\,.$ □

From part (b) of Theorem 3.4 we see that the variance of $\tilde{\mu}_T$ based on two-stage cluster samples consists of two components. The first one is $\big(1-k/K\big)\sigma_T^2/k$, which is the variance component due to the first stage sample selection of clusters. The second one is W/k, where $W = K^{-1}\sum_{i=1}^{K} M_i^2(1-m_i/M_i)\sigma_i^2/m_i$, which is the variance component due to the second stage sample selection of units within the k selected clusters. The second component becomes small when m_i is close to M_i but could be a substantial part of $V(\tilde{\mu}_T)$ otherwise. Ignoring the second stage sample selection of units will result in serious under-evaluation of the variance.

Note that in the formula for estimation of the variance, the first term estimates not only the first term in the variance but a substantial part of the second term in the variance as well. The first term in the variance estimator is of order $1/k$ while the second term in the variance estimator is of order $1/K$. If K/k is large, the second term in the variance estimator becomes negligible.

An unbiased estimator for the population total T_y is given by $\tilde{T}_y = K\tilde{\mu}_T$, with an unbiased variance estimator given by $v\big(\tilde{T}_y\big) = K^2 v\big(\tilde{\mu}_T\big)$.

3.4.3 Estimation of the Population Mean μ_y

The situation is similar to that of single-stage cluster sampling. If the overall population size N is known, an unbiased estimator for μ_y is given by $\tilde{\mu}_y = \tilde{T}_y/N$, with an unbiased variance estimator given by $v\big(\tilde{\mu}_y\big) = v\big(\tilde{T}_y\big)/N^2$. If N is unknown, we need to estimate $N = \sum_{i=1}^{K} M_i$ using the known M_i for $i \in \mathbf{S}_c$. The final estimator of μ_y is given by a ratio estimator $\tilde{\mu}_y = \tilde{\mu}_T/\hat{\mu}_M$, where $\tilde{\mu}_T = k^{-1}\sum_{i\in\mathbf{S}_c} M_i\bar{y}_i$ and $\hat{\mu}_M = k^{-1}\sum_{i\in\mathbf{S}_c} M_i$. Again, even if N is known, it is sometimes more efficient to use the ratio estimator.

3.5 Stratified Two-Stage Cluster Sampling

Stratified multistage sampling designs are commonly used for large scale complex surveys. We present results for stratified two-stage cluster sampling, with SRSWOR used at both stages for each stratum. The main focus is to highlight the notational complexities for the design. The theoretical development is a straightforward combination of the results presented in Sects. 3.1 and 3.4.

Suppose that the survey population has a stratified clustered structure. Let H be the total number of strata. Let N_h be the stratum population size and $W_h = N_h/N$ be the stratum weight, where $N = \sum_{h=1}^{H} N_h$ is the overall population size. Let

- K_h be the total number of clusters in stratum h;
- M_{hi} be the total number of units in cluster i within stratum h;
- y_{hij} be the value of the y-variable for unit j in cluster i within stratum h.

The population total is given by

$$T_y = \sum_{h=1}^{H}\sum_{i=1}^{K_h}\sum_{j=1}^{M_{hi}} y_{hij} = \sum_{h=1}^{H} T_{yh}\,, \quad \text{where} \quad T_{yh} = \sum_{i=1}^{K_h}\sum_{j=1}^{M_{hi}} y_{hij}\,.$$

The population mean is given by

$$\mu_y = \frac{T_y}{N} = \sum_{h=1}^{H} W_h\mu_{yh}\,, \quad \text{where} \quad \mu_{yh} = \frac{T_{yh}}{N_h}\,.$$

Following the general principles of estimation under stratified sampling, the population total and mean can be estimated by

$$\hat{T}_y = \sum_{h=1}^{H} \hat{T}_{yh} \quad \text{and} \quad \hat{\mu}_y = \sum_{h=1}^{H} W_h \hat{\mu}_{yh} \, ,$$

where $\hat{T}_{yh}$ and $\hat{\mu}_{yh}$ are estimates of the stratum population total and mean using survey data from stratum h. For stratified multistage sampling, the stratum population size N_h may not be known. In this case N_h, N and W_h also need to be estimated.

Suppose that two-stage cluster sampling, with SRSWOR at both stages, is used within each stratum. Let

- k_h be the number of clusters selected in stratum h;
- $\mathbf{S}_{ch}$ be the set of labels of the k_h selected clusters in stratum h;
- m_{hi} be the number of units selected in cluster $i \in \mathbf{S}_{ch}$;
- $\mathbf{S}_{hi}$ be the set of m_{hi} selected units for $i \in \mathbf{S}_{ch}$;
- $\bar{y}_{hi} = m_{hi}^{-1} \sum_{j \in \mathbf{S}_{hi}} y_{hij}$ be the cluster sample mean;
- $s_{hi}^2 = (m_{hi} - 1)^{-1} \sum_{j \in \mathbf{S}_{hi}} (y_{hij} - \bar{y}_{hi})^2$ be the cluster sample variance.

An unbiased estimator of T_{yh} is given by

$$\hat{T}_{yh} = K_h \cdot \frac{1}{k_h} \sum_{i \in \mathbf{S}_{ch}} M_{hi} \bar{y}_{hi} \, .$$

Note that K_h and M_{hi}, $i \in \mathbf{S}_{ch}$ are all known as part of the sampling frame information under stratified two-stage sampling. If the stratum population size $N_h = \sum_{i=1}^{K_h} M_{hi}$ is unknown, it can be unbiasedly estimated by

$$\hat{N}_h = K_h \cdot \frac{1}{k_h} \sum_{i \in \mathbf{S}_{ch}} M_{hi} \, .$$

It can easily be argued that $\hat{T}_y = \sum_{h=1}^{H} \hat{T}_{yh}$ is an unbiased estimator of T_y. For variance estimation, we have

$$v\big(\hat{T}_y\big) = \sum_{h=1}^{H} v\big(\hat{T}_{yh}\big) \, ,$$

where $v\big(\hat{T}_{yh}\big) = K_h^2 v\big(\tilde{\mu}_{Th}\big)$, and $v\big(\tilde{\mu}_{Th}\big)$ can be constructed based on Part (c) of Theorem 3.4 using data from stratum h, with parallel notation

$$\mu_{Th} = \frac{1}{K_h}\sum_{i=1}^{K_h} T_{hi} \quad \text{and} \quad \tilde{\mu}_{Th} = \frac{1}{k_h}\sum_{i\in \mathbf{S}_{ch}} \hat{T}_{hi} \,,$$

where $T_{hi} = \sum_{j=1}^{M_{hi}} y_{hij}$ and $\hat{T}_{hi} = M_{hi}\bar{y}_{hi}$.

3.6 Problems

3.1 (Stratified Simple Random Sampling) Consider the overall sample mean from a stratified simple random sample given by

$$\bar{y} = \frac{1}{n}\sum_{h=1}^{H}\sum_{i\in \mathbf{S}_h} y_{hi} \,,$$

where $n = \sum_{h=1}^{H} n_h$ is the overall sample size. Argue that $\bar{y}$ is not a design unbiased estimator of the population mean μ_y unless the sample sizes n_h are allocated proportional to N_h.

3.2 (Neyman Allocation for Stratified Sampling)

(a) Show that the minimum variance $V_{neym}\big(\bar{y}_{st}\big)$ under Neyman allocation is given by (3.6).

(b) Argue without using any technical details that $V_{neym}\big(\bar{y}_{st}\big) \le V_{prop}\big(\bar{y}_{st}\big)$.

(c) Show that

$$V_{prop}\big(\bar{y}_{st}\big) - V_{neym}\big(\bar{y}_{st}\big) = \frac{1}{n}\sum_{h=1}^{H} W_h\big(\sigma_{yh} - \bar{\sigma}_y\big)^2 \,,$$

where $\bar{\sigma}_y = \sum_{h=1}^{H} W_h\sigma_{yh}$.

(d) Discuss scenarios under which Neyman allocation would provide much more efficient estimate for μ_y than proportional allocation.

3.3 (Optimal Allocation with Fixed C_0 or V_0) The optimal allocation is given by $n_h \propto W_h\sigma_{yh}/\sqrt{c_h}$. Let n be the overall sample size, i.e., $n = \sum_{h=1}^{H} n_h$. It follows that

$$n_h = n\,\frac{W_h\sigma_{yh}/\sqrt{c_h}}{\sum_{k=1}^{H} W_k\sigma_{yk}/\sqrt{c_k}} = n\,\frac{N_h\sigma_{yh}/\sqrt{c_h}}{\sum_{k=1}^{H} N_k\sigma_{yk}/\sqrt{c_k}} \,.$$

The overall sample size n, however, is not predetermined. It depends on the fixed C_0 or V_0.

(a) Show that, with fixed C_0, the optimal allocation under constraint (3.7) leads to

$$n = C_1\Big(\sum_{k=1}^{H} W_k\sigma_{yk}/\sqrt{c_k}\Big)/\Big(\sum_{k=1}^{H} W_k\sigma_{yk}\sqrt{c_k}\Big)\,.$$

(b) Show that, with fixed V_0, the optimal allocation under constraint (3.8) leads to

$$n = \Big(\sum_{k=1}^{H} W_k\sigma_{yk}/\sqrt{c_k}\Big)\Big(\sum_{k=1}^{H} W_k\sigma_{yk}\sqrt{c_k}\Big)/\,V_1\,.$$

3.4 (Sample Size Allocation for Stratified Sampling) Consider a hypothetical situation where we have a stratified population with $H = 3$, $N_1 = 500$, $N_2 = 1000$ and $N_3 = 1500$. Suppose that we know the stratum population variances $\sigma_{y1}^2 = 100$, $\sigma_{y2}^2 = 200$ and $\sigma_{y3}^2 = 225$. The goal is to make comparisons among different sample size allocations for stratified simple random sampling.

(a) Let $n = 84$. Find the proportional sample size allocation and compute $V_{prop}(\bar{y}_{st})$.
(b) Let $n = 84$. Find the Neyman allocation for stratum sample sizes and compute $V_{neym}(\bar{y}_{st})$.
(c) Suppose that the three stratum population means μ_{yh}, $h = 1, 2, 3$ are all equal, and we take a sample of size $n = 84$ from the (unstratified) population using SRSWOR. Let $\bar{y}$ be the sample mean. Find $V(\bar{y})$. **Hint**: Find σ_y^2 first.
(d) Compare the values of $V_{prop}(\bar{y}_{st})$, $V_{neym}(\bar{y}_{st})$, $V(\bar{y})$ and make appropriate comments.

3.5 (Neyman Allocation Based on an Auxiliary Variable) Neyman allocation requires the knowledge of stratum population variances σ_{yh}^2, $h = 1, 2, \cdots, H$ of the study variable y. In practical situations those quantities are unknown, but information on an auxiliary variable x can often be obtained from other sources. For simplicity, we assume that the coefficient of variation σ_{yh}/μ_{yh} is a constant across all strata, and y_{hi}/x_{hi} is also a constant for all elements in the population.

(a) Show that the Neyman allocation is given by $n_h = nT_{xh}/T_x$, where $T_{xh} = \sum_{i=1}^{N_h} x_{hi}$ and $T_x = \sum_{h=1}^{H}\sum_{i=1}^{N_h} x_{hi} = \sum_{h=1}^{H} T_{xh}$.
(b) Show that, for some constant c,

$$V_{neym}\big(\bar{y}_{st}\big) = c\Big(\frac{1}{n} - \frac{1}{N}\Big)\mu_x^2 - \frac{c}{N^2}\sum_{h=1}^{H} N_h\big(\mu_{xh} - \mu_x\big)^2\,,$$

where $\mu_x = T_x/N$.
(c) If we can choose to stratify the population by the values of the x variable, what kind of stratification will make $V_{neym}(\bar{y}_{st})$ small and hence result in efficient stratified sampling?

3.6 (Simple Systematic Sampling) Suppose that $N = nK$, where n is the sample size. Argue that the simple systematic sampling method described in Sect. 2.3 is a special case of single-stage cluster sampling.

3.7 (Post-Stratification) Let $\mathbf{S}$ be a sample of size n selected by SRSWOR from a stratified population. The stratified structure of the population is not used for sample selection due to lack of information about stratum membership of units on the sampling frame. Suppose that the stratum membership can be correctly identified for each $i \in \mathbf{S}$ and the information on the stratum weights W_h is also available. The sample $\mathbf{S}$ can be post-stratified as $\mathbf{S} = \mathbf{S}_1 \cup \cdots \cup \mathbf{S}_{\mathrm{H}}$ with corresponding break-down of the sample size $n = n_1 + \cdots + n_{\mathrm{H}}$. Note that the post-stratified sample sizes n_h are random numbers. Let $\bar{y}_h$ and s_{yh}^2 be the sample mean and sample variance computed from the post-stratified samples. The post-stratified sample mean is defined as

$$\bar{y}_{post} = \sum_{h=1}^{H} W_h \bar{y}_h \,.$$

(a) Show that the post-stratified sample mean is a design-unbiased estimator of μ_y.

(b) Show that the variance of the post-stratified mean is given by

$$V(\bar{y}_{post}) = \sum_{h=1}^{H} W_h^2 \Big\{ E\Big(\frac{1}{n_h}\Big) - \frac{1}{N_h} \Big\} \sigma_{yh}^2 \,.$$

(c) Show that $E(n_h) = nW_h$ and argue that an approximately unbiased variance estimator is given by

$$v(\bar{y}_{post}) = \Big(1 - \frac{n}{N}\Big) \frac{1}{n} \sum_{h=1}^{H} W_h s_{yh}^2 \,.$$

(d) Use the result of (c) to argue that the post-stratified mean is likely to be more efficient than the simple sample mean $\bar{y}$.

3.8 (Intra-Cluster Correlation Coefficient) Suppose that the survey population consists of K clusters of equal size $M_i = M$. The overall population size is $N = KM$. Suppose that we randomly select a cluster from the population, and then randomly select two units without replacement from the selected cluster. Let Z_1 and Z_2 be the y-values of the two selected units. The *intra-cluster correlation coefficient* is defined as

$$\rho = \frac{Cov(Z_1, Z_2)}{\sqrt{V(Z_1)V(Z_2)}} \,.$$

(a) Show that $E(Z_1) = E(Z_2) = \mu_y$, $V(Z_1) = V(Z_2) = (N-1)\sigma_y^2/N$ and

$$Cov(Z_1, Z_2) = \frac{1}{KM(M-1)} \sum_{i=1}^{K} \sum_{j=1}^{M} \sum_{l \neq j}^{M} (y_{ij} - \mu_y)(y_{il} - \mu_y) .$$

(b) Show that

$$\sigma_T^2 = \frac{KM-1}{K-1} \sigma_y^2 \{1 + (M-1)\rho\} ,$$

where $\sigma_T^2 = (K-1)^{-1} \sum_{i=1}^{K} (T_i - \mu_T)^2$. This leads to $M^{-1}\sigma_T^2 \doteq \sigma_y^2\{1 + (M - 1)\rho\}$ if we treat $(KM-1)/\{M(K-1)\} \doteq 1$.

3.9 (Estimation of σ_y^2 Under Single-Stage Cluster Sampling) Consider single-stage cluster sampling with clusters of equal size ($M_i = M$). The k clusters are selected by SRSWOR from the K clusters in the population. The overall population size is $N = KM$ and the overall sample size is $n = kM$. Let the "within cluster variance" and the "between cluster variance" be respectively defined as

$$\sigma_W^2 = \frac{1}{K} \sum_{i=1}^{K} \sigma_i^2 \quad \text{and} \quad \sigma_B^2 = \frac{1}{K-1} \sum_{i=1}^{K} \left(\mu_i - \mu_y\right)^2 ,$$

where σ_i^2 are the cluster variances and μ_i are the cluster means. Note that the μ_i and the σ_i^2 are available for $i \in \mathbf{S}_c$ under single-stage cluster sampling.

(a) Argue that $s_W^2 = k^{-1} \sum_{i \in \mathbf{S}_c} \sigma_i^2$ is an unbiased estimator of σ_W^2.
(b) Argue that $s_B^2 = (k-1)^{-1} \sum_{i \in \mathbf{S}_c} \left(\mu_i - \hat{\mu}_y\right)^2$ is an unbiased estimator of σ_B^2, where $\hat{\mu}_y = k^{-1} \sum_{i \in \mathbf{S}_c} \mu_i$.
(c) Find an unbiased estimator for the overall population variance σ_y^2.
Hint: Show that $(KM-1)\sigma_y^2 = K(M-1)\sigma_W^2 + M(K-1)\sigma_B^2$.
(d) Argue that the overall sample variance $s_y^2 = (n-1)^{-1} \sum_{i \in \mathbf{S}_c} \sum_{j=1}^{M} (y_{ij} - \bar{y})^2$, where $\bar{y} = n^{-1} \sum_{i \in \mathbf{S}_c} \sum_{j=1}^{M} y_{ij}$, is usually not an unbiased estimator of σ_y^2.
Hint: Consider a decomposition of s_y^2 in terms of s_W^2 and s_B^2.

3.10 (Single-Stage Cluster Sampling) An accounting firm is interested in estimating the error rate in a compliance audit it is conducting. The population contains 828 claims, and each claim consists of 215 fields. The firm audits 85 claims selected from the 828 claims using SRSWOR, and all 215 fields are checked for errors for each selected claim. It is found that, among the 85 audited claims, one claim has errors in 4 of the 215 fields, one claim has three errors, 4 claims have two errors, 22 claims have one error, and the remaining 57 claims have no errors.

(a) Treating this as single-stage cluster sampling, estimate the error rate among all 828 × 215 = 178,020 fields. Compute the standard error for your estimate.
(b) Estimate (with standard error) the total number of errors in the 828 claims.

(c) Suppose that, instead of taking a cluster sample, the firm takes 85 × 215 = 18,275 fields from the 828 × 215 = 178,020 fields in the population using SRSWOR. If the estimated error rate for all fields from the SRSWOR sample is the same as in part (a), what will be the estimated variance? How does this compare with the estimated variance from part (a)?

3.11 (Proof of Part (c) of Theorem 3.4) Let $\sigma_T^2 = (K-1)^{-1}\sum_{i=1}^{K}\left(T_i - \mu_T\right)^2$ be the population variance of cluster totals. Let $\hat{\sigma}_T^2 = (k-1)^{-1}\sum_{i\in\mathbf{S}_c}\left(\hat{T}_i - \tilde{\mu}_T\right)^2$ be a "plug-in" moment estimator for σ_T^2, where $\hat{T}_i = M_i\bar{y}_i$ and $\tilde{\mu}_T = k^{-1}\sum_{i\in\mathbf{S}_c}\hat{T}_i$. Show that $E(\hat{\sigma}_T^2) = \sigma_T^2 + W$, where $W = K^{-1}\sum_{i=1}^{K} M_i^2(1 - m_i/M_i)\sigma_i^2/m_i$.
Hint: Re-write $\hat{\sigma}_T^2$ as

$$\hat{\sigma}_T^2 = \frac{k}{k-1}\left\{\frac{1}{k}\sum_{i\in\mathbf{S}_c} M_i^2(\bar{y}_i)^2\right\} - \frac{k}{k-1}\left(\tilde{\mu}_T\right)^2,$$

and use $E_2\{(\bar{y}_i)^2\} = \mu_i^2 + V_2(\bar{y}_i)$, $E\{(\tilde{\mu}_T)^2\} = \mu_T^2 + V(\tilde{\mu}_T)$, plus the result from Part (b) of Theorem 3.4 to complete the proof.

3.12 (Single-Stage and Two-Stage Cluster Sampling)

(a) Suppose that the population consists of $K = 40$ clusters. Let $T_i = \sum_{j=1}^{M_i} y_{ij}$ and $T_i^{(2)} = \sum_{j=1}^{M_i} y_{ij}^2$ for cluster i where M_i is the cluster size and y_{ij} is the value of the response variable for unit j in cluster i. A single-stage cluster sample of $k = 3$ clusters is selected by SRSWOR, and the data from the three sampled clusters are summarized as follows:

Cluster (i)	M_i	T_i	$T_i^{(2)}$
1	4	20	102
2	3	15	110
3	7	49	351

Compute an estimate of the population mean μ_y with an estimated variance, assuming $N = \sum_{i=1}^{K} M_i = 200$ is known.

(b) Compute an estimate of the population mean μ_y, assuming N is unknown.

(c) Suppose that for each of the three selected clusters, only two elements are chosen at the second stage, i.e., $m_i = 2$, by SRSWOR, and the y values are 3.0 and 7.0 for cluster 1, 4.0 and 8.0 for cluster 2, and 6.0 and 10.0 for cluster 3. Compute an estimate of μ_y with an estimated variance, assuming $N = 200$ is known.

(d) Continued from part (c): Compute an estimate of the population mean μ_y, assuming N is unknown.

Chapter 4
General Theory and Methods of Unequal Probability Sampling

This chapter covers two important aspects of modern survey sampling theory and methods: (i) A unified estimation theory using the first and the second order inclusion probabilities under an arbitrary probability sampling design; (ii) Sampling procedures for selecting units with unequal probabilities. The specific sampling procedures and estimation methods discussed in Chaps. 2 and 3 can be viewed as part of the general theory and methods under the unified framework.

4.1 Sample Inclusion Probabilities

A probability sampling design can be defined through one of the two general concepts: (i) A probability measure $\mathscr{P}$ over the set of all possible candidate samples, Ω; (ii) A sampling scheme specified through draw-by-draw procedures from the sampling frames. The first concept is more of theoretical interest while the second concept is what we use in practice to select a probability survey sample.

It will become clear in the next section that the most crucial part of a probability sampling design, specified through either (i) or (ii), are the sample inclusion probabilities. Let $\mathbf{S}$ be the survey sample. The first order inclusion probabilities are defined as

$$\pi_i = P\big(i \in \mathbf{S}\big)\,, \quad i = 1, 2, \cdots, N\,.$$

The second order inclusion probabilities are defined as

$$\pi_{ij} = P\big(i, j \in \mathbf{S}\big)\,, \quad i, j = 1, 2, \cdots, N \quad \text{and} \quad i \neq j\,.$$

The inclusion probabilities are defined for all units in the population. However, it is shown in the next section that constructing a point estimator of μ_y or T_y typically

C. Wu, M. E. Thompson, *Sampling Theory and Practice*, ICSA Book
Series in Statistics, https://doi.org/10.1007/978-3-030-44246-0_4

only requires π_i for $i \in \mathbf{S}$, and computing a variance estimator usually requires π_{ij} for $i, j \in \mathbf{S}$. It will also help to simplify certain expressions if we let

$$\pi_{ii} = P(i, i \in \mathbf{S}) = P(i \in \mathbf{S}) = \pi_i \, .$$

This amounts to putting the π_i's as the diagonal elements on the matrix of inclusion probabilities with the π_{ij}'s as off-diagonal elements. It is apparent that $\pi_{ij} = \pi_{ji}$ for any i and j.

The first and the second order inclusion probabilities are uniquely defined by the probability sampling design, but different sampling designs may have the same first and/or second order inclusion probabilities. For SRSWOR, we have $\pi_i = n/N$ for any i and $\pi_{ij} = n(n-1)/\{N(N-1)\}$ for any $i \neq j$. For stratified simple random sampling, we have $\pi_i = n_h/N_h$ if unit $i \in \mathbf{U}_h$, i.e., unit i belongs to stratum h. The second order inclusion probabilities can be expressed as

$$\pi_{ij} = \begin{cases} n_h(n_h-1)/\{N_h(N_h-1)\} & \text{if} \quad i, j \in \mathbf{U}_h \quad \text{and} \quad i \neq j \, , \\ n_h n_l/(N_h N_l) & \text{if} \quad i \in \mathbf{U}_h \, , \;\; j \in \mathbf{U}_l \quad \text{and} \quad h \neq l \, . \end{cases} \tag{4.1}$$

The inclusion probabilities for some other commonly used sampling designs are given in Problem 4.1.

For any probability sampling designs, the first and the second order inclusion probabilities satisfy the following equalities:

$$\sum_{i=1}^N \pi_i = E(n) \, , \qquad \sum_{i=1}^N \sum_{j=1}^N \pi_{ij} = E\big(n^2\big) \, .$$

For sampling designs with fixed sample sizes, the above equalities reduce to

$$\sum_{i=1}^N \pi_i = n \, , \qquad \sum_{i=1}^N \sum_{j=1}^N \pi_{ij} = n^2 \, . \tag{4.2}$$

The second equality can also be written as $\sum_{i \neq j, i=1}^N \sum_{j=1}^N \pi_{ij} = n(n-1)$. When the sample size n is fixed, we have a third equality given by

$$\sum_{j=1}^N \pi_{ij} = n\pi_i \, , \quad i = 1, 2, \cdots, N \, . \tag{4.3}$$

Equations (4.2) and (4.3) are useful tools for checking for computational errors in calculating the inclusion probabilities.

Let A_i be the sample indicator variables defined in Sect. 2.1, i.e., $A_i = 1$ if $i \in \mathbf{S}$ and $A_i = 0$ otherwise. We have $\pi_i = E(A_i)$ and $\pi_{ij} = E(A_i A_j)$. Note that we also have $\pi_{ii} = E(A_i A_i) = E(A_i) = \pi_i$. The three equalities on π_i and π_{ij} in (4.2) and (4.3) follow from the three parallel equalities on the A_i's:

$$\sum_{i=1}^{N} A_i = n \,, \qquad \sum_{j=1}^{N} A_i A_j = n A_i \,, \qquad \sum_{i=1}^{N}\sum_{j=1}^{N} A_i A_j = n^2 \,.$$

Equation (4.3) uses $E(nA_i) = n\pi_i$ when n is fixed. When n is random, $E(nA_i)$ does not reduce to $E(n)\pi_i$ since n and A_i are correlated to each other.

In some practical situations, it might be desirable to include certain "special units" into the sample with certainty. This is equivalent to setting $\pi_i = 1$ if i is always part of the sample. For nonrespondents (who never participate in any surveys) or other units which can never be included in the final sample, we have $\pi_i = 0$. This latter case needs to be treated with caution. See a formal result presented in Part (c) of Problem 4.3. Any units with $\pi_i = 0$ do not belong to the sampled population. Statistical inferences based on the observed survey data under such scenarios are not necessarily valid for the population that includes those units. Additional information or acceptable assumptions are required to extend the results.

4.2 The Horvitz-Thompson Estimator

Consider the estimation of the population total $T_y = \sum_{i=1}^{N} y_i$ under a general probability sampling design. Let $\{y_i, i \in \mathbf{S}\}$ be the survey sample data and let π_i and π_{ij} be the first and the second order inclusion probabilities under the design. We assume that $\pi_i > 0$ for every i. The Horvitz-Thompson estimator of T_y, from the well-known paper by Horvitz and Thompson (1952), is given by

$$\hat{T}_{y\text{HT}} = \sum_{i \in \mathbf{S}} \frac{y_i}{\pi_i} = \sum_{i \in \mathbf{S}} d_i y_i \,,$$

where $d_i = 1/\pi_i$ is called the *basic design weight* and the subscript "HT" indicates "Horvitz-Thompson". Using the sample indicator variables A_i, we can rewrite the HT estimator as $\hat{T}_{y\text{HT}} = \sum_{i=1}^{N} A_i d_i y_i$. Noting that $E(A_i) = \pi_i = d_i^{-1}$ for every i, we have the following basic result on the HT estimator.

Theorem 4.1 *For any probability sampling design satisfying $\pi_i > 0$ for all i, the Horvitz-Thompson estimator $\hat{T}_{y\text{HT}}$ is design-unbiased for the population total T_y.*

The Horvitz-Thompson estimator for the population total is one of the fundamental building blocks for analysis of complex survey data. It can be verified that many commonly used estimators of T_y under specific sampling designs are special cases of the Horvitz-Thompson estimator. Horvitz and Thompson's (1952) paper was published by the Journal of the American Statistical Association (JASA). In 1951, Narain had a paper published by the Journal of the Indian Society of Agricultural Statistics and proposed the same estimator under an arbitrary probability sampling design. The term "*Horvitz-Thompson estimator*" has been used by the large majority of survey statisticians, partly due to the influence of the JASA paper and many subsequent research papers on the topic. The contribution from Narain's (1951) paper, however, should be rightfully credited from a historical point of view. Rao (2005) advocated to use the term "the Narain-Horvitz-Thompson (NHT) estimator" rather than the HT estimator.

The basic design weights $d_i = 1/\pi_i$ play a crucial role for the general estimation theory. When the sum of the d_i values in the sample is equal to N, the value of d_i can be interpreted as

> the number of units in the survey population which are represented by unit i selected for the survey sample.

In such a case the HT estimator $\hat{T}_{y\text{HT}} = \sum_{i\in\mathbf{S}} d_i y_i$ is also called an *expansion estimator*, with the weight d_i being treated as the expansion factor from unit i in the sample to the group of units in the population which are represented by the unit. Expansion estimators are linear estimators in terms of their structure.

Consider the following general class of linear estimators

$$\mathscr{G} = \Big\{\hat{T}_{yc} \mid \hat{T}_{yc} = \sum_{i\in\mathbf{S}} c_i(\mathbf{S})y_i\,, \quad c_i(\mathbf{S}) \;\text{ depends on }\; i \;\text{ and }\; \mathbf{S}\Big\}\,.$$

This is the so-called *Godambe class of linear estimators*. A fundamental (negative) result in design-based survey sampling theory is that minimum variance unbiased linear estimator does not exist in the general class $\mathscr{G}$ (Godambe 1955). Consider the following subclass of $\mathscr{G}$:

$$\mathscr{G}_1 = \Big\{\hat{T}_{yc} \mid \hat{T}_{yc} = \sum_{i\in\mathbf{S}} c_i y_i\,, \quad c_i \;\text{ is pre-specified and independent of }\; \mathbf{S}\Big\}\,.$$

It can be shown that the HT estimator is the only design-unbiased estimator among this subclass of linear estimators (Problem 4.3).

Theorem 4.2 *Let $\hat{T}_{y\text{HT}} = \sum_{i\in\mathbf{S}} y_i/\pi_i$ be the Horvitz-Thompson estimator of the population total T_y where the survey design satisfies $\pi_i > 0$ and $\pi_{ij} > 0$ for all i and j.*

(a) The theoretical variance of $\hat{T}_{y\text{HT}}$ is given by

$$V\big(\hat{T}_{y\text{HT}}\big) = \sum_{i=1}^{N}\sum_{j=1}^{N}\big(\pi_{ij} - \pi_i\pi_j\big)\frac{y_i}{\pi_i}\frac{y_j}{\pi_j}\,. \tag{4.4}$$

(continued)

Theorem 4.2 (continued)
(b) An unbiased variance estimator for $\hat{T}_{\mathrm{yHT}}$ is given by

$$v\big(\hat{T}_{\mathrm{yHT}}\big) = \sum_{i\in\mathbf{S}}\sum_{j\in\mathbf{S}} \frac{\pi_{ij}-\pi_i\pi_j}{\pi_{ij}}\,\frac{y_i}{\pi_i}\,\frac{y_j}{\pi_j}\,. \tag{4.5}$$

Proof

(a) Using the sample indicator variables A_i, we have

$$V\big(\hat{T}_{\mathrm{yHT}}\big) = V\Big(\sum_{i=1}^{N} A_i\,\frac{y_i}{\pi_i}\Big) = \sum_{i=1}^{N}\sum_{j=1}^{N} Cov\big(A_i, A_j\big)\,\frac{y_i}{\pi_i}\,\frac{y_j}{\pi_j}\,.$$

The expression given by (4.4) follows from $Cov(A_i, A_j) = \pi_{ij} - \pi_i\pi_j$. It should be noted that $\pi_{ii} = \pi_i$ and the terms with $i = j$ correspond to the individual variance terms, i.e.,

$$\big(\pi_{ii}-\pi_i\pi_i\big)\,\frac{y_i}{\pi_i}\,\frac{y_i}{\pi_i} = \pi_i\big(1-\pi_i\big)\Big(\frac{y_i}{\pi_i}\Big)^2 = V\Big(A_i\,\frac{y_i}{\pi_i}\Big)\,.$$

(b) To show that $v\big(\hat{T}_{\mathrm{yHT}}\big)$ is an unbiased variance estimator, we re-write the expression from (4.5) as

$$v\big(\hat{T}_{\mathrm{yHT}}\big) = \sum_{i=1}^{N}\sum_{j=1}^{N} A_iA_j\,\frac{\pi_{ij}-\pi_i\pi_j}{\pi_{ij}}\,\frac{y_i}{\pi_i}\,\frac{y_j}{\pi_j}\,.$$

It follows from $E(A_iA_j) = \pi_{ij}$ that $E\big\{v(\hat{T}_{\mathrm{yHT}})\big\} = V\big(\hat{T}_{\mathrm{yHT}}\big)$.

□

It is sometimes of interest to separate the terms with $i = j$ from (4.4) and (4.5). The alternative expressions for the theoretical variance and the variance estimator are given by

$$V\big(\hat{T}_{\mathrm{yHT}}\big) = \sum_{i=1}^{N}\frac{1-\pi_i}{\pi_i}y_i^2 + 2\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\big(\pi_{ij}-\pi_i\pi_j\big)\,\frac{y_i}{\pi_i}\,\frac{y_j}{\pi_j}\,,$$

$$v\big(\hat{T}_{\mathrm{yHT}}\big) = \sum_{i\in\mathbf{S}}\frac{1-\pi_i}{\pi_i^2}y_i^2 + 2\sum_{i\in\mathbf{S}}\sum_{j>i,j\in\mathbf{S}}\frac{\pi_{ij}-\pi_i\pi_j}{\pi_{ij}}\,\frac{y_i}{\pi_i}\,\frac{y_j}{\pi_j}\,.$$

In practice, the most challenging part of variance estimation for the HT estimator is to compute the second order inclusion probabilities π_{ij} for $i, j \in \mathbf{S}$. For sampling designs with unequal selection probabilities, this can be a daunting task. See Sects. 4.4–4.6 for further discussion.

4.2.1 The Yates-Grundy-Sen Variance Formula for the HT Estimator

For sampling designs with fixed sample sizes, a practically useful alternative expression for the theoretical variance of the HT estimator and the related variance estimator, due to Yates and Grundy (1953) and Sen (1953), are given below.

Theorem 4.3 *Under sampling designs with fixed sample sizes, we have*

$$V\big(\hat{T}_{y\mathrm{HT}}\big) = \sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\big(\pi_i\pi_j - \pi_{ij}\big)\Big(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\Big)^2 \tag{4.6}$$

$$= \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\big(\pi_i\pi_j - \pi_{ij}\big)\Big(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\Big)^2 . \tag{4.7}$$

An unbiased variance estimator is given by

$$v\big(\hat{T}_{y\mathrm{HT}}\big) = \sum_{i\in\mathbf{S}}\sum_{j>i, j\in\mathbf{S}}\frac{\pi_i\pi_j - \pi_{ij}}{\pi_{ij}}\Big(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\Big)^2 \tag{4.8}$$

$$= \frac{1}{2}\sum_{i\in\mathbf{S}}\sum_{j\in\mathbf{S}}\frac{\pi_i\pi_j - \pi_{ij}}{\pi_{ij}}\Big(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\Big)^2 . \tag{4.9}$$

Proof We show that the second expression (4.7) for $V\big(\hat{T}_{y\mathrm{HT}}\big)$ holds. The first expression (4.6) is obviously equivalent to the second one. The design unbiasedness of the variance estimators (4.8) and (4.9) follows from the same argument as for (4.5).

When the sampling design has a fixed sample size, we have

$$\sum_{j=1}^{N}\big(\pi_i\pi_j - \pi_{ij}\big) = \pi_i\sum_{j=1}^{N}\pi_j - \sum_{j=1}^{N}\pi_{ij} = n\pi_i - n\pi_i = 0 .$$

This leads to

$$\sum_{i=1}^{N}\sum_{j=1}^{N}\left(\pi_i\pi_j-\pi_{ij}\right)\left(\frac{y_i}{\pi_i}\right)^2=\sum_{j=1}^{N}\sum_{i=1}^{N}\left(\pi_i\pi_j-\pi_{ij}\right)\left(\frac{y_j}{\pi_j}\right)^2=0\,.$$

Consequently,

$$\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left(\pi_i\pi_j-\pi_{ij}\right)\left(\frac{y_i}{\pi_i}-\frac{y_j}{\pi_j}\right)^2=\sum_{i=1}^{N}\sum_{j=1}^{N}\left(\pi_{ij}-\pi_i\pi_j\right)\frac{y_i}{\pi_i}\frac{y_j}{\pi_j}\,,$$

which is the expression for $V\left(\hat{T}_{y\mathrm{HT}}\right)$ given by (4.4). □

It should be noted that the expressions from Theorem 4.3 involve terms $\pi_i\pi_j-\pi_{ij}$ while the expressions from Theorem 4.2 involve terms $-(\pi_i\pi_j-\pi_{ij})=\pi_{ij}-\pi_i\pi_j$. For $i=j$, $\pi_i\pi_j-\pi_{ij}=-\pi_i(1-\pi_i)\le 0$, but none of the expressions in (4.6) - (4.9) requires terms with $i=j$ since $(y_i/\pi_i-y_j/\pi_j)^2=0$ for $i=j$.

One of the implications from the expressions (4.8) or (4.9) for the variance estimator is that the estimated variance is guaranteed to be nonnegative if the sampling design satisfies $\pi_i\pi_j-\pi_{ij}\ge 0$, i.e., $Cov(A_i,A_j)\le 0$ for all $i\ne j$. Many of the commonly used sampling designs do satisfy this property. The Yates-Grundy-Sen variance estimator (4.8) or (4.9) is less sensitive to rounding errors and therefore is computationally more stable than the general variance estimator given in (4.5).

4.2.2 *The Hájek Variance Estimator*

The variance formulas presented in Sect. 4.2.1 require the second order inclusion probabilities π_{ij}. It can be seen from the unequal probability sampling procedures described in Sect. 4.4 that computing the second order inclusion probabilities can be a very difficult task for certain survey designs.

Hájek (1964) presented a variance estimator for the Horvitz-Thompson estimator which does not involve the second order inclusion probabilities. The Hájek variance estimator is given by

$$v_{\mathrm{H}}\left(\hat{T}_{y\mathrm{HT}}\right)=\sum_{i\in\mathbf{S}}c_i\left(\frac{y_i}{\pi_i}-\hat{B}\right)^2\,,$$

where

$$\hat{B}=\frac{\sum_{i\in\mathbf{S}}c_iy_i/\pi_i}{\sum_{i\in\mathbf{S}}c_i}\quad\text{and}\quad c_i=\frac{n}{n-1}(1-\pi_i)\,.$$

The variance estimator is derived based on an approximation to the π_{ij} using the first order inclusion probabilities π_i and π_j and is shown to perform well for high entropy sampling designs (Haziza et al. 2008), including the conditional Poisson sampling method, the Rao-Sampford sampling method and the Rao-Hartley-Cochran method discussed in Sects. 4.4 and 4.6.

4.2.3 Estimation of the Population Mean μ_y and the Hájek Estimator

When the population size N is known, the Horvitz-Thompson estimator for the population mean μ_y is given by

$$\hat{\mu}_{y\mathrm{HT}} = \frac{1}{N}\sum_{i\in\mathbf{S}}\frac{y_i}{\pi_i} = \frac{1}{N}\hat{T}_{y\mathrm{HT}} .$$

It is a design-unbiased estimator for μ_y with theoretical variance and variance estimator given respectively by

$$V\big(\hat{\mu}_{y\mathrm{HT}}\big) = \frac{1}{N^2}V\big(\hat{T}_{y\mathrm{HT}}\big) \quad \text{and} \quad v\big(\hat{\mu}_{y\mathrm{HT}}\big) = \frac{1}{N^2}v\big(\hat{T}_{y\mathrm{HT}}\big) .$$

When the population size N is unknown, which is often the case for two-stage or multi-stage cluster sampling, an exactly design-unbiased estimator of μ_y might not be available. Let

$$\hat{N} = \sum_{i\in\mathbf{S}}\frac{1}{\pi_i} = \sum_{i\in\mathbf{S}} d_i .$$

The $\hat{N}$ can be viewed as the HT estimator $\hat{T}_{z\mathrm{HT}} = \sum_{i\in\mathbf{S}} d_i z_i$ for the population total $T_z = \sum_{i=1}^N z_i$ where $z_i = 1$ for every i. In other words, the $\hat{N}$ is a design-unbiased estimator for $T_z = N$. The population mean μ_y can be estimated by

$$\hat{\mu}_{y\mathrm{H}} = \frac{1}{\hat{N}}\sum_{i\in\mathbf{S}}\frac{y_i}{\pi_i} = \frac{\sum_{i\in\mathbf{S}} d_i y_i}{\sum_{i\in\mathbf{S}} d_i} .$$

This is the so-called Hájek estimator of the population mean (Hájek 1971), as indicated by the "H" in the subscript. The Hájek estimator is a ratio of two HT estimators. Its design-based properties will be discussed in Sect. 5.3 after the introduction to the generalized ratio estimator.

Although it is not exactly unbiased in general, the Hájek estimator for the population mean μ_y has the following attractive feature: When $y_i = c$ for every i, we have $\mu_y = c$. The Hájek estimator also gives $\hat{\mu}_{y\text{H}} = c$ for any sampling designs. This makes the Hájek estimator as a practically useful alternative method even if N is available, as demonstrated below for the estimation of distribution functions and quantiles.

4.2.4 Estimation of the Population Distribution Function and Quantiles

The finite population distribution function for the study variable y is defined as

$$F_y(t) = \frac{1}{N}\sum_{i=1}^{N} I(y_i \le t)\,, \qquad t \in (-\infty\,,\ \infty)\,,$$

where $I(\cdot)$ is an indicator function, so that $I(y_i \le t) = 1$ if $y_i \le t$ and $I(y_i \le t) = 0$ if $y_i > t$. Note that $F_y(t)$ is a step-function with jumps at distinct values of the y_i's.

For a given t, the value of $F_y(t)$ is the proportion of units in the population with y-values less than or equal to t. Suppose that the unit is "family" and the y-value is the "annual family income" (in \$, from all sources). Suppose that a family is categorized as a "*Low Income Family*" if the annual family income is less than or equal to $t_0 = \$30{,}000$. In this case $F_y(t_0)$ is the proportion of low-income families in the survey population.

Let $\alpha \in (0,\ 1)$. The 100αth population quantile t_α for the y-variable is defined through the inversion of $F_y(t)$. Due to the discontinuity of the finite population distribution function, the exact definition of t_α is given by

$$t_\alpha = F_y^{-1}(\alpha) = \inf\left\{t \mid F_y(t) \ge \alpha\right\}.$$

Note that this implies $F_y(t_\alpha) \ge \alpha$. The 100αth population quantile t_α is the cut-off of the y-values for units belonging to the lower 100α percent of the population. Suppose that the government has an overall budget to provide a benefit package to 10% of the families in the country, and the eligibility is based on family income in the previous year. The tenth population quantile $t_{0.10}$ for the family income variable y would be a natural cut-off to decide eligible families for the program, and the cut-off would sometimes be conservative.

The population distribution function $F_y(t)$ for a fixed t is the population mean for the indicator variable $I(y_i \le t)$. The Hájek estimator of $F_y(t)$ is given by

$$\hat{F}_{y\text{H}}(t) = \frac{\sum_{i\in\mathbf{S}} d_i I(y_i \le t)}{\sum_{i\in\mathbf{S}} d_i} = \sum_{i\in\mathbf{S}} w_i\, I(y_i \le t)\,,$$

where $w_i = d_i / \sum_{k \in \mathbf{S}} d_k$ which depends on the sample $\mathbf{S}$. It is apparent that $w_i > 0$ and $\sum_{i \in \mathbf{S}} w_i = 1$. The Hájek estimator $\hat{F}_{y\mathrm{H}}(t)$ itself is a true distribution function: it is nondecreasing and satisfies $\hat{F}_{y\mathrm{H}}(-\infty) = 0$ and $\hat{F}_{y\mathrm{H}}(\infty) = 1$. This is not the case if we use the HT estimator for $F_y(t)$ where $\hat{N} = \sum_{i \in \mathbf{S}} d_i$ is replaced by N.

The fact that $\hat{F}_{y\mathrm{H}}(t)$ is a true distribution function is important for the estimation of population quantiles. The Hájek estimator of t_α for a given α is defined as

$$\hat{t}_\alpha = \hat{F}_{y\mathrm{H}}^{-1}(\alpha) = \inf\left\{ t \mid \hat{F}_{y\mathrm{H}}(t) \geq \alpha \right\}.$$

Let n be the sample size and let $y_{(1)} \leq y_{(2)} \leq \cdots \leq y_{(n)}$ be the ordered values of $\{y_i, i \in \mathbf{S}\}$. Let $w_{[i]}$ be the value of w_i associated with $y_{(i)}$, i.e.,

$$\left(w_{[1]}, y_{(1)}\right), \left(w_{[2]}, y_{(2)}\right), \cdots, \left(w_{[n]}, y_{(n)}\right)$$

are the ordered pairs from $(w_1, y_1), (w_2, y_2), \cdots, (w_n, y_n)$ based on the ordered values of the y_i's. It can be seen that $\hat{t}_\alpha = y_{(1)}$ if $\alpha < w_{[1]}$. For any $\alpha \geq w_{[1]}$, we have

$$\hat{t}_\alpha = y_{(j)}\,, \quad \text{where} \quad \sum_{i=1}^{j} w_{[i]} \geq \alpha \quad \text{and} \quad \sum_{i=1}^{j-1} w_{[i)]} < \alpha\,.$$

The design-based properties of $\hat{F}_{y\mathrm{H}}(t)$ and $\hat{t}_\alpha$ will be discussed in Sect. 5.3 after introduction of the generalized ratio estimators.

4.3 PPS Sampling and the HT Estimator: An Optimal Strategy

The Horvitz-Thompson estimator for the population total T_y is design-unbiased under an arbitrary probability sampling design. One of the related research topics is to identify sampling designs which together with the HT estimator provide more efficient estimation for T_y.

4.3.1 PPS Sampling Designs

Suppose that $y_i > 0$ for all i. Consider a hypothetical situation where we can select the sample $\mathbf{S}$ with the first order inclusion probabilities proportional to the values of the y variable, i.e., $\pi_i \propto y_i$, $i = 1, 2, \cdots, N$. Let n be the planned sample size. The constraint $\sum_{i=1}^{N} \pi_i = n$ would lead to

$$\pi_i = n\frac{y_i}{T_y}\,, \quad i = 1, 2, \cdots, N\,.$$

The HT estimator of T_y under this hypothetical sampling design will satisfy

$$\hat{T}_{y\text{HT}} = \sum_{i\in\mathbf{S}} \frac{y_i}{\pi_i} = T_y \sum_{i\in\mathbf{S}} \frac{y_i}{ny_i} = T_y\,.$$

In other words, the HT estimator reduces to the exact value of the population total with any survey sample selected under the design.

In practice, the hypothetical sampling design can never be used, since the values of the y_i's are required for the sampling design before the selection of the sample while the objective of the survey is to measure the y_i after sample units are selected. Nevertheless, the unrealistic feature $\pi_i \propto y_i$ points in the direction of obtaining potentially more efficient sampling designs, which select "large units" (in terms of the y values) with bigger probabilities.

Let z be an auxiliary variable which provides a measure for the "size" of sampling units. For instance, in agricultural surveys, the study variable y may be the yield of a farm product, and the z variable the acreage of the farm. In business surveys, the z variable could simply be the number of workers of the company. For family expenditure surveys, the z variable could be the family income of the previous year. For two-stage or multi-stage cluster sampling, the z variable could be the actual size of the clusters. Our discussions in the remaining part of the chapter focus on single-stage sampling methods but the methodological development can be used for more general sampling designs. We assume that

- The values $z_1, z_2, \cdots, z_N$ are available at the survey design stage;
- The value z_i provides a measure of the size for unit i, and $z_i > 0$ for all i;
- The size variable z and the study variable y are positively correlated.

We consider sampling designs for which the first order inclusion probabilities are proportional to the size variable, i.e., $\pi_i \propto z_i$, $i = 1, 2, \cdots, N$. These are termed as PPS (*Probability Proportional to Size*) sampling designs. Let n be the expected sample size. The constraint $\sum_{i=1}^N \pi_i = n$ leads to

$$\pi_i = n\frac{z_i}{T_z}\,, \quad i = 1, 2, \cdots, N\,.$$

If we assume the size variable z is re-scaled such that $\sum_{i=1}^N z_i = 1$, then $\pi_i = nz_i$. Since $\pi_i \le 1$ for all i, we require that the re-scaled size variable z satisfies $z_i \le 1/n$ for all i.

Finding sampling designs and sampling procedures with first order inclusion probabilities proportional to a size measure has been an active research topic ever since Horvitz and Thompson's (1952) JASA paper. There are a number of PPS sampling methods which can be used in practice for general purposes. Three of those methods are presented in Sect. 4.4. Two alternative methods are presented in Sects. 4.5 and 4.6.

4.3.2 An Optimal Strategy

As noted in Sect. 2.1, a sampling design paired with an estimation method is called a *sampling strategy*. We show that a PPS sampling design coupled with the HT estimator is an optimal sampling strategy under the following framework.

Let $\{(y_i, z_i), i = 1, 2, \cdots, N\}$ be the values of (y, z) for the N units in the survey population, and assume that $z_i > 0$ for all i. The two variables y and z are assumed to follow a *superpopulation model*, denoted as ξ, and the finite population values (y_i, z_i), $i = 1, 2, \cdots, N$, are viewed as a particular "sample" from the superpopulation model ξ. More specifically, we assume that

$$y_i = \beta z_i + z_i \epsilon_i \,, \quad i = 1, 2, \cdots, N \,, \tag{4.10}$$

where ϵ_i, $i = 1, 2, \cdots, N$ are independent and identically distributed error terms with $E_\xi(\epsilon_i) = 0$ and $V_\xi(\epsilon_i) = \tau^2$. The E_ξ and V_ξ refer to expectation and variance under the superpopulation model, ξ. The parameters β and $\tau^2 > 0$ are referred to as *superpopulation parameters*. Under the model (4.10), we have $E_\xi\big(y_i \mid z_i\big) = \beta z_i$ and $V_\xi\big(y_i \mid z_i\big) = z_i^2 \tau^2$.

The concept of superpopulation models was first introduced in the landmark paper by Cochran (1939). The use of models in survey sampling will be further discussed in Chap. 5 as well as Chap. 7 in Part II of the book. Under the model (4.10), the values of the y variable should be viewed as random variables with the first two moments specified by (4.10). The design-based variance of the HT estimator is denoted as $V_p\big(\hat{T}_{y\text{HT}}\big)$, where the subscript "$p$" indicates "*probability sampling design*". The general expression of $V_p\big(\hat{T}_{y\text{HT}}\big)$ is given by (4.4) for any sampling design and is also given by (4.7) when the sampling design has a fixed sample size. While $V_p\big(\hat{T}_{y\text{HT}}\big)$ has a fixed value for the survey population at hand, it should also be viewed as a random quantity under model (4.10) due to its dependence on $(y_1, y_2, \cdots, y_N)$.

The expected value of the design-based variance of the HT estimator under the assumed superpopulation model is given by $E_\xi\big\{V_p(\hat{T}_{y\text{HT}})\big\}$. This is also called the *average variance* or the *expected variance* by Rao (1966a,b) and the *anticipated variance* by Isaki and Fuller (1982). Due to Godambe's result (1955) on the non-existence of best linear unbiased estimator under the design-based framework for finite populations, optimality criteria involving *minimum anticipated variance* under the joint model-design framework have been considered by many survey researchers. See, for instance, Godambe (1955), Godambe and Thompson (1973), Cassel et al. (1976), Isaki and Fuller (1982) and Wu (2003), among many others.

We assume that the inclusion probabilities π_i are independent of the y variable. The following result was originally due to Godambe (1955) but is presented here under the explicitly stated model (4.10).

Theorem 4.4 *The sampling strategy of a PPS sampling design paired with the HT estimator is optimal in that the anticipated variance $E_\xi\{V_p(\hat{T}_{y\text{HT}})\}$ under model (4.10) is minimized under any fixed sample size PPS sampling designs with $\pi_i \propto z_i$.*

Proof Under the model (4.10) and for $i \neq j$, we have

$$E_\xi\Big\{\Big(\frac{y_i}{\pi_i}-\frac{y_j}{\pi_j}\Big)^2\Big\} = \Big\{E_\xi\Big(\frac{y_i}{\pi_i}-\frac{y_j}{\pi_j}\Big)\Big\}^2 + V_\xi\Big(\frac{y_i}{\pi_i}-\frac{y_j}{\pi_j}\Big)$$
$$= \beta^2\Big(\frac{z_i}{\pi_i}-\frac{z_j}{\pi_j}\Big)^2 + \tau^2\Big(\frac{z_i^2}{\pi_i^2}+\frac{z_j^2}{\pi_j^2}\Big).$$

Using the same arguments as in the proof of Theorem 4.3, we can show that

$$\sum_{i=1}^N \sum_{j\neq i, j=1}^N (\pi_i\pi_j-\pi_{ij})\frac{z_i^2}{\pi_i^2} = \sum_{i=1}^N \sum_{j\neq i, j=1}^N (\pi_i\pi_j-\pi_{ij})\frac{z_j^2}{\pi_j^2} = \sum_{i=1}^N \pi_i(1-\pi_i)\frac{z_i^2}{\pi_i^2},$$

which holds for sampling designs with fixed sample sizes. Under the model (4.10) and by the Yates-Grundy-Sen variance formula (4.7), we have

$$E_\xi\{V_p(\hat{T}_{y\text{HT}})\} = \frac{1}{2}\sum_{i=1}^N \sum_{j\neq i, j=1}^N (\pi_i\pi_j-\pi_{ij})E_\xi\Big\{\Big(\frac{y_i}{\pi_i}-\frac{y_j}{\pi_j}\Big)^2\Big\}$$
$$= \frac{\beta^2}{2}\sum_{i=1}^N \sum_{j\neq i, j=1}^N (\pi_i\pi_j-\pi_{ij})\Big(\frac{z_i}{\pi_i}-\frac{z_j}{\pi_j}\Big)^2 + \tau^2\sum_{i=1}^N \pi_i(1-\pi_i)\frac{z_i^2}{\pi_i^2}$$
$$= \beta^2 D_1 + \tau^2 D_2 - \tau^2 D_3,$$

where

$$D_1 = \frac{1}{2}\sum_{i=1}^N \sum_{j\neq i, j=1}^N (\pi_i\pi_j-\pi_{ij})\Big(\frac{z_i}{\pi_i}-\frac{z_j}{\pi_j}\Big)^2, \quad D_2 = \sum_{i=1}\frac{z_i^2}{\pi_i} \quad \text{and} \quad D_3 = \sum_{i=1}^N z_i^2.$$

We first note that D_3 is independent of the choice of probability sampling designs. We also note that D_1 is the design-based variance for $\hat{T}_{z\text{HT}} = \sum_{i\in \mathbf{S}} z_i/\pi_i$ and hence $D_1 \geq 0$. It is straightforward to show that D_2 is minimized with respect to the π_i's subject to $\sum_{i=1}^{N} \pi_i = n$ when the sampling design satisfies $\pi_i \propto z_i$. Such a design also leads to $D_1 = 0$ which is the minimum value of D_1. □

One of the implications of Theorem 4.4 is as follows. For commonly encountered practical situations where $y_i \geq 0$ for all i, the model (4.10) implies that $\beta > 0$. The optimal sampling strategy, which combines a PPS sampling design with the use of the HT estimator, requires that the size variable z and the study variable y are positively correlated. When this rule is broken, the resulting unequal probability sampling strategy can be surprisingly inefficient. Rao (1966a) compared the HT estimator with the unweighted estimator and showed that the unweighted estimator has smaller average variance under the superpopulation model $y_i = \beta_0 + \beta_1 z_i + \varepsilon$ when the response variable y is unrelated to the size variable z (i.e., $\beta_1 = 0$). Scott and Smith (1969) considered the estimator with minimum anticipated mean squared error under the same model among the class of model-design unbiased estimators. See also the well-known Basu's elephant example below.

4.3.3 Basu's Elephant Example

At the Symposium on Foundations of Statistical Inference organized by the Department of Statistics, University of Waterloo, March 31 to April 9, 1970, Basu presented the following example as part of his talk. The following text is taken directly without any alteration from Basu's paper published in the proceedings of the symposium (Basu 1971).

Example 4.1 The circus owner is planning to ship his 50 elephants and so he needs a rough estimate of the total weight of the elephants. As weighing an elephant is a cumbersome process, the owner wants to estimate the total weight by weighing just one elephant. Which elephant should he weigh? So the owner looks back on his records and discovers a list of the elephants' weights taken 3 years ago. He finds that 3 years ago Sambo the middle-sized elephant was the average (in weight) elephant in the herd. He checks with the elephant trainer who reassures him (the owner) that Sambo may still be considered to be the average elephant in the herd. Therefore, the owner plans to weigh Sambo and take $50y$ (where y is the present weight of Sambo) as an estimate of the total weight $Y = Y_1 + Y_2 + \ldots + Y_{50}$ of the 50 elephants. But the circus statistician is horrified when he learns of the owner's purposive sampling plan. "How can you get an unbiased estimate of Y this way?" protests the statistician. So, together they work out a compromise sampling plan. With the help of a table of random numbers they devise a plan that allots a selection probability of $99/100$ to Sambo, and equal selection probabilities of $1/4900$ to each of the other 49 elephants. Naturally, Sambo is selected and the owner is happy. "How are you going to estimate Y?", asks the statistician. "Why? The estimate ought to be $50y$ of course," says the owner. "Oh! No! That cannot possibly be right," says the statistician, "I recently read an article in the *Annals* *of Mathematical Statistics* where it is proved that the Horvitz-Thompson estimator is the unique hyperadmissible estimator in the class of all generalized polynomial unbiased estimators." "What is the Horvitz-Thompson estimate in this case?" asks the owner, duly impressed. "Since the selection probability of Sambo in our plan was $99/100$," says the statistician, "the proper estimate of Y is $100y/99$ and not $50y$." "And, how would you have estimated Y," inquires the incredulous owner, "if our sampling plan made us select, say, the big elephant Jumbo?" "According to what I understand of the Horvitz-Thompson estimation method," says the unhappy statistician, "the proper estimate of Y would then have been $4900y$, where y is Jumbo's weight." That is how the statistician lost his circus job (and perhaps became a teacher of statistics!). ◇

As a historical note, one of the discussants for Basu's paper presented at the 1970 Waterloo Symposium was Jaroslav Hájek. The term *Hájek estimator* was coined following Hájek's discussion at the symposium (Hájek 1971).

4.4 PPS Sampling Procedures

Let $p_i > 0$ for all i and $\sum_{i=1}^{N} p_i = 1$. It is a simple task to select one unit from $\{1, 2, \cdots, N\}$ with the discrete probability measure $(p_1, p_2, \cdots, p_N)$. Here is one commonly used procedure:

- Let $b_0 = 0$; let $b_j = \sum_{k=1}^{j} p_k$ for $j = 1, 2, \cdots, N$, with $b_N = 1$;
- Generate r from $U[0,\ 1]$, the uniform distribution over $[0,\ 1]$;
- Select unit i if $b_{i-1} < r \le b_i$.

This procedure is called *the cumulative sum method*. It can be shown that $P(i$ is selected$) = p_i$, $i = 1, 2, \cdots, N$. In other words, selecting a sample of $n = 1$ can easily be done for any pre-specified inclusion probabilities.

It turns out that selecting a survey sample $\mathbf{S}$ of sample size $n \ge 2$ with inclusion probabilities π_i proportional to the pre-specified size measures z_i is a much harder task. This can be seen from the simplest case of selecting a PPS sample with $n = 2$ described below. For the rest of this section, we assume that $z_i > 0$ is the re-scaled size variable such that $\sum_{i=1}^{N} z_i = 1$.

4.4.1 PPS Sampling with n = 2

Suppose that $z_i < 1/2$ for all i and the goal is to select a sample $\mathbf{S}$ consisting of two units (i.e. $n = 2$) such that $\pi_i = P(i \in \mathbf{S}) = 2z_i$, $i = 1, 2, \cdots, N$.

A typical sampling procedure for selecting a sample of $n = 2$ units from the list of N units involves two draw-by-draw steps:

1. Select the first unit from the list of N units on the sampling frame based on a set of specified probabilities p_i: $p_i \geq 0$ and $\sum_{i=1}^N p_i = 1$;
2. Select the second unit from the list of $N - 1$ remaining units on the sampling frame based on another set of specified probabilities $q_{i|j}$: $q_{i|j} \geq 0$ and $\sum_{i \neq j, i=1}^N q_{i|j} = 1$, given that unit j is the unit selected in the first step.

The first and second order inclusion probabilities $\pi_i = P(i \in \mathbf{S})$ and $\pi_{ij} = P(i, j \in \mathbf{S})$ under this general sampling procedure are computed as follows:

$$
\begin{aligned}
\pi_i &= P(i \text{ is selected 1st}) + P(i \text{ is selected 2nd}) \\
&= p_i + \sum_{j \neq i, j=1}^{N} p_j\, q_{i|j}\,; \\
\pi_{ij} &= P(i \text{ selected 1st, } j \text{ selected 2nd}) + P(j \text{ selected 1st, } i \text{ selected 2nd}) \\
&= p_i\, q_{j|i} + p_j\, q_{i|j}\,.
\end{aligned}
$$

It becomes clear but a bit surprising that, if we select the two units with probabilities proportional to z_i at each step, i.e., $p_i = z_i$ and $q_{i|j} = z_i/(1 - z_j)$, the final inclusion probabilities π_i are not equal to $2z_i$ unless all z_i are equal. In the latter case $\pi_i = 2/N$ and the procedure is SRSWOR with $n = 2$. In general, sampling procedures with unequal selection probabilities at different steps create an asymmetric picture, where units selected at later steps depend on units selected at earlier steps. It requires some fine mathematical tricks to achieve the ultimate goal of having $\pi_i = nz_i$ for the final sample even for the simplest case of $n = 2$.

Example 4.2 Brewer's PPS Sampling Procedure for $n = 2$ (Brewer 1963a):

1. Draw the first unit from $\{1, 2, \cdots, N\}$ with probabilities $p_i \propto z_i(1 - z_i)/(1 - 2z_i)$, i.e.,
$$p_i = \frac{1}{D}\,\frac{z_i(1 - z_i)}{1 - 2z_i}\,, \quad i = 1, 2, \cdots, N\,,$$
where D is the normalization constant such that $\sum_{i=1}^N p_i = 1$.
2. Draw the second unit from the remaining $N - 1$ units with probabilities $q_{i|j} = z_i/(1 - z_j)$, $i = 1, 2, \cdots, N$ and $i \neq j$, given that j is selected on the first draw.

The normalization constant is given by

$$D = \sum_{i=1}^{N} \frac{z_i(1 - z_i)}{1 - 2z_i} = \frac{1}{2} \sum_{i=1}^{N} \frac{z_i\{1 + (1 - 2z_i)\}}{1 - 2z_i} = \frac{1}{2}\Big(1 + \sum_{i=1}^{N} \frac{z_i}{1 - 2z_i}\Big)\,.$$

The first order inclusion probabilities are given by

$$\begin{aligned}\pi_i &= \frac{z_i(1-z_i)}{D(1-2z_i)} + \sum_{j=1, j\neq i}^{N} \frac{z_j(1-z_j)}{D(1-2z_j)} \cdot \frac{z_i}{1-z_j} \\ &= \frac{z_i}{D}\Big[1 + \frac{z_i}{1-2z_i} + \sum_{j=1, j\neq i}^{N} \frac{z_j}{1-2z_j}\Big] \\ &= 2z_i\,.\end{aligned}$$

The second order inclusion probabilities are given by

$$\pi_{ij} = \frac{z_i(1-z_i)}{D(1-2z_i)} \cdot \frac{z_j}{1-z_i} + \frac{z_j(1-z_j)}{D(1-2z_j)} \cdot \frac{z_i}{1-z_j} = \frac{2z_iz_j}{D} \cdot \frac{1-z_i-z_j}{(1-2z_i)(1-2z_j)}\,.$$

Brewer's sampling procedure satisfies $\pi_i\pi_j - \pi_{ij} > 0$ for all $i \neq j$. This property was first pointed out by Rao (1965). See Problem 4.5 for further details. ◇

There exist several PPS sampling procedures for $n = 2$. Brewer and Hanif (1983) provided a comprehensive review of PPS sampling methods, including methods for $n = 2$. In practice, those methods are often used in stratified cluster sampling where the number of strata H is large and only two clusters are selected within each stratum.

The mathematical details presented in Example 4.2 for Brewer's PPS sampling method for $n = 2$ reveal the difficulties of draw-by-draw procedures for selecting a PPS sample without replacement with a fixed $n \geq 2$. In the rest of this section, we introduce three popular unequal probability sampling methods which represent three different approaches to PPS sampling without replacement: (i) systematic sampling; (ii) rejective sampling; and (iii) random grouping.

4.4.2 *The Randomized Systematic PPS Sampling Method*

There exist a few general sampling procedures for selecting a PPS sample with any desirable sample size n. The randomized systematic PPS sampling procedure is the simplest one. It is a generalization from the simple systematic sampling method described in Sect. 2.3.

Suppose that the N units on the sampling frame have been arranged in a particular order: $\{1, 2, \cdots, N\}$. Let $n \geq 2$ be the planned sampling size. Let $\pi_i = nz_i$ be the intended first order inclusion probabilities. We assume that $z_i \leq 1/n$ for each i. The *systematic PPS sampling* procedure consists of two steps:

Step 1 Calculate the accumulated inclusion probabilities for the given order of units on the sampling frame: Let $b_0 = 0$; let $b_j = \sum_{i=1}^{j} \pi_i$, $j = 1, 2, \cdots, N$, with $b_N = n$.

Step 2 Generate a random number r from $U[0,\ 1]$, the uniform distribution over $[0,\ 1]$, and select unit i into the sampled set of units, $\mathbf{S}$, if

$$b_{i-1} < r + k \leq b_i \quad \text{for some integer } k, \quad k = 0, 1, 2, \cdots, n-1\,.$$

Since $b_0 = 0$ and $b_N = n$, there will be exactly n units selected for $\mathbf{S}$. It can be shown that $P(i \in \mathbf{S}) = b_i - b_{i-1} = \pi_i$, where $\pi_i = nz_i$ are specified prior to the sample selections (Problem 4.6).

The systematic PPS sampling procedure depends on the particular order of units on the sampling frame, which forms the base for Step 1. The procedure was first described by Goodman and Kish (1950) as a controlled sample selection method, because each selected unit represents a block of units on the list and the n sampled units are spread over the sampling frame. A major consequence from the pre-ordered sampling frame is that the second order inclusion probabilities π_{ij} are zero for certain pairs (i, j). For instance, if $b_2 = \pi_1 + \pi_2 < 1$, we will have $\pi_{12} = 0$, because the units 1 and 2 can never be simultaneously selected for the same sample.

The *randomized systematic PPS sampling* method (Hartley and Rao 1962) adds the following initial step before selecting the final sample using the aforementioned Steps 1 and 2:

Step 0 Randomize the order of the N units on the sampling frame; Denote the randomized order as $\{1, 2, \cdots, N\}$.

The final sample selected through the randomized systematic PPS sampling procedure maintains the basic property $\pi_i = nz_i$ with fixed sample size n. It also satisfies $\pi_{ij} > 0$ for all pairs (i, j). The latter property ensures unbiased variance estimation for the Horvitz-Thompson estimator. Unfortunately, there are no closed form expressions available for π_{ij}. Hartley and Rao (1962) provided an approximation formula for π_{ij}. In practice, PPS sampling methods are often used for selecting clusters in multistage cluster sampling where the total number of clusters on the sampling frame may not be too large. Under such scenarios, the second order inclusion probabilities π_{ij} can be obtained through a simulation-based approach (Thompson and Wu 2008). An R function for the randomized systematic PPS sampling is included in Appendix A.1.

4.4.3 The Rao-Sampford Method

The term *rejective sampling* was formally employed by Hájek (1964) but the approach itself had been used at least since Yates and Grundy (1953). Rejective procedures usually involve drawing units with replacement, a scheme which is easy to implement for any pre-specified selection probabilities. During the sample selection process a partial sample is abandoned whenever a unit is selected more than once. The process then starts afresh until n distinct units are selected for the final sample. The most crucial theoretical component for any rejective sampling procedure, however, is to specify the selection probabilities at each draw so that the final sample has the desired first order inclusion probabilities.

The Rao-Sampford PPS sampling method (Rao 1965; Sampford 1967) can be executed through a rejective procedure. It selects n units from the frame list of N units with replacement: The first unit is selected with probabilities $p_i = z_i$, $i = 1, 2, \cdots, N$, and all $n - 1$ subsequent units are selected with probabilities $q_i = cz_i/(1 - nz_i)$, $i = 1, 2, \cdots, N$, where the constant c satisfies $\sum_{i=1}^{N} q_i = 1$. Any sample containing duplicated units is rejected immediately, and the sampling process restarts with a new set of n draws. The process continues until the final sample of n distinct units is selected.

The Rao-Sampford sampling method has been popular among survey practitioners due to its practical simplicity for implementation and several other attractive theoretical features. It is an exact PPS sampling method with $\pi_i = nz_i$ for any fixed $n \geq 2$, with closed form expressions available for the second order inclusion probabilities π_{ij}, which also satisfy $\pi_i\pi_j - \pi_{ij} > 0$ for all $i \neq j$ (Sampford 1967). It is more efficient than the systematic PPS sampling method of Goodman and Kish (1950) in terms of the variance of the Horvitz-Thompson estimator (Asok and Sukhatme 1976). R functions for selecting a sample by the Rao-Sampford method and for computing the second order inclusion probabilities can be found in Appendix A.2.

4.4.4 The Rao-Hartley-Cochran Method

Selecting one unit from a list can easily be done with any pre-specified probabilities. This leads to the idea of random grouping for which the N population units are randomly divided into n groups and only one unit is selected from each group.

The Rao-Hartley-Cochran (RHC) method (Rao et al. 1962) selects a sample of n units with the following two steps:

Step 1 Randomly split the population into n groups G_1, G_2, $\cdots$, G_n with pre-specified group sizes N_1, N_2, $\cdots$, N_n $(\sum_{g=1}^{n} N_g = N)$.

Step 2 Select one unit from the gth group with probabilities $p_i = z_i/T_{zg}$, $i \in G_g$ where $T_{zg} = \sum_{j \in G_g} z_j$, and do this independently for $g = 1, 2, \cdots, n$.

The RHC method is not an exact PPS sampling procedure because the final inclusion probabilities π_i are not equal exactly to nz_i. There are two stages of randomizations involved in the sampling process: random grouping of population units and random selection of a unit within each group. Let $\mathbf{S}$ be the set of the n units selected in the final sample. For any given grouping at the first stage, we have

$$P\big(i \in \mathbf{S} \mid i \in G_g\big) = \frac{z_i}{T_{zg}} \,.$$

If the group sizes $N_1, N_2, \cdots$ and N_n are all the same or chosen to be as close to each other as possible, we would expect $T_{z1} \doteq \cdots \doteq T_{zn}$ under random grouping. Noting that $\sum_{g=1}^{n} T_{zg} = \sum_{i=1}^{N} z_i = 1$, we have $T_{zg} \doteq 1/n$ and $P(i \in \mathbf{S} \mid i \in G_g) \doteq nz_i$.

It turns out that we only require the conditional inclusion probabilities $P(i \in \mathbf{S} \mid i \in G_g)$ in order to develop a theory for the estimation of T_y. For the realized grouping used to select the final sample, let $y_g^* = y_i$, $z_g^* = z_i$ and $p_g^* = z_i/T_{zg}$, where unit i is selected in the gth group. The RHC estimator of the population total T_y is given by

$$\hat{T}_{y\text{RHC}} = \sum_{g=1}^{n} \frac{y_g^*}{p_g^*} = \sum_{g=1}^{n} \frac{T_{zg}}{z_g^*} y_g^* .$$

In addition to the survey sample data $\{(y_g^*, z_g^*), g = 1, 2, \cdots, n\}$, the group totals T_{zg} and the group sizes N_g, $g = 1, 2, \cdots, n$ are also required to compute the variance estimator given below.

Theorem 4.5 *Under the Rao-Hartley-Cochran sampling design,*

(a) The RHC estimator is design-unbiased for T_y, i.e.,

$$E\big(\hat{T}_{y\text{RHC}}\big) = T_y .$$

(b) The design-based variance of $\hat{T}_{y\text{RHC}}$ is given by

$$V\big(\hat{T}_{y\text{RHC}}\big) = \frac{\sum_{g=1}^{n} N_g^2 - N}{N(N-1)} \left(\sum_{i=1}^{N} \frac{y_i^2}{z_i} - T_y^2 \right) . \tag{4.11}$$

(c) An unbiased variance estimator for $\hat{T}_{y\text{RHC}}$ is given by

$$v\big(\hat{T}_{y\text{RHC}}\big) = \frac{\sum_{g=1}^{n} N_g^2 - N}{N^2 - \sum_{g=1}^{n} N_g^2} \sum_{g=1}^{n} T_{zg} \left(\frac{y_g^*}{z_g^*} - \hat{T}_{y\text{RHC}} \right)^2 . \tag{4.12}$$

Proof Let $E_2(\cdot \mid G)$ and $V_2(\cdot \mid G)$ denote the conditional expectation and variance given the first stage groups; let E_1 and V_1 denote the expectation and variance with respect to the random grouping.

(a) For any given group G_g, the sampling distribution of y_g^*/p_g^* is given by

$$P\left(\frac{y_g^*}{p_g^*} = \frac{y_i}{p_i} \right) = p_i , \quad i \in G_g .$$

It follows that

$$E_2\left(\frac{y_g^*}{p_g^*} \mid G\right) = \sum_{i\in G_g} \frac{y_i}{p_i} p_i = \sum_{i\in G_g} y_i$$

and

$$E_2\big(\hat{T}_{y\mathrm{RHC}} \mid G\big) = \sum_{g=1}^{n} E_2\left(\frac{y_g^*}{p_g^*} \mid G\right) = \sum_{g=1}^{n}\sum_{i\in G_g} y_i = T_y\,,$$

which further leads to $E\big(\hat{T}_{y\mathrm{RHC}}\big) = E_1\big\{E_2\big(\hat{T}_{y\mathrm{RHC}} \mid G\big)\big\} = T_y$.

(b) Noting that $V_1\big\{E_2\big(\hat{T}_{y\mathrm{RHC}} \mid G\big)\big\} = V_1\big(T_y\big) = 0$ and y_g^*/p_g^*, $g = 1, 2, \cdots, n$ are conditionally independent given the groups, we have

$$V\big(\hat{T}_{y\mathrm{RHC}}\big) = E_1\big\{V_2\big(\hat{T}_{y\mathrm{RHC}} \mid G\big)\big\} = E_1\left\{\sum_{g=1}^{n} V_2\left(\frac{y_g^*}{p_g^*} \mid G\right)\right\}.$$

It is shown in Problem 4.8 that

$$\begin{aligned} V_2\left(\tfrac{y_g^*}{p_g^*} \mid G\right) &= \frac{1}{2}\textstyle\sum_{i\in G_g}\sum_{j\in G_g} z_i z_j \left(\frac{y_i}{z_i} - \frac{y_j}{z_j}\right)^2 \\ &= \frac{1}{2}\textstyle\sum_{i=1}^{N}\sum_{j=1}^{N} A_{ij}(g)\, z_i z_j \left(\frac{y_i}{z_i} - \frac{y_j}{z_j}\right)^2, \end{aligned}$$

where $A_{ij}(g) = 1$ if $i, j \in G_g$ and $A_{ij}(g) = 0$ otherwise. Since $E_1\big\{A_{ij}(g)\big\} = \{N_g(N_g - 1)\}/\{N(N-1)\}$ for any $i \neq j$, the final expression for $V\big(\hat{T}_{y\mathrm{RHC}}\big)$ follows from $\sum_{g=1}^{n} N_g(N_g - 1) = \sum_{g=1}^{n} N_g^2 - N$ and

$$\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} z_i z_j \left(\frac{y_i}{z_i} - \frac{y_j}{z_j}\right)^2 = \sum_{i=1}^{N} \frac{y_i^2}{z_i} - T_y^2\,.$$

(c) The proof is left as an exercise in Problem 4.8.

□

The variance formula given by (4.11) indicates that the optimal grouping with respect to $N_1, \cdots, N_n$ is to minimize $\sum_{g=1}^{n} N_g^2$ subject to $\sum_{g=1}^{n} N_g = N$. When N/n is an integer, the optimal choice is to set $N_g = N/n$ for all g. If $N = nR + k$ for some $k \geq 1$, the nearly optimal grouping strategy is to set k groups with size $R + 1$ and the remaining groups with size R.

The Rao-Hartley-Cochran strategy, i.e., the random grouping sampling design coupled with the use of the RHC estimator, has been used in several fields such as forestry, labor force surveys, audit sampling and other accounting applications due to its remarkable simplicity and good efficiency properties (Rao 2005). It is always more efficient than the PPS sampling with replacement strategy; see Example 4.4 of the next section.

4.5 PPS Sampling with Replacement

The concept of PPS sampling designs is motivated by the efficient estimation of the population total using the Horvitz-Thompson estimator under survey designs with fixed sample sizes. In practice, there are two major technical difficulties associated with any PPS sampling methods for drawing a sample with a fixed sample size n: The sample selection procedure to ensure that $\pi_i = nz_i$ and the calculation of second order inclusion probabilities π_{ij} for variance estimation.

The *PPS sampling with replacement* procedure is a natural alternative approach with both practical and theoretical appeal. It selects a unit from $\{1, 2, \cdots, N\}$ with the given probabilities $p_i = z_i$, and the process is repeated n times, independently, with each unit selected from the same full list $\{1, 2, \cdots, N\}$. Let $\mathbf{S}^*$ be the set of n selected units, including possibly repeated units under the with-replacement sample selection procedure. Let $\mathbf{S}$ be the set of distinct units included in $\mathbf{S}^*$.

Let Y_i and Z_i be the values of the study variable y and the size variable z for the ith selected unit, $i = 1, 2, \cdots, n$. The following estimator for the population total T_y under PPS sampling with replacement was referred to as the Hansen-Hurwitz estimator (Hansen and Hurwitz 1943), although the term "PPS sampling with replacement" was not used at that time:

$$\hat{T}_{y\text{HH}} = \sum_{i \in \mathbf{S}^*} \frac{y_i}{nz_i} = \frac{1}{n}\sum_{i=1}^{n} \frac{Y_i}{Z_i} = \frac{1}{n}\sum_{i=1}^{n} R_i \,, \tag{4.13}$$

where the HH in the subscript indicates "Hansen-Hurwitz". Note that we have used $\sum_{i=1}^n R_i$ instead of $\sum_{i \in \mathbf{S}^*} R_i$ to indicate the sequential selection of the n units. It should also be noted that $R_i = Y_i/Z_i$ is a random variable associated with the ith selection while y_i/z_i has a fixed value attached to unit i. The notation $\sum_{i \in \mathbf{S}^*} y_i/z_i$ emphasizes that the sum is over the random set $\mathbf{S}^*$.

Theorem 4.6 *Under PPS sampling with replacement,*

(a) The Hansen-Hurwitz estimator $\hat{T}_{y\text{HH}}$ is unbiased for T_y, i.e.,

$$E\big(\hat{T}_{y\text{HH}}\big) = T_y \,.$$

(continued)

Theorem 4.6 (continued)
(b) The theoretical variance of $\hat{T}_{y\text{HH}}$ *is given by*

$$V\big(\hat{T}_{y\text{HH}}\big) = \frac{1}{n}\sum_{i=1}^{N} z_i\Big(\frac{y_i}{z_i} - T_y\Big)^2 .$$

(c) An unbiased variance estimator is given by

$$v\big(\hat{T}_{y\text{HH}}\big) = \frac{1}{n(n-1)}\sum_{i\in\mathbf{S}^*}\Big(\frac{y_i}{z_i} - \hat{T}_{y\text{HH}}\Big)^2 .$$

Proof The most crucial observation is that $R_i, i = 1, 2, \cdots, n$ are independent and identically distributed random variables under PPS sampling with replacement, with the common distribution of the random variable R given by

$$P\Big(R = \frac{y_i}{z_i}\Big) = z_i , \quad i = 1, 2, \cdots, N ,$$

where the values $y_i/z_i, i = 1, 2, \cdots, N$ are the ratios of the y and the z values for the N units in the population. It follows that

$$E(R) = \sum_{i=1}^{N}\frac{y_i}{z_i}\cdot z_i = T_y \quad \text{and} \quad V(R) = E[\{R - E(R)\}^2] = \sum_{i=1}^{N} z_i\Big(\frac{y_i}{z_i} - T_y\Big)^2 .$$

The Hansen-Hurwitz estimator $\hat{T}_{y\text{HH}} = n^{-1}\sum_{i=1}^{n} R_i$ is the "sample mean" of n iid random variables. We have

$$E\big(\hat{T}_{y\text{HH}}\big) = E(R) = T_y \quad \text{and} \quad V\big(\hat{T}_{y\text{HH}}\big) = \frac{V(R)}{n} = \frac{1}{n}\sum_{i=1}^{N} z_i\Big(\frac{y_i}{z_i} - T_y\Big)^2 .$$

The unbiased variance estimator $v\big(\hat{T}_{y\text{HH}}\big)$ is obtained by replacing $V(R)$ in $V\big(\hat{T}_{y\text{HH}}\big)$ by the "sample variance"

$$s_R^2 = \frac{1}{n-1}\sum_{i\in\mathbf{S}^*}\Big(\frac{y_i}{z_i} - \hat{T}_{y\text{HH}}\Big)^2 .$$

□

An alternative proof of Theorem 4.5 is to let t_i be the number of times unit i is selected in the sample $\mathbf{S}^*$, $i = 1, 2, \cdots, N$, and to re-write $\hat{T}_{y\text{HH}}$ as

$$\hat{T}_{y\mathrm{HH}} = \frac{1}{n}\sum_{i=1}^{N} t_i \frac{y_i}{z_i} .$$

The t_i's are the only random quantities in the expression, and results of the theorem can be derived based on the fact that $(t_1, t_2, \cdots, t_N)$ follows $Multinomial(n;\ z_1, z_2, \cdots,\ z_N)$, with $E(t_i) = nz_i$, $V(t_i) = nz_i(1 - z_i)$ and $Cov(t_i, t_j) = -nz_i z_j$ for $i \neq j$.

Example 4.3 Single-stage cluster sampling with clusters selected by PPS sampling with replacement: Suppose that the population consists of K clusters, with M_i being the size for cluster i, $i = 1, 2, \cdots, K$. The overall population size is $N = \sum_{i=1}^{K} M_i$. Consider single-stage cluster sampling, for which we select k clusters by PPS sampling with replacement, with selection probabilities $z_i = M_i/N$, $i = 1, 2, \cdots, K$. Let $\mathbf{S}_c^*$ be the set of k clusters selected, including the possibly repeated clusters. The cluster total $T_i = \sum_{j=1}^{M_i} y_{ij}$ is observed for $i \in \mathbf{S}_c^*$.

The Hansen-Hurwitz estimator of $T_y = \sum_{i=1}^{K} T_i$ is given by

$$\hat{T}_{y\mathrm{HH}} = \sum_{i\in\mathbf{S}_c^*} \frac{T_i}{kz_i} = \frac{N}{k}\sum_{i\in\mathbf{S}_c^*} \mu_i ,$$

where $\mu_i = T_i/M_i$ is the cluster mean. The theoretical variance of $\hat{T}_{y\mathrm{HH}}$ is given by

$$V\big(\hat{T}_{y\mathrm{HH}}\big) = \frac{1}{k}\sum_{i=1}^{K} z_i\Big(\frac{T_i}{z_i} - T_y\Big)^2 = \frac{N}{k}\sum_{i=1}^{K} M_i\big(\mu_i - \mu_y\big)^2 .$$

For the purpose of comparison, consider the basic sampling design where the k clusters are selected by SRSWOR. Let $\mathbf{S}_c$ be the set of k selected clusters. Under this design the unbiased estimator of $T_y = K\mu_T$, where $\mu_T = K^{-1}\sum_{i=1}^{K} T_i$, is given by

$$\hat{T}_y = K\hat{\mu}_T = K\,\frac{1}{k}\sum_{i\in\mathbf{S}_c} T_i ,$$

and the related theoretical variance is given by

$$V\big(\hat{T}_y\big) = K^2\Big(1 - \frac{k}{K}\Big)\frac{1}{k}\frac{1}{K-1}\sum_{i=1}^{K}\big(T_i - \mu_T\big)^2 .$$

A crucial observation for the comparison between the two theoretical variances under the two sampling designs is that $V\big(\hat{T}_{y\mathrm{HH}}\big)$ depends on variations among the cluster means μ_i's while $V\big(\hat{T}_y\big)$ depends on the variations among the cluster totals T_i's. When the cluster sizes M_i are very different among different clusters, the variations among the T_i's can be dramatic, and consequently, the value of $V\big(\hat{T}_y\big)$ can be much bigger than $V\big(\hat{T}_{y\text{HH}}\big)$ since the cluster means μ_i's are less affected by the cluster sizes. In the extreme case of $y_{ij} = c$ for all i, j, we have $\mu_i = c$ for all i and $V\big(\hat{T}_{y\text{HH}}\big) = 0$. It can be seen that $V\big(\hat{T}_y\big) > 0$ in this case unless all cluster sizes M_i are the same. The PPS sampling with replacement method for selecting $\mathbf{S}_c^*$ is more efficient than SRSWOR for selecting $\mathbf{S}_c$. ◇

PPS sampling with replacement, however, is generally less efficient than PPS sampling without replacement with fixed sample sizes. This is similar to the comparison between SRSWOR and SRSWR. The following example provides partial evidence on efficiency comparisons between the two types of sampling designs.

Example 4.4 The Rao-Hartley-Cochran sampling method can be viewed as an approximate PPS sampling procedure with fixed sample size. The variance of the RHC estimator is given by (4.11). The variance of the Hansen-Hurwitz estimator under PPS sampling with replacement (Theorem 4.6) can alternatively be written as

$$V\big(\hat{T}_{y\text{HH}}\big) = \frac{1}{n}\left(\sum_{i=1}^{N}\frac{y_i^2}{z_i} - T_y^2\right).$$

It follows that

$$V\big(\hat{T}_{y\text{RHC}}\big) = n\frac{\sum_{g=1}^{n} N_g^2 - N}{N(N-1)}\, V\big(\hat{T}_{y\text{HH}}\big).$$

Consider scenarios where N/n is an integer and $N_g = N/n$ for all g. We have

$$n\frac{\sum_{g=1}^{n} N_g^2 - N}{N(N-1)} = \frac{N-n}{N-1} < 1$$

when $n > 1$. The RHC estimation strategy is generally more efficient than the HH estimator under PPS sampling without replacement and the same number of draws. When N is very large and n/N is small, the difference between the two methods becomes negligible. ◇

The methodology and estimation theory of PPS sampling with replacement, however, provide an important alternative tool with practical implications. Variance estimation is one of the major tasks for survey data analysis and requires second order inclusion probabilities π_{ij} for all pairs (i, j) in the sample. It is known that calculations of the π_{ij}'s under certain without-replacement sampling methods can be extremely difficult and sometime even impossible. It is common practice in survey data analysis to assume that sampled units are selected by a with-replacement sampling procedure for the purpose of variance estimation, even if the original survey design involves a without-replacement sampling method. This can be justified under the conditions that the population size N is large and the sampling fraction n/N is small.

Suppose that the original sample is selected by a PPS sampling method (without replacement) with a fixed sample size n. The first order inclusion probabilities are given by $\pi_i = nz_i$, where $z_i > 0$ and $\sum_{i=1}^N z_i = 1$. If we (incorrectly) assume that the original sample is obtained by the PPS sampling with replacement procedure where the ith unit is selected with probabilities $p_i = z_i$, what will be the consequences? Let $\mathbf{S}$ be the set of distinct units from the n selected units under the assumed with-replacement procedure. Let $m = |\mathbf{S}|$ be the number of distinct units selected. It is apparent that $m \le n$ and $\pi_i^\circ = P(i \in \mathbf{S}) \neq nz_i$. However, when N is large and $z_i = O(1/N)$, we have

$$\pi_i^\circ = P(i \in \mathbf{S}) = 1 - (1 - z_i)^n \doteq 1 - (1 - nz_i) = nz_i \,,$$

where the approximation $(1 - z_i)^n \doteq 1 - nz_i$ is valid for any fixed n with $z_i = o(1)$. The random sample size m will also be very close to n when n/N is small, because the probability of selecting repeated units is small (See Problem 4.2 for details on SRSWR). In other words, the first order inclusion probabilities and the realized sample size under the assumed with-replacement sampling procedure remain approximately the same as those under the original sampling design. The two estimators

$$\hat{T}_{y\mathrm{HH}} = \sum_{i \in \mathbf{S}^*} \frac{y_i}{nz_i} \qquad \text{and} \qquad \hat{T}_{y\mathrm{HT}} = \sum_{i \in \mathbf{S}} \frac{y}{\pi_i^\circ}$$

have similar performances. For finite samples, the potentially duplicated units in $\mathbf{S}^*$ used in $\hat{T}_{y\mathrm{HH}}$ are compensated by the fact that the π_i° used in $\hat{T}_{y\mathrm{HT}}$ satisfy $\pi_i^\circ = 1 - (1 - z_i)^n < nz_i$ for all i.

For large scale complex surveys, the sampling design often involves multi-stage sampling, with the first stage clusters selected by an unequal probability sampling method. When the total number of first stage sampling units is reasonably large and the sampling fraction is relatively small, it is often assumed for the purpose of variance estimation that the first stage clusters are selected by PPS sampling with replacement to avoid the difficulties of computing the second order inclusion probabilities.

Example 4.5 (Multi-Stage Cluster Sampling with First Stage Clusters Selected by PPS Sampling with Replacement) Let $T_y = \sum_{i=1}^K T_i$ be the population total, where T_i is the total for cluster i, $i = 1, 2, \cdots, K$. Let z_i be the size measure for the first stage clusters such that $z_i > 0$ and $\sum_{i=1}^K z_i = 1$. We assume that the sampling design has the following three features:

- The first stage sample consists of k first stage clusters selected by PPS sampling with replacement, with selection probabilities $z_1, z_2, \cdots, z_K$.
- The sampling design used for selecting a sample within a selected first stage cluster permits unbiased estimation of the cluster total.

- Samples selected from different clusters are independent of each other.

Let $E_1(\cdot)$ and $V_1(\cdot)$ be the expectation and variance under the first stage sampling design; let $E_2(\cdot)$ and $V_2(\cdot)$ denote the expectation and variance under the sampling design within the first stage clusters, possibly unequal probability multi-stage sampling designs as well. Let $\hat{T}_i$ be the estimator of T_i using data selected within cluster i, such that $E_2(\hat{T}_i) = T_i$. Without loss of generality, we assume that $i = 1, 2, \cdots, k$ are the first stage sampled clusters with some possibly duplicated. The Hansen-Hurwitz type estimator for T_y is given by

$$\hat{T}_{y\text{HH}} = \sum_{i=1}^{k} \frac{\hat{T}_i}{kz_i} = \frac{1}{k} \sum_{i=1}^{k} \hat{r}_i \,,$$

where $\hat{r}_i = \hat{T}_i/z_i$. Under the first stage sampling design, the $\hat{r}_i$'s are independent and identically distributed random variables. It follows that

$$E_p(\hat{T}_{y\text{HH}}) = E_1\Big\{ \sum_{i=1}^{k} \frac{E_2(\hat{T}_i)}{kz_i} \Big\} = E_1\Big\{ \sum_{i=1}^{k} \frac{T_i}{kz_i} \Big\} = T_y \,.$$

A design-unbiased variance estimator is given by

$$v_p(\hat{T}_{y\text{HH}}) = \frac{1}{k(k-1)} \sum_{i=1}^{k} \Big(\frac{\hat{T}_i}{z_i} - \hat{T}_{y\text{HH}} \Big)^2 \,.$$

Proof of the unbiasedness of the variance estimator is left as an exercise in Problem 4.10. ◇

4.6 Poisson Sampling

Another alternative PPS sampling procedure is Poisson sampling. The sampling procedure consists of N independent Bernoulli trials over the N units in the population. Let $\pi_i \in (0, 1)$ be the desired probability to include unit i in the sample, $i = 1, 2, \cdots, N$. Those inclusion probabilities are specified before the sample selection takes place.

Step 1 For unit i, generate a random number r_i from $U[0, 1]$; select unit i if $r_i \le \pi_i$.
Step 2 Repeat Step 1 for $i = 1, 2, \cdots, N$, independently.

Let $\mathbf{S}$ be the set of sampled units. It is clear that under Poisson sampling no unit is selected more than once. It is also apparent that $P(i \in \mathbf{S}) = \pi_i$, $i = 1, 2, \cdots, N$.

Due to independent selections among different units, the second order inclusion probabilities are given by $\pi_{ij} = \pi_i \pi_j$ for $i \neq j$.

Theorem 4.7 *Under Poisson sampling with pre-specified first order inclusion probabilities π_i, $i = 1, 2, \cdots, N$:*

(a) The Horvitz-Thompson estimator $\hat{T}_{y\text{HT}} = \sum_{i\in\mathbf{S}} y_i/\pi_i$ is unbiased for T_y.
(b) The theoretical variance of $\hat{T}_{y\text{HT}}$ is given by

$$V\big(\hat{T}_{y\text{HT}}\big) = \sum_{i=1}^{N} \pi_i(1-\pi_i)\Big(\frac{y_i}{\pi_i}\Big)^2 = \sum_{i=1}^{N} \frac{1-\pi_i}{\pi_i} y_i^2 .$$

(c) An unbiased variance estimator is given by

$$v\big(\hat{T}_{y\text{HT}}\big) = \sum_{i\in\mathbf{S}} \frac{1-\pi_i}{\pi_i^2} y_i^2 .$$

The proof of Theorem 4.6 follows from the same arguments used in the proof of Theorem 4.2, plus the fact that $A_1, A_2, \cdots, A_N$ are independent, where $A_i = I(i \in \mathbf{S})$ are the sample inclusion indicators.

One of the major drawbacks of Poisson sampling is that the final sample size $n = |\mathbf{S}|$ is random. Using the expression $n = \sum_{i=1}^{N} A_i$, we have

$$E(n) = \sum_{i=1}^{N} \pi_i \quad \text{and} \quad V(n) = \sum_{i=1}^{N} \pi_i(1-\pi_i) .$$

It turns out that the variation induced by the random sample size is quite large and the Horvitz-Thompson estimator under Poisson sampling is usually not efficient.

Example 4.6 (Bernoulli Sampling) Let n_0 be the planned sample size and N be the population size. The so-called Bernoulli sampling is a special case of Poisson sampling where $\pi_i = n_0/N$ for all i. Let n be the final sample size. It follows that $E(n) = n_0$ and $V(n) = n_0(1 - n_0/N)$. Suppose that $n_0 = 100$ and N is large. Then we have $E(n) = 100$ and $V(n) \doteq 100$. A 95% prediction interval for the final sample size n is given approximately by (80, 120), which demonstrates the potential amount of variation in estimators caused by the variability in the final sample size n. ◇

Poisson sampling is seldom used in practice due to its inefficiency. However, it has rich theoretical value through a related rejective sampling procedure. Hájek (1964, 1981) considered a version of rejective sampling which was a conditional Poisson sampling, where samples with sizes not equal to the pre-specified n

are rejected, and the sampling selection process continues until a final sample with the exact sample size n is obtained. With the increased computing power of recent years, rejective sampling procedures can be used in practice, and their theoretical properties have attracted more attention from survey statisticians. Further discussions on Poisson sampling and conditional Poisson sampling with additional references can be found in Tillé (2006).

4.7 Problems

4.1 (Sample Inclusion Probabilities)

(a) Let the sample $\mathbf{S}$ be the set of distinct units from simple random sampling with replacement (SRSWR) described in Sect. 2.2 with n selections. Let $\pi_i = P(i \in \mathbf{S})$ and $\pi_{ij} = P(i, j \in \mathbf{S})$. Show that

$$\pi_i = 1 - \big(1 - N^{-1}\big)^n\,, \quad i = 1, 2, \cdots, N\,,$$
$$\pi_{ij} = 1 - 2\big(1 - N^{-1}\big)^n + \big(1 - 2/N\big)^n\,, \quad i, j = 1, 2, \cdots, N \text{ and } i \neq j\,.$$

(b) Continued from (a): Show that, when N is large, the first and the second order inclusion probabilities under SRSWR are very close to those under SRSWOR.

(c) Use the format of (4.1) to express the first and the second order inclusion probabilities for single-stage cluster sampling with clusters selected by SRSWOR (Sect. 3.3).

(d) Use the format of (4.1) to express the first and the second order inclusion probabilities for two-stage cluster sampling with SRSWOR at both stages (Sect. 3.4).

4.2 (Simple Random Sampling with Replacement) Let $\mathbf{S}^*$ be the set of n units selected from $\{1, 2, \cdots, N\}$ using SRSWR, including all repeated units. Let m be the number of distinct units in $\mathbf{S}^*$.

(a) Find an algebraic expression for $P(m = n)$, and denote it as $p(N, n)$. This is the probability that no units are selected more than once under SRSWR.

(b) Suppose that $N = 100$. Compute the values of $p(N, n)$ for $n = 5, 4, 3$ and 2. This gives an approximate picture for the probabilities $p(N, n)$ when the sampling fractions (n/N) are 5%, 4%, 3% and 2%.

4.3 (Godambe Class of Linear Estimators) Consider the subclass of the Godambe class of linear estimators, $\mathscr{G}_1$, defined in Sect. 4.2.

(a) Find $E\big(\hat{T}_{yc}\big)$ and $V\big(\hat{T}_{yc}\big)$ under a general sampling design with inclusion probabilities π_i and π_{ij}, where $\hat{T}_{yc} = \sum_{i \in \mathbf{S}} c_i y_i \in \mathscr{G}_1$.

(b) Argue that the Horvitz-Thompson estimator is the only unbiased estimator of T_y in the class $\mathscr{G}_1$.

(c) Show that for any sampling design there exists an unbiased estimator for the population total T_y if and only if $\pi_i = P(i \in \mathbf{S}) > 0$ for $i = 1, 2, \cdots, N$.

4.4 (Simple Random Sampling Without Replacement) Let $\bar{y}$ be the sample mean and $\hat{T}_{y\text{HT}} = N\bar{y}$ be the HT estimator for T_y.

(a) Show that the Yates-Grundy-Sen variance formula (4.7) reduces to $N^2V(\bar{y})$ where $V(\bar{y}) = (1 - n/N)\sigma_y^2/n$.

(b) Show that the Yates-Grundy-Sen variance estimator (4.9) reduces to $N^2v(\bar{y})$ where $v(\bar{y}) = (1 - n/N)s_y^2/n$.

4.5 (Brewer's PPS Sampling Method for $n = 2$) For Brewer's PPS sampling method described in Example 4.2, show that $\pi_i\pi_j - \pi_{ij} > 0$ for all $i \neq j$. **Hint**: Show that

$$2D(1-2z_i)(1-2z_j) - (1 - z_i - z_j) = \left(\sum_{k=1,k\neq i,j}^{N} \frac{z_k}{1-2z_k}\right)(1-2z_i)(1-2z_j)\,.$$

4.6 (The Systematic PPS Sampling Method) Suppose that the sample $\mathbf{S}$ is selected by the systematic PPS sampling method (Steps 1 and 2) described in Sect. 4.4. Show that $P(i \in \mathbf{S}) = b_i - b_{i-1} = \pi_i$.

Hint: Consider two possible scenarios: (i) There exists an integer k $(k = 0, 1, \cdots, n-1)$ such that $k \leq b_{i-1} < b_i \leq k + 1$; (ii) There exists an integer k $(k = 1, \cdots, n-1)$ such that $b_{i-1} \leq k < b_i$.

4.7 (Midzuno's Method for Unequal Probability Sampling (Midzuno 1952)) Select the first unit with probability q_i, $i = 1, 2, \cdots, N$: $q_i > 0$ and $\sum_{i=1}^{N} q_i = 1$. Keep the first selected unit aside, and draw a sample of size $n-1$ from the remaining $N-1$ elements in the population using SRSWOR. Let $\mathbf{S}$ be the set of n units selected in the final sample.

(a) Show that the sampling design satisfies

$$P(\mathbf{S}) = \binom{N-1}{n-1}^{-1} \sum_{i\in\mathbf{S}} q_i\,.$$

(b) Show that the first order inclusion probabilities are given by

$$\pi_i = \frac{N-n}{N-1}q_i + \frac{n-1}{N-1}\,.$$

(c) Show that the second order inclusion probabilities are given by

$$\pi_{ij} = \frac{(n-1)(N-n)}{(N-1)(N-2)}(q_i + q_j) + \frac{(n-1)(n-2)}{(N-1)(N-2)}\,.$$

(d) Show that for a pre-specified set of inclusion probabilities $\pi_i, i = 1, 2, \cdots, N$, we might be able to choose q_i accordingly. Are there any restrictions on π_i?

4.8 (The Rao-Hartley-Cochran Method) Further details for the proofs of Parts (b) and (c) in Theorem 4.5:

(a) Show that

$$V_2\left(\frac{y_g^*}{p_g^*} \mid G\right) = \frac{1}{2}\sum_{i\in G_g}\sum_{j\in G_g} z_i z_j \left(\frac{y_i}{z_i} - \frac{y_j}{z_j}\right)^2 .$$

Hint: Consider the Yates-Grundy-Sen variance formula for the HT estimator with $n = 1$, $\pi_i = p_i$ and $\pi_{ij} = 0$ for $i \neq j$.

(b) Show that $v(\hat{T}_{y\text{RHC}})$ given in (4.12) is an unbiased variance estimator for the RHC estimator.

4.9 (Comparisons Among Different Sampling Methods) Let the finite population consist of $N = 3$ units, with values x_i and y_i given below, where x_i is the size measure for the ith unit.

Unit (i)	x_i	y_i	y_i/x_i
1	20	1.20	0.06
2	30	2.10	0.07
3	15	0.75	0.05

(a) Describe how to take a PPS sample of $n = 2$ using Brewer's method. Compute the theoretical variance $V(\hat{T}_{y\text{HT}})$.
(b) Describe how to take a PPS sample with expected sample sine $n = 2$ using Poisson sampling. Compute the theoretical variance $V(\hat{T}_{y\text{HT}})$.
(c) Describe how to take a PPS sample of $n = 2$ under selection with replacement. Compute the theoretical variance $V(\hat{T}_{y\text{HH}})$.
(d) Compute $V(\bar{y})$ for $n = 2$ under SRSWOR.
(e) Compare the results from (a), (b), (c), (d) and comment.

4.10 (Multi-Stage Cluster Sampling with PPS Sampling with Replacement) Consider the sampling design described in Example 4.5 where the first stage clusters are selected by PPS sampling with replacement. Let $\hat{T}_{y\text{HH}} = \sum_{i=1}^{k} \hat{T}_i/(kz_i) = k^{-1}\sum_{i=1}^{k} \hat{r}_i$, where $\hat{r}_i = \hat{T}_i/z_i$.

(a) Show that the design-based variance of $\hat{T}_{y\text{HH}}$ can be expressed as

$$V_p\big(\hat{T}_{y\text{HH}}\big) = V_1\Big(\frac{1}{k}\sum_{i=1}^{k}\frac{T_i}{z_i}\Big) + E_1\Big\{\frac{1}{k^2}\sum_{i=1}^{k} V_2\Big(\frac{\hat{T}_i}{z_i}\Big)\Big\}$$

(b) Show that the variance estimator given by

$$v_p\big(\hat{T}_{y\mathrm{HH}}\big) = \frac{1}{k(k-1)}\sum_{i=1}^{k}\Big(\frac{\hat{T}_i}{z_i} - \hat{T}_{y\mathrm{HH}}\Big)^2$$

is design-unbiased, i.e., $E_p\big(v_p\big\{\hat{T}_{y\mathrm{HH}}\big)\big\} = V_p\big(\hat{T}_{y\mathrm{HH}}\big)$.

4.11 (Two-Stage Cluster Sampling with Self-Weighting) Suppose that the finite population consists of K clusters with size M_i for cluster i. Let $N = \sum_{i=1}^{K} M_i$ be the population size. Let $n = km$ be the overall sample size, where k is the size of the first stage sample $\mathbf{S}_c$ and m is the fixed size of the second stage sample $\mathbf{S}_i$ for all i. See Sect. 3.4 for further notational details.

(a) Suppose that the first stage sample $\mathbf{S}_c$ is selected by a PPS sampling method without replacement, with inclusion probabilities proportional to the cluster sizes M_i; the second stage sample $\mathbf{S}_i$ is selected by SRSWOR. Show that the Horvitz-Thompson estimator of the population mean μ_y is given by

$$\hat{\mu}_{y\mathrm{HT}} = \bar{y} = \frac{1}{n}\sum_{i\in\mathbf{S}_c}\sum_{j\in\mathbf{S}_i} y_{ij}\,.$$

This is an example of the so-called self-weighting survey designs where the final survey weights are equal for all units.

(b) Suppose that the first stage sampling fraction is small and the variance estimator given in Part (b) of Problem 4.10 can be used. Show that the variance estimator for $\hat{\mu}_{y\mathrm{HT}}$ is given by

$$v_p\big(\hat{\mu}_{y\mathrm{HT}}\big) = \frac{1}{k(k-1)}\sum_{i=1}^{k}\big(\bar{y}_i - \bar{y}\big)^2\,,$$

where $\bar{y}_i = m^{-1}\sum_{j\in\mathbf{S}_i} y_{ij}$ and $\bar{y} = \hat{\mu}_{y\mathrm{HT}}$.

Chapter 5
Model-Based Prediction and Model-Assisted Estimation

This chapter provides an introduction to the model-based prediction approach to survey sampling and model-assisted estimation in the presence of auxiliary population information. The two approaches, model-based prediction and design-based estimation, are conceptually different and have been the subject of debate among survey statisticians. The general model-assisted framework, however, enables the strength of plausible models to be built into design-based inferences for survey samples.

Information on certain covariates $\mathbf{x}$ is often available at the population level prior to the planning and execution of the survey. This is commonly referred to as *auxiliary information*. There are two common scenarios in practical situations:

(i) Values $\{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N\}$ are known for the entire survey population. This is the so-called *complete auxiliary information*.
(ii) The population means $\mu_{\mathbf{x}} = N^{-1}\sum_{i=1}^N \mathbf{x}_i$ or the population totals $T_{\mathbf{x}} = \sum_{i=1}^N \mathbf{x}_i$ are known.

Auxiliary information may be obtained from previous censuses and their projections for the current time. It could be available from administrative data such as tax files, business registries and health records. It can also be obtained from previous surveys where the sample sizes are large and the estimates from the survey data are deemed to be very reliable and can be treated as relevant population information. In natural resource inventory surveys, useful auxiliary information can be derived from preliminary analyses of aerial pictures or satellite images.

In general, auxiliary information can be used at either the survey design stage or the estimation stage or both. Variables defining stratum membership are used for stratified sampling designs, and the variable defining the size measure of units is required for PPS sampling. The focus of this chapter is on how to use auxiliary information at the estimation stage. The amount of available auxiliary information often dictates the type of estimation strategies to be used. By default, if auxiliary variables $\mathbf{x}$ are involved, values of both the response variable y and $\mathbf{x}$ are collected

C. Wu, M. E. Thompson, *Sampling Theory and Practice*, ICSA Book Series in Statistics, https://doi.org/10.1007/978-3-030-44246-0_5

for units included in the current survey sample. In this case survey data consist of $\{(y_i, \mathbf{x}_i), i \in \mathbf{S}\}$ plus the auxiliary population information on $\mathbf{x}$.

5.1 Model-Based Prediction Methods

One of the best-known quotes in statistics is "*All models are wrong, but some are useful*" by the late Professor George E. P. Box. The true message behind the quote is that models need to be used with caution and, when models are deemed to be plausible for the problem under study, they can be very useful. As a matter of fact, no statistical methodologies are completely model-free, and use of models often becomes unavoidable in many practical situations.

The use of statistical models in finite population sampling, however, has been a subject for debate among survey statisticians. The cornerstone of the foundation for design-based inference in survey sampling was laid out by Neyman (1934). Under the design-based framework, the survey population is a fixed finite set of N units. Values of the response variable y and covariates $\mathbf{x}$ associated with a particular unit are also viewed as fixed. As a consequence, finite population parameters such as μ_y and T_y are fixed quantities as well. The primary source of randomization is induced by the probability sampling design for selecting the survey sample. Probability models under the design-based framework are discrete and nonparametric in nature, suitable for dealing with the estimation of simple descriptive population parameters. More sophisticated models may also be used under the design-based framework for analytic use of survey data; see Part II of the book for further detail.

Model-based inferences for survey sampling are built on the concept of a *superpopulation model*, which was first introduced by Cochran (1939) and was briefly mentioned in Sect. 4.3. Suppose that the finite population values $\{y_1, y_2, \cdots, y_N\}$ can be viewed as a random sample from a superpopulation model, ξ. Under the model, the y_i's are generally viewed as random variables with probability distributions specified by the assumed model.

Model I The common mean model for the survey population:

$$y_i = \mu + \varepsilon_i \,, \quad i = 1, 2, \cdots, N \,,$$

where ε_1, $\cdots$, ε_N are independent and identically distributed error terms with $E_\xi(\varepsilon_i) = 0$ and $V_\xi(\varepsilon_i) = \sigma^2$. Here E_ξ and V_ξ refer to expectation and variance under the model, ξ. The μ and σ^2 are superpopulation parameters, which are conceptually different from the finite population parameters μ_y and σ_y^2.

Let $x > 0$ be a covariate and (y_i, x_i) be the values of y and x associated with unit i. The relation between y and x may be conveniently described by a simple linear regression model.

Model II The simple linear regression model without an intercept:

$$y_i = \beta x_i + \varepsilon_i \,, \quad i = 1, 2, \cdots, N \,,$$

where $\varepsilon_1, \cdots, \varepsilon_N$ are independent error terms with $E_\xi(\varepsilon_i) = 0$ and $V_\xi(\varepsilon_i) = v_i \sigma^2$. For any i, the v_i is a known constant often depending on x_i. The β and σ^2 are superpopulation parameters. The model (4.10) used in Sect. 4.3 is a special case of Model II with $z_i = x_i$ and $v_i = x_i$.

Model III The simple linear regression model with an intercept:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \,, \quad i = 1, 2, \cdots, N \,,$$

where $\varepsilon_1, \cdots, \varepsilon_N$ are independent error terms with $E_\xi(\varepsilon_i) = 0$ and $V_\xi(\varepsilon_i) = v_i \sigma^2$. The superpopulation parameters in this case are β_0, β_1 and σ^2.

5.1.1 Model-Unbiased Prediction Estimator for μ_y

Under the assumed superpopulation model, the finite population mean μ_y needs to be viewed as a random quantity since it is defined through $y_1, y_2, \cdots, y_N$. Estimation of μ_y under the model becomes a prediction problem.

We assume that the superpopulation ξ, which is assumed to hold for the survey population, also holds for the survey sample **S**, regardless of how **S** is selected. Under such scenarios the probability sampling design is referred to as "*ignorable*" for the model-based prediction approach.

Let $\hat{\mu}_y$ be an estimator of μ_y, computed based on the survey sample data and the available auxiliary information. We term $\hat{\mu}_y$ a *model-unbiased prediction estimator* of μ_y if

$$E_\xi\big(\hat{\mu}_y - \mu_y\big) = 0 \,.$$

It is apparent that construction of a model-based prediction estimator depends on the assumed model as well as the type of auxiliary information that is available. Model-unbiasedness needs to be evaluated under the assumed model.

Example 5.1 Suppose that the survey population follows the common mean model (Model I). Let $\{y_i, i \in \mathbf{S}\}$ be the survey sample data, which also follow Model I, i.e., $y_i = \mu + \varepsilon_i, i \in \mathbf{S}$, where $\varepsilon_i, i \in \mathbf{S}$ are independent and identically distributed with $E_\xi(\varepsilon_i) = 0$ and $V_\xi(\varepsilon_i) = \sigma^2$. Under the assumed model, we have $E_\xi(y_i) = \mu$. Let $\hat{\mu}_y = \bar{y} = n^{-1} \sum_{i \in \mathbf{S}} y_i$ be the sample mean. It follows that

$$E_\xi\big(\hat{\mu}_y - \mu_y\big) = \mu - \mu = 0 \,.$$

The sample mean $\hat{\mu}_y = \bar{y}$ is a model-unbiased prediction estimator for μ_y under the common mean model. ◇

Example 5.2 Suppose that the survey population follows the simple linear regression model without an intercept (Model II). Under the assumed model, we have

$$E_\xi(y_i \mid x_i) = \beta x_i \quad \text{and} \quad V_\xi(y_i \mid x_i) = v_i\sigma^2\,, \quad i = 1, 2, \cdots, N\,.$$

For simplicity of notation, we will consider x to be non-random and use $E_\xi(y_i)$ and $V_\xi(y_i)$ to denote the conditional expectation and the conditional variance under regression models. The y_i's are also independent conditional on the given x_i's.

Let $\{(y_i, x_i), i \in \mathbf{S}\}$ be the survey sample data, which also follow Model II. Suppose that $v_i = 1$ and $V_\xi(\varepsilon_i) = \sigma^2$. The ordinary least squares estimator for β is given by

$$\hat{\beta} = \frac{\sum_{i\in\mathbf{S}} x_i y_i}{\sum_{i\in\mathbf{S}} x_i^2}\,. \tag{5.1}$$

It can be verified that $E_\xi\big(\hat{\beta}\big) = \beta$. Suppose that the population mean $\mu_x = N^{-1}\sum_{i=1}^N x_i$ is known. A model-based prediction estimator for μ_y can be computed as $\hat{\mu}_y = \hat{\beta}\mu_x$. We have

$$E_\xi\big(\hat{\mu}_y - \mu_y\big) = E_\xi\big(\hat{\beta}\big)\mu_x - \frac{1}{N}\sum_{i=1}^N E_\xi(y_i) = \beta\mu_x - \beta\mu_x = 0\,.$$

The estimator $\hat{\mu}_y = \hat{\beta}\mu_x$ is model-unbiased under Model II. ◇

Example 5.3 Another model-based prediction estimator for μ_y under Model II can be constructed as follows. Re-write μ_y as

$$\mu_y = \frac{1}{N}\Big(\sum_{i\in\mathbf{S}} y_i + \sum_{i\notin\mathbf{S}} y_i\Big)\,.$$

For $i \notin \mathbf{S}$, the value y_i can be predicted as $\hat{y}_i = \hat{\beta}x_i$. We can therefore use the following prediction estimator

$$\hat{\mu}_y = \frac{1}{N}\Big(\sum_{i\in\mathbf{S}} y_i + \sum_{i\notin\mathbf{S}} \hat{y}_i\Big) = \frac{1}{N}\big\{n\bar{y} + \hat{\beta}(N\mu_x - n\bar{x})\big\} = \hat{\beta}\mu_x + \frac{n}{N}\big(\bar{y} - \hat{\beta}\bar{x}\big)\,,$$

where $\bar{x} = n^{-1}\sum_{i\in\mathbf{S}} x_i$. It is straightforward to show that $E_\xi\big(\hat{\mu}_y - \mu_y\big) = 0$. ◇

Example 5.4 Suppose that the survey population follows the simple linear regression model with an intercept (Model III). Under the assumed model, we have

$$E_\xi(y_i) = \beta_0 + \beta_1 x_i \quad \text{and} \quad V_\xi(y_i) = v_i\sigma^2\,, \quad i = 1, 2, \cdots, N\,.$$

Consider cases where $v_i = 1$. The ordinary least squares estimators of β_1 and β_0 are given by

$$\hat{\beta}_1 = \frac{s_{xy}}{s_x^2} \quad \text{and} \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \,,$$

where $\bar{y} = n^{-1} \sum_{i \in \mathbf{S}} y_i$, $\bar{x} = n^{-1} \sum_{i \in \mathbf{S}} x_i$, $s_{xy} = (n-1)^{-1} \sum_{i \in \mathbf{S}} (x_i - \bar{x})(y_i - \bar{y})$ and $s_x^2 = (n-1)^{-1} \sum_{i \in \mathbf{S}} (x_i - \bar{x})^2$. Under Model III, we have $E_\xi(\hat{\beta}_1) = \beta_1$ and $E_\xi(\hat{\beta}_0) = \beta_0$. Let

$$\hat{\mu}_y = \hat{\beta}_0 + \hat{\beta}_1 \mu_x = \bar{y} + \hat{\beta}_1 \big(\mu_x - \bar{x}\big) \,. \tag{5.2}$$

It can be shown that $E_\xi\big(\hat{\mu}_y - \mu_y\big) = 0$. The estimator $\hat{\mu}_y$ is an unbiased prediction estimator under Model III. ◇

5.1.2 Variance Estimation for Model-Based Prediction

Under the model-based prediction approach, estimation of $V_\xi\big(\hat{\mu}_y\big)$ is usually not of interest. This is due to the fact that μ_y is a random quantity and a prediction interval for μ_y often relies on

$$Z = \frac{\hat{\mu}_y - \mu_y}{\big\{V_\xi\big(\hat{\mu}_y - \mu_y\big)\big\}^{1/2}} \,.$$

When $\hat{\mu}_y$ is a model-unbiased prediction estimator for μ_y, we have $E_\xi(Z) = 0$ and $V_\xi(Z) = 1$. Using $N(0, 1)$ to approximate the distribution of Z becomes possible under certain conditions. Consequently, we are interested in estimating $V_\xi\big(\hat{\mu}_y - \mu_y\big)$ rather than $V_\xi\big(\hat{\mu}_y\big)$.

Example 5.1 (Continued) Under Model I, we have $V_\xi(y_i) = \sigma^2$ and $y_1, y_2, \cdots, y_N$ are independent, leading to

$$\begin{aligned} V_\xi\big(\hat{\mu}_y - \mu_y\big) &= V_\xi \left\{ \Big(\frac{1}{n} - \frac{1}{N}\Big) \sum_{i \in \mathbf{S}} y_i - \frac{1}{N} \sum_{i \notin \mathbf{S}} y_i \right\} \\ &= \Big(\frac{1}{n} - \frac{1}{N}\Big)^2 n\sigma^2 + \frac{N-n}{N^2}\sigma^2 \\ &= \Big(1 - \frac{n}{N}\Big)\frac{\sigma^2}{n} \,. \end{aligned}$$

This model-based variance has the same form as $V_p(\bar{y})$ under SRSWOR, except that σ^2 is the variance for the superpopulation. An unbiased variance estimator can be obtained if we replace σ^2 by the sample variance s_y^2. ◇

Example 5.2 (Continued) Under Model II with $v_i = 1$, we have $V_\xi(y_i) = \sigma^2$ and $y_1, y_2, \cdots, y_N$ are conditionally independent given all the x_i's. Let $c_i = x_i / \sum_{k \in \mathbf{S}} x_k^2$. We can re-write $\hat{\beta}$ as $\sum_{i \in \mathbf{S}} c_i y_i$, which further leads to $\hat{\mu}_y = \hat{\beta}\mu_x = \sum_{i \in \mathbf{S}} c_i \mu_x y_i$. Consequently,

$$\begin{aligned} V_\xi\left(\hat{\mu}_y - \mu_y\right) &= V_\xi\left\{\sum_{i \in \mathbf{S}} \left(c_i \mu_x - N^{-1}\right) y_i - \frac{1}{N} \sum_{i \notin \mathbf{S}} y_i\right\} \\ &= \left\{\sum_{i \in \mathbf{S}} \left(c_i \mu_x - N^{-1}\right)^2 + \frac{N-n}{N^2}\right\} \sigma^2 . \end{aligned}$$

Under the assumed regression model, the superpopulation variance σ^2 can be estimated by $\hat{\sigma}^2 = (n-1)^{-1} \sum_{i \in \mathbf{S}} \left(y_i - \hat{\beta} x_i\right)^2$. ◇

Brewer (1963b) was one of the first to use the model-based prediction approach in survey sampling in the context of Model II. Royall (1970) and his collaborators made a systematic study of model-based prediction methods for finite populations; see Valliant et al. (2000) for comprehensive coverage of the topic. As a general framework, however, the model-based prediction approach faces some serious challenges. The first major issue is that the validity of model-based inferences depends on the validity of the model. When the assumed model is questionable, all the inferences based on the model become questionable. Hansen et al. (1983) used Model II with $v_i = x_i$ (i.e. $V_\xi\left(\varepsilon_i\right) \propto x_i$) and showed that the model-based ratio estimator performed very poorly in large samples even if the deviations from the true model were small. The second major issue is that sampling design features for selecting the survey sample are typically ignored by model-based prediction methods. In addition to the assumed model for the survey population, they also assume that the survey sample follows the same superpopulation model. Model building, checking and diagnostics using survey sample data, which are related to the first issue of validity, have been shown to be a difficult task. There are also scenarios where the survey sample data do not follow the same model as the survey population, due to a so-called informative sampling design. Statistical modelling and analysis using survey data under complex survey designs, especially informative sampling designs, is an active research area pursued by survey statisticians.

5.2 Model-Assisted Estimation Methods

Inferences based on reliable models can be very efficient but results from misspecified models could be disastrous under the model-based prediction approach. Design-based inferences for survey sampling, on the other hand, impose no model assumptions, and the probability sampling design is chosen by the survey sampler based on the particular survey population under study. Typically, normal theory confidence intervals for large samples are asymptotically valid "*whatever the unknown properties of the population*" (Neyman 1934).

The strength of plausible models may be built into design-based inferences through a model-assisted approach. Under the model-assisted framework, a plausible model is used to motivate the construction of the estimator but the evaluation of the estimator takes place under both the model-based and the design-based framework. More specifically, an estimator is termed as *model-assisted* if

(i) It is model-unbiased prediction estimator under the assumed model.
(ii) It is approximately design-unbiased under the probability sampling design, irrespective of the model.

Part (i) can sometimes be weakened down to "approximately model-unbiased" and part (ii) may be replaced by "design-consistent". Design-consistency, i.e. the property that the estimator converges in probability to the parameter of interest under the sampling design, is a stronger concept than approximate design-unbiasedness. The former typically requires that the design-based variance goes to zero as the sample size gets large, assuming the finite parameter to be estimated is of order $O(1)$.

If the sample is selected by SRSWOR, the simplest model-assisted estimator of the population mean μ_y is $\hat{\mu}_y = \bar{y}$ under the common mean model (Model I). In this case we have $E_\xi\big(\hat{\mu}_y - \mu_y\big) = 0$ under Model I and $E_p\big(\hat{\mu}_y\big) = \mu_y$ regardless of the model.

Model-assisted estimators generally require that certain auxiliary population information is available. For the rest of this section, we assume that the sample $\mathbf{S}$ is selected by SRSWOR and the sample data are given by $\{(y_i, x_i), i \in \mathbf{S}\}$, where x is a single auxiliary variable. In addition, the population mean μ_x is available as the known auxiliary information.

5.2.1 The Simple Ratio Estimator

The simple ratio estimator for the population mean μ_y can be motivated through the regression model without an intercept (Model II). Under Model II, $\mu_y = \beta\mu_x + \bar{\varepsilon}_N$, where $\bar{\varepsilon}_N = N^{-1}\sum_{i=1}^N \varepsilon_i$ which is usually small for large N. This leads to $\mu_y \doteq \beta\mu_x$. Since μ_x is known, an estimator of μ_y can be obtained if we replace β by a suitable estimate. The simple ratio estimator of μ_y is defined as

$$\hat{\mu}_{yR} = \frac{\bar{y}}{\bar{x}}\,\mu_x\,,$$

where the subscript "R" denotes "Ratio". We have the following results on $\hat{\mu}_{yR}$.

Theorem 5.1 *Under the model-based prediction approach, the simple ratio estimator $\hat{\mu}_{yR}$ is model-unbiased for μ_y under Model II. Under the design-based framework with SRSWOR:*

(a) It is approximately design-unbiased for μ_y, i.e., $E_p(\hat{\mu}_{yR}) \doteq \mu_y$, regardless of the model.

(b) Its design-based variance is given by

$$V_p(\hat{\mu}_{yR}) \doteq \Big(1 - \frac{n}{N}\Big)\frac{1}{n}\big(\sigma_y^2 + R^2\sigma_x^2 - 2\,R\,\sigma_{xy}\big)\,,$$

where $R = \mu_y/\mu_x$, $\sigma_{xy} = (N-1)^{-1}\sum_{i=1}^N (x_i - \mu_x)(y_i - \mu_y)$ is the finite population covariance between x and y, and σ_x^2 and σ_y^2 are the respective finite population variance for x and y.

(c) A design-based variance estimator for $\hat{\mu}_{yR}$ is given by

$$v_p(\hat{\mu}_{yR}) = \Big(1 - \frac{n}{N}\Big)\frac{1}{n}\big(s_y^2 + \hat{R}^2 s_x^2 - 2\,\hat{R}\,s_{xy}\big)\,,$$

where $\hat{R} = \bar{y}/\bar{x}$, $s_{xy} = (n-1)^{-1}\sum_{i\in\mathbf{S}}(x_i - \bar{x})(y_i - \bar{y})$ is the sample covariance between x and y, and s_x^2 and s_y^2 are the respective sample variance for x and y.

Proof Under Model II, we have $E_\xi(y_i) = \beta x_i$, $E_\xi(\bar{y}) = \beta\bar{x}$ and $E_\xi(\hat{R}) = \beta$, where $E_\xi(\cdot)$ is conditional on x. It follows that

$$E_\xi\big(\hat{\mu}_{yR} - \mu_y\big) = \beta\mu_x - \beta\mu_x = 0\,.$$

The simple ratio estimator is model-unbiased under Model II.

Under the design-based framework, both $\bar{y}$ and $\bar{x}$ depend on the sampled set $\mathbf{S}$ and therefore both are random quantities. We need to deal with the nonlinear statistic $\hat{R} = \bar{y}/\bar{x}$, which can be viewed as an estimator for the population ratio $R = \mu_y/\mu_x$. Noting that $E_p(\hat{R}) \neq R$, we need suitable approximations to develop theoretical design-based properties for $\hat{R}$. Under SRSWOR, we have

$$\bar{y} - \mu_y = O_p\Big(\frac{1}{\sqrt{n}}\Big) \quad \text{and} \quad \bar{x} - \mu_x = O_p\Big(\frac{1}{\sqrt{n}}\Big)\,,$$

where $O_p(\cdot)$ and also $o_p(\cdot)$ in equation (5.3) below are the conventional notation for stochastic orders. For large n, the values of $\bar{y}$ and $\bar{x}$ remain in the neighbourhood of μ_y and μ_x, respectively. Using a Taylor series expansion of $\hat{R} = \bar{y}/\bar{x}$ at (μ_y, μ_x), we have

$$\hat{R} = \frac{\bar{y}}{\bar{x}} = \frac{\mu_y}{\mu_x} + \frac{1}{\mu_x}\big(\bar{y} - R\bar{x}\big) + o_p\Big(\frac{1}{\sqrt{n}}\Big) . \tag{5.3}$$

Note that $E_p(\bar{y}) = \mu_y$, $E_p(\bar{x}) = \mu_x$ and $E_p(\bar{y} - R\bar{x}) = 0$. This gives $E_p\big(\hat{R}\big) \doteq \mu_y/\mu_x$ and $E_p\big(\hat{\mu}_{yR}\big) \doteq \mu_y$. It also follows from (5.3) that

$$V_p\big(\hat{R}\big) \doteq \frac{1}{\mu_x^2} V_p(\bar{y} - R\bar{x}) = \frac{1}{\mu_x^2}\Big(1 - \frac{n}{N}\Big)\frac{1}{n}\big(\sigma_y^2 + R^2\sigma_x^2 - 2\,R\,\sigma_{xy}\big) , \tag{5.4}$$

which further leads to the design-based variance $V_p\big(\hat{\mu}_{yR}\big) = \mu_x^2 V_p\big(\hat{R}\big)$. The design-based variance estimator $v_p\big(\hat{\mu}_{yR}\big)$ is obtained by replacing unknown population quantities by suitable estimates. Note that the sample covariance s_{xy} is a design-unbiased estimator for σ_{xy} under SRSWOR (Problem 5.1). □

Equation (5.3) provides a linearized form for the nonlinear statistic $\hat{R}$. This technique is called *linearization* and is one of the major approaches to variance estimation in survey sampling. The variance formula given in (5.4) for $\hat{R}$ as well as the variance formula for $\hat{\mu}_{yR}$ given in the theorem are called linearization variance. They are valid when the sample size n is large.

When both μ_y and μ_x are unknown and $R = \mu_y/\mu_x$ is the parameter of interest, equation (5.4) leads to the following design-based variance estimator for $\hat{R}$ under SRSWOR:

$$v_p\big(\hat{R}\big) = \frac{1}{\bar{x}^2}\Big(1 - \frac{n}{N}\Big)\frac{1}{n}\big(s_y^2 + \hat{R}^2 s_x^2 - 2\,\hat{R}\,s_{xy}\big) . \tag{5.5}$$

Variance estimation for the ratio estimator generated a number of research papers on the finite sample performances of alternative variance estimators. The relationship $V_p\big(\hat{\mu}_{yR}\big) = \mu_x^2 V_p\big(\hat{R}\big)$ coupled with (5.5) leads to the following alternative variance estimator for the ratio estimator,

$$v_p\big(\hat{\mu}_{yR}\big) = \frac{\mu_x^2}{\bar{x}^2}\Big(1 - \frac{n}{N}\Big)\frac{1}{n}\big(s_y^2 + \hat{R}^2 s_x^2 - 2\,\hat{R}\,s_{xy}\big) . \tag{5.6}$$

Even though the factor $\mu_x^2/\bar{x}^2$ is close to be one under large samples, it has been shown by several authors that the variance estimator given in (5.6) is much more stable for samples with small or moderate sizes. See, for instance, Rao and Rao (1971), Royall and Cumberland (1978) and Wu (1982).

Example 5.5 (Single-Stage Cluster Sampling with Unknown Population Size) Let $\mathbf{S}_c$ be a set of k clusters selected from a population of K clusters using SRSWOR. The cluster totals T_i and the cluster sizes M_i are observed for $i \in \mathbf{S}_c$. The population size N is unknown. The estimator of μ_y, given by (3.10) in Sect. 3.3, can be written as

$$\hat{\mu}_y = \frac{k^{-1}\sum_{i\in \mathbf{S}_c} T_i}{k^{-1}\sum_{i\in \mathbf{S}_c} M_i} = \frac{\bar{T}}{\bar{M}} ,$$

where $\bar{T} = k^{-1}\sum_{i\in \mathbf{S}_c} T_i$ and $\bar{M} = k^{-1}\sum_{i\in \mathbf{S}_c} M_i$ are the two sample means computed using variables T_i and M_i respectively, with "sample size" k and "population size" K. A design-based variance estimator for $\hat{\mu}_y$ can be obtained from the general formula (5.5) where $\hat{R} = \bar{T}/\bar{M}$. ◇

The variance formula $V_p(\hat{\mu}_{yR})$ given in Theorem 5.1 can be used to provide a comparison between the two estimators $\hat{\mu}_{yR}$ and $\hat{\mu}_y = \bar{y}$ under SRSWOR. The sample mean $\bar{y}$ is design-unbiased. The simple ratio estimator $\hat{\mu}_{yR}$ is approximately design-unbiased and the bias is typically negligible when n is large. For comparison of design-based variances, we will have $V_p(\hat{\mu}_{yR}) < V_p(\bar{y})$ if $R^2\sigma_x^2 - 2R\sigma_{xy} < 0$. Consider scenarios where $\mu_y > 0$ and $\mu_x > 0$. The simple ratio estimator $\hat{\mu}_{yR}$ is more efficient than $\bar{y}$ if

$$\rho = \frac{\sigma_{xy}}{\sigma_x\sigma_y} > \frac{R}{2}\frac{\sigma_x}{\sigma_y} = \frac{1}{2}\frac{CV(x)}{CV(y)} ,$$

where ρ is the *population correlation coefficient*, and $CV(x) = \sigma_x/\mu_x$ and $CV(y) = \sigma_y/\mu_y$ are the *coefficients of variation* for the y and the x variables.

The simple ratio estimator $\hat{\mu}_{yR}$, which uses the known population mean μ_x from the auxiliary variable x, will have better performance than the sample mean $\bar{y}$ when y and x are highly and positively correlated and the sample size n is relatively large. The estimator $\hat{\mu}_{yR}$ could have non-negligible bias when n is small. The theoretical development on the estimation of the ratio $R = \mu_y/\mu_x$, however, has its own value. Practical applications of the theory include single-stage cluster sampling with unknown population size N as described in Example 5.5, and two-stage cluster sampling with unknown N, discussed in Sect. 3.4. It is also required for developing a variance estimator for the Hájek estimator to be discussed in Sect. 5.3 under a general unequal probability sampling design.

5.2.2 The Simple Regression Estimator

The simple linear regression model with an intercept (Model III) is commonly used in practice. We denote the model-based prediction estimator (5.2) discussed in Example 5.4 as

$$\hat{\mu}_{y\text{REG}} = \bar{y} + \hat{\beta}_1(\mu_x - \bar{x}) ,$$

where $\hat{\beta}_1 = s_{xy}/s_x^2$. This is called the simple regression estimator of μ_y. The subscript "REG" indicates "regression".

Under Model III, we have $E_\xi(\hat{\beta}_1) = \beta_1$, which further leads to $E_\xi\big(\hat{\mu}_{y\text{REG}} - \mu_y\big) = 0$. For design-based properties under SRSWOR, we have $E_p(s_x^2) = \sigma_x^2$ and $E_p(s_{xy}) = \sigma_{xy}$. The design-based expectation of the nonlinear statistic $\hat{\beta}_1 = s_{xy}/s_x^2$ cannot be computed exactly. Using a Taylor series expansion (i.e. linearization) for $\hat{\beta}_1$ similar to the one used in the development for $\hat{R}$, we have

$$\hat{\beta}_1 = B_1 + O_p\Big(\frac{1}{\sqrt{n}}\Big),$$

where $B_1 = \sigma_{xy}/\sigma_x^2$. It follows that $E_p\big(\hat{\beta}_1\big) \doteq B_1$. Let $B_0 = \mu_y - B_1\mu_x$. The B_0 and B_1 are the finite population regression coefficients, which are the ordinary least square estimators of β_0 and β_1 based on the entire survey population. The B_0 and B_1 are finite population parameters while β_0 and β_1 are the parameters of the superpopulation.

Theorem 5.2 *Under the model-based prediction approach, the simple regression estimator $\hat{\mu}_{y\text{REG}}$ is model-unbiased for μ_y under Model III. Under the design-based framework with SRSWOR:*

(a) It is approximately design-unbiased for μ_y, i.e., $E_p\big(\hat{\mu}_{y\text{REG}}\big) \doteq \mu_y$, regardless of the model.
(b) Its design-based variance is given by

$$V_p\big(\hat{\mu}_{y\text{REG}}\big) \doteq \Big(1 - \frac{n}{N}\Big)\frac{1}{n}\big(\sigma_y^2 + B_1^2\sigma_x^2 - 2\,B_1\,\sigma_{xy}\big).$$

(c) A design-based variance estimator for $\hat{\mu}_{y\text{REG}}$ is given by

$$v_p\big(\hat{\mu}_{y\text{REG}}\big) = \Big(1 - \frac{n}{N}\Big)\frac{1}{n}\big(s_y^2 + \hat{\beta}_1^2 s_x^2 - 2\,\hat{\beta}_1\,s_{xy}\big).$$

Proof Design-based properties of the estimator can be derived from the following linearization step:

$$\hat{\mu}_{y\text{REG}} = \bar{y} + B_1\big(\mu_x - \bar{x}\big) + o_p\Big(\frac{1}{\sqrt{n}}\Big).$$

The design-based variance is given by $V_p\big(\hat{\mu}_{y\text{REG}}\big) \doteq V_p\big(\bar{y} - B_1\bar{x}\big)$. Details of the proof are similar to the proof of Theorem 5.1 and left as an exercise. □

An important alternative expression for $V_p\big(\hat{\mu}_{y\text{REG}}\big)$ is based on the following relation between the finite population regression coefficient B_1 and the finite population correlation coefficient ρ:

$$B_1 = \frac{\sigma_{xy}}{\sigma_x^2} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \cdot \frac{\sigma_y}{\sigma_x} = \rho \frac{\sigma_y}{\sigma_x} \,.$$

It is straightforward to show that $V_p\big(\hat{\mu}_{y\text{REG}}\big)$ given in Theorem 5.2 can be re-written as

$$V_p\big(\hat{\mu}_{y\text{REG}}\big) \doteq \Big(1 - \frac{n}{N}\Big)\frac{1}{n}\sigma_y^2\big(1 - \rho^2\big) \,.$$

The design-based variance estimator $v_p\big(\hat{\mu}_{y\text{REG}}\big)$ can be re-written as

$$v_p\big(\hat{\mu}_{y\text{REG}}\big) \doteq \Big(1 - \frac{n}{N}\Big)\frac{1}{n}s_y^2\big(1 - r^2\big) \,,$$

where $r = s_{xy}/(s_x s_y)$ is the sample correlation coefficient. Note that we require the sample size n to be large to justify the approximations used in the linearization steps. An immediate consequence from the alternative expression for $V_p\big(\hat{\mu}_{y\text{REG}}\big)$ is that the simple regression estimator $\hat{\mu}_{y\text{REG}}$ is more efficient than $\hat{\mu}_y = \bar{y}$ when the sample size n is large. The stronger the correlation between y and x, the more gain of efficiency from using $\hat{\mu}_{y\text{REG}}$. In addition, it does not matter whether the correlation is positive or negative. The estimated regression coefficient $\hat{\beta}_1$ can automatically take care of either situation. It can also be shown that the simple regression estimator is generally more efficient than the simple ratio estimator (Problem 5.8).

Model-assisted estimators are motivated by a plausible working model. They are more efficient if the model holds but remain valid under the design-based framework even if the model is wrong. The framework for inferences is design-based, and variance estimators are only developed under the sampling design rather than the joint randomization involving both the model and the design, as demonstrated in Theorems 5.1 and 5.2. Chapter 7 contains further discussion on variance estimation under the joint randomization framework.

5.3 The Generalized Ratio Estimator

The simple ratio estimator $\hat{\mu}_{yR} = \big(\bar{y}/\bar{x}\big)\mu_x$ is motivated by Model II specified in Sect. 5.1. The survey sample data $\{(y_i, x_i), i \in \mathbf{S}\}$ are collected by SRSWOR and the population mean μ_x of the single x variable is assumed to be known. Under SRSWOR, $\hat{R} = \bar{y}/\bar{x}$ is a suitable design-based estimator of $R = \mu_y/\mu_x$. The simple ratio estimator can alternatively be motivated by replacing R in $\mu_y = R\mu_x$ by $\hat{R}$.

Under a general unequal probability sampling design with $\pi_i = P(i \in \mathbf{S})$ and $\pi_{ij} = P(i, j \in \mathbf{S})$, the HT estimators for μ_y and μ_x are given respectively by

$$\hat{\mu}_{y\text{HT}} = \frac{1}{N}\sum_{i\in\mathbf{S}}\frac{y_i}{\pi_i} = \frac{1}{N}\sum_{i\in\mathbf{S}} d_i y_i \quad \text{and} \quad \hat{\mu}_{x\text{HT}} = \frac{1}{N}\sum_{i\in\mathbf{S}}\frac{x_i}{\pi_i} = \frac{1}{N}\sum_{i\in\mathbf{S}} d_i x_i\,,$$

where $d_i = 1/\pi_i$ is the basic design weight. A suitable design-based estimator for $R = \mu_y/\mu_x$ is given by

$$\hat{R} = \frac{\hat{\mu}_{y\text{HT}}}{\hat{\mu}_{x\text{HT}}} = \frac{\sum_{i\in\mathbf{S}}\frac{y_i}{\pi_i}}{\sum_{i\in\mathbf{S}}\frac{x_i}{\pi_i}} = \frac{\sum_{i\in\mathbf{S}} d_i y_i}{\sum_{i\in\mathbf{S}} d_i x_i}\,.$$

The generalized ratio estimator of μ_y under the general sampling design is defined as

$$\hat{\mu}_{y\text{GR}} = \left(\frac{\hat{\mu}_{y\text{HT}}}{\hat{\mu}_{x\text{HT}}}\right)\mu_x = \left(\frac{\sum_{i\in\mathbf{S}} d_i y_i}{\sum_{i\in\mathbf{S}} d_i x_i}\right)\mu_x\,,$$

where the subscript GR indicates "Generalized Ratio".

We assume that the sampling design is non-informative in that the response variable y_i and the first order inclusion probability π_i are conditionally independent given x_i. It follows that under Model II,

$$E_\xi\big(\hat{\mu}_{y\text{HT}}\big) = \beta\hat{\mu}_{x\text{HT}} \quad \text{and} \quad E_\xi\big(\hat{\mu}_{y\text{GR}} - \mu_y\big) = 0\,.$$

In other words, the generalized ratio estimator $\hat{\mu}_{y\text{GR}}$ is a model-unbiased prediction estimator of μ_y under Model II.

Development of design-based properties for nonlinear statistics such as $\hat{\mu}_{y\text{GR}}$ requires certain regularity conditions on the sampling design as well as conditions on the finite population; see Fuller (2009, Sect. 1.3) for further discussion. We assume that the sampling design and the survey population (in terms of variables x and y) satisfy

$$\hat{\mu}_{y\text{HT}} - \mu_y = O_p\Big(\frac{1}{\sqrt{n}}\Big) \quad \text{and} \quad \hat{\mu}_{x\text{HT}} - \mu_x = O_p\Big(\frac{1}{\sqrt{n}}\Big)\,.$$

We have the following linearization of $\hat{R}$ by using a Taylor series expansion of the ratio of two HT estimators:

$$\hat{R} = \frac{\hat{\mu}_{y\text{HT}}}{\hat{\mu}_{x\text{HT}}} = R + \frac{1}{\mu_x}\Big(\hat{\mu}_{y\text{HT}} - R\,\hat{\mu}_{x\text{HT}}\Big) + o_p\Big(\frac{1}{\sqrt{n}}\Big)\,, \tag{5.7}$$

where $R = \mu_y/\mu_x$. It follows from (5.7) that

$$E_p\big(\hat{R}\big) \doteq R \quad \text{and} \quad V_p\big(\hat{R}\big) \doteq \frac{1}{\mu_x^2}\,V_p\big(\hat{\mu}_{e\text{HT}}\big)\,, \tag{5.8}$$

where $\hat{\mu}_{e\text{HT}} = N^{-1}\sum_{i\in\mathbf{S}} d_i e_i$ and $e_i = y_i - Rx_i$. The basic design-based properties of $\hat{R} = \hat{\mu}_{y\text{HT}}/\hat{\mu}_{x\text{HT}}$ are summarized by (5.8). It also leads to the design-based properties of the generalized ratio estimator of μ_y:

$$E_p\big(\hat{\mu}_{y\text{GR}}\big) \doteq \mu_y \quad \text{and} \quad V_p\big(\hat{\mu}_{y\text{GR}}\big) \doteq V_p\big(\hat{\mu}_{e\text{HT}}\big)\,.$$

It is now clear that the efficiency of the generalized ratio estimator depends on the variance of the HT estimator for $e_i = y_i - Rx_i$. Note that the variable z_i mimics the residual variable $\varepsilon_i = y_i - \beta x_i$ under Model II. If Model II adequately describes the relation between y and x for the survey population, it is expected that the residuals are less variable than the responses y_i themselves. Consequently, it is expected that $V_p\big(\hat{\mu}_{e\text{HT}}\big)$ will be smaller than $V_p\big(\hat{\mu}_{y\text{HT}}\big)$.

A design-based variance estimator for $\hat{\mu}_{y\text{GR}}$ can be constructed by using the variance estimator $v_p\big(\hat{\mu}_{e\text{HT}}\big)$ of the HT estimator $\hat{\mu}_{e\text{HT}} = N^{-1}\sum_{i\in\mathbf{S}} d_i e_i$ for the variable $e_i = y_i - Rx_i$, with R replaced by $\hat{R} = \hat{\mu}_{y\text{H}}/\hat{\mu}_{x\text{H}}$ and N substituted by $\hat{N} = \sum_{i\in\mathbf{S}} d_i$ in the final form. An alternative variance estimator using the same concept of (5.6) is to multiply $v_p\big(\hat{\mu}_{e\text{HT}}\big)$ by $\mu_x^2/\hat{\mu}_{x\text{H}}^2$.

Example 5.6 (Variance Estimation for the Hájek Estimator) The Hájek estimator of μ_y is computed as $\hat{\mu}_{y\text{H}} = \sum_{i\in\mathbf{S}} d_i y_i / \sum_{i\in\mathbf{S}} d_i$. It can alternatively be written as

$$\hat{\mu}_{y\text{H}} = \frac{N^{-1}\sum_{i\in\mathbf{S}} d_i y_i}{N^{-1}\sum_{i\in\mathbf{S}} d_i} = \frac{\hat{\mu}_{y\text{HT}}}{\hat{\mu}_{x\text{HT}}}\,,$$

where $x_i = 1$. This is a special case of (5.8) where $\mu_x = 1$ and $R = \mu_y$. The theoretical variance of the Hájek estimator is given by $V_p\big(\hat{\mu}_{y\text{H}}\big) \doteq V_p\big(\hat{\mu}_{e\text{HT}}\big)$ where $e_i = y_i - \mu_y$. A design-based variance estimator for $\hat{\mu}_{y\text{H}}$ is given by

$$v_p\big(\hat{\mu}_{y\text{H}}\big) = \frac{1}{\hat{N}^2}\sum_{i\in\mathbf{S}}\sum_{j\in\mathbf{S}} \frac{\pi_{ij} - \pi_i\pi_j}{\pi_{ij}}\,\frac{\hat{e}_i}{\pi_i}\,\frac{\hat{e}_j}{\pi_j}\,, \tag{5.9}$$

where $\hat{N} = \sum_{i\in\mathbf{S}} d_i$ and $\hat{e}_i = y_i - \hat{\mu}_{y\text{H}}$. Approximate variance estimators for the Horvitz-Thompson estimator which do not involve the second order inclusion probabilities can also be used for the Hájek estimator, which amounts to replacing y_i by $\hat{e}_i = y_i - \hat{\mu}_{y\text{H}}$ and N by $\hat{N}$. ◇

Example 5.7 (Confidence Intervals for the Distribution Function) The Hájek estimator for the finite population distribution function $F_y(t)$ at a fixed t is given by $\hat{F}_{y\text{H}}(t) = \{\sum_{i\in\mathbf{S}} d_i I(y_i \le t)\}/\{\sum_{i\in\mathbf{S}} d_i\}$, defined through the indicator variable $I(y_i \le t)$. The design-based variance estimator $v_p\big\{\hat{F}_{y\text{H}}(t)\big\}$ can be computed using the general variance formula described in Example 5.6 for the Hájek estimator using $e_i = I(y_i \le t) - F_y(t)$, with $F_y(t)$ replaced by $\hat{F}_{y\text{H}}(t)$ in the final form. A $(1-\alpha)$-level confidence interval for $F_y(t)$ based on normal approximations to the Wald statistic can be constructed as

$$\Big(\hat{F}_{y\mathrm{H}}(t) - Z_{\alpha/2}\big[v_p\{\hat{F}_{y\mathrm{H}}(t)\}\big]^{1/2}\,, \quad \hat{F}_{y\mathrm{H}}(t) + Z_{\alpha/2}\big[v_p\{\hat{F}_{y\mathrm{H}}(t)\}\big]^{1/2}\Big)\,, \tag{5.10}$$

where $Z_{\alpha/2}$ is the upper $\alpha/2$ quantile from the standard normal distribution. ◇

It turns out that variance estimation for the estimated population quantiles $\hat{t}_\gamma = \hat{F}_{y\mathrm{H}}^{-1}(\gamma)$, $\gamma \in (0,1)$ is a very difficult task. A useful technical tool in deriving asymptotic properties of quantile estimators is the Bahadur representation which relates quantile estimates directly to the estimated distribution function (Bahadur 1966). With complex survey data, however, Bahadur representations are difficult to establish even under very restrictive regularity conditions (Chen and Wu 2002). See Sect. 7.2 for further discussions.

Woodruff (1952) proposed a very simple method for constructing a $(1-\alpha)$-level confidence interval for the population quantile t_γ which does not involve variance estimation for $\hat{t}_\gamma$. Let $\gamma \in (0,1)$ and t_γ be the parameter of interest. The $(1-\alpha)$-level Woodruff confidence interval for t_γ is constructed as follows:

- Compute the point estimate $\hat{t}_\gamma = \hat{F}_{y\mathrm{H}}^{-1}(\gamma)$.
- Compute the estimated variance $v_p\{\hat{F}_{y\mathrm{H}}(t)\}$ for the distribution function at $t = \hat{t}_\gamma$.
- Compute the $(1-\alpha)$-level confidence interval (5.10) for the distribution function using $t = \hat{t}_\gamma$; denote the resulting interval as $(\hat{F}_L, \hat{F}_U)$.
- Compute the confidence interval for t_γ as $(\hat{t}_{\gamma L}\,,\ \hat{t}_{\gamma U})$, where

$$\hat{t}_{\gamma L} = \hat{F}_{y\mathrm{H}}^{-1}(\hat{F}_L) \quad \text{and} \quad \hat{t}_{\gamma U} = \hat{F}_{y\mathrm{H}}^{-1}(\hat{F}_U)\,.$$

It can be shown that $P\big(\hat{t}_{\gamma L} \le t_\gamma \le \hat{t}_{\gamma U}\big) \doteq 1-\alpha$ (Woodruff 1952; Lohr 2010, pp. 389–392). The Woodruff confidence interval is constructed through the inversion of the confidence interval for the distribution function. Sitter and Wu (2001) demonstrated a surprising property of the Woodruff confidence interval for quantiles t_γ: it performs well even when γ is small or large and sample size is moderate.

5.4 The Generalized Regression Estimator

Suppose that auxiliary information is available on k covariates $x_1, x_2, \cdots, x_k$. Let $(y_i, x_{i1}, x_{i2}, \cdots, x_{ik})$ be the values of $(y, x_1, x_2, \cdots, x_k)$ attached to unit i. Consider the following superpopulation model:

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \varepsilon_i\,, \quad i = 1, 2, \cdots, N\,, \tag{5.11}$$

where the ε_i's are independent with $E_\xi(\varepsilon_i) = 0$ and $V_\xi(\varepsilon_i) = v_i\sigma^2$. The v_i's are known constants depending on $\mathbf{x}_i = (x_{i1}, x_{i2}, \cdots, x_{ik})'$. Let $\mathbf{z}_i =$

$(1, x_{i1}, x_{i2}, \cdots, x_{ik})' = (1, \mathbf{x}_i')'$; let $\beta = (\beta_0, \beta_1, \cdots, \beta_k)'$. The parameters β as well as σ^2 are superpopulation parameters. Model (5.11) can be written as

$$y_i = \mathbf{z}_i'\beta + \varepsilon_i\,, \quad i = 1, 2, \cdots, N\,. \tag{5.12}$$

Let $\mu_{xl} = N^{-1}\sum_{i=1}^{N} x_{il}$ be the known population mean for the lth covariate, $l = 1, \cdots, k$. Let $\mu_{\mathbf{x}} = (\mu_{x1}, \cdots, \mu_{xk})'$ and $\mu_{\mathbf{z}} = (1, \mu_{\mathbf{x}}')'$.

The most popular model-assisted estimator of μ_y is the generalized regression (GREG) estimator (Cassel et al. 1976; Särndal 1980). A large body of literature on model-assisted inferences has been developed under the linear model (5.11). The book by Särndal et al. (1992) provides comprehensive coverage of generalized regression estimation under linear regression models with additional references on the topic. Fuller (2002) contains discussions on practical and theoretical aspects of regression estimation for survey samples.

5.4.1 GREG Under SRSWOR

Suppose that the survey sample $\mathbf{S}$ of size n is selected by SRSWOR. Let $\{(y_i, \mathbf{x}_i), i \in \mathbf{S}\}$ be the survey sample data. Let $\bar{y} = n^{-1}\sum_{i\in\mathbf{S}} y_i$ and $\bar{\mathbf{x}} = n^{-1}\sum_{i\in\mathbf{S}} \mathbf{x}_i$ be the sample means. Let $\hat{\mathbf{B}} = (\hat\beta_0, \hat\beta_1, \cdots, \hat\beta_k)'$ be the ordinary least squares estimators of β:

$$\hat{\mathbf{B}} = \Big(\sum_{i\in\mathbf{S}} \mathbf{z}_i\mathbf{z}_i'\Big)^{-1}\Big(\sum_{i\in\mathbf{S}} \mathbf{z}_i y_i\Big)\,.$$

Under model (5.12), we have $E_\xi\big(\hat{\mathbf{B}}\big) = \beta$ by the properties of the least squares estimators. Let

$$\hat\mu_{y\text{GREG}} = \hat{\mathbf{B}}'\mu_{\mathbf{z}}\,, \tag{5.13}$$

where the subscript GREG indicates "Generalized Regression". It can be shown that

$$E_\xi\big(\hat\mu_{y\text{GREG}} - \mu_y\big) = 0\,,$$

the GREG estimator is model-unbiased prediction estimator of μ_y under model (5.12).

We now examine the design-based properties of the GREG estimator. Under SRSWOR, the estimator $\hat{\mathbf{B}}$ is approximately design-unbiased for the so-called "*population regression coefficients*"

$$\mathbf{B} = \Big(\sum_{i=1}^{N} \mathbf{z}_i\mathbf{z}_i'\Big)^{-1}\Big(\sum_{i=1}^{N} \mathbf{z}_i y_i\Big)\,.$$

The population regression coefficients **B** are finite population parameters and can be viewed as the least squares estimators of β using data from the entire finite population. They are also called the "*census parameters*" defined through census estimating equations; see Chap. 7 for further details.

Let $\hat{\mathbf{B}} = (\hat{B}_0, \hat{\mathbf{B}}_1')'$ where $\hat{B}_0 = \hat{\beta}_0$ and $\hat{\mathbf{B}}_1 = (\hat{\beta}_1, \cdots, \hat{\beta}_k)'$. Noting that $\mathbf{z}_i = (1, \mathbf{x}_i')'$, it can be shown that

$$\hat{\mathbf{B}}_1 = \left\{ \sum_{i \in \mathbf{S}} \big(\mathbf{x}_i - \bar{\mathbf{x}}\big)\big(\mathbf{x}_i - \bar{\mathbf{x}}\big)' \right\}^{-1} \left\{ \sum_{i \in \mathbf{S}} \big(\mathbf{x}_i - \bar{\mathbf{x}}\big) y_i \right\}$$

and $\hat{B}_0 = \bar{y} - \hat{\mathbf{B}}_1'\bar{\mathbf{x}}$. Since $\mu_{\mathbf{z}} = (1, \mu_{\mathbf{x}}')'$, we can re-write $\hat{\mu}_{y\text{GREG}} = \hat{\mathbf{B}}'\mu_{\mathbf{z}}$ as

$$\hat{\mu}_{y\text{GREG}} = \bar{y} + \hat{\mathbf{B}}_1'\big(\mu_{\mathbf{x}} - \bar{\mathbf{x}}\big)\,. \tag{5.14}$$

The census parameters **B** can similarly be partitioned into $\mathbf{B} = (B_0, \mathbf{B}_1')'$, where $B_0 = \mu_y - \mathbf{B}_1'\mu_{\mathbf{x}}$ and

$$\mathbf{B}_1 = \left\{ \sum_{i=1}^{N} \big(\mathbf{x}_i - \mu_{\mathbf{x}}\big)\big(\mathbf{x}_i - \mu_{\mathbf{x}}\big)' \right\}^{-1} \left\{ \sum_{i=1}^{N} \big(\mathbf{x}_i - \mu_{\mathbf{x}}\big) y_i \right\}\,. \tag{5.15}$$

Under the survey design, the GREG estimator given by (5.14) is a nonlinear statistic involving the product of two random variables $\hat{\mathbf{B}}_1$ and $\bar{\mathbf{x}}$. The GREG estimator can be approximated by

$$\hat{\mu}_{y\text{GREG}} \doteq \bar{y} + \mathbf{B}_1'\big(\mu_{\mathbf{x}} - \bar{\mathbf{x}}\big)\,, \tag{5.16}$$

where the approximation amounts to dropping the term

$$\big(\hat{\mathbf{B}}_1 - \mathbf{B}_1\big)'\big(\mu_{\mathbf{x}} - \bar{\mathbf{x}}\big) = o_p\Big(\frac{1}{\sqrt{n}}\Big)\,.$$

It follows from (5.16) that

$$E_p\big(\hat{\mu}_{y\text{GREG}}\big) \doteq \mu_y \quad \text{and} \quad V_p\big(\hat{\mu}_{y\text{GREG}}\big) \doteq V_p\big(\bar{y} - \mathbf{B}_1'\bar{\mathbf{x}}\big) = V_p\big(\bar{e}\big)\,,$$

where $\bar{e} = n^{-1}\sum_{i \in \mathbf{S}} e_i$ and $e_i = y_i - B_0 - B_1 x_{i1} - \cdots - B_k x_{ik} = y_i - \mathbf{z}_i'\mathbf{B}$ is the finite population residual variable. If the linear regression model (5.11) is adequate for the relation between y and $\mathbf{x}$, the variance of the residual variables e is typically smaller than the variance of the response variable y. More formally, we can show that $V_p\big(\bar{y} - \mathbf{B}_1'\bar{\mathbf{x}}\big) \le V_p\big(\bar{y}\big)$ (Problem 5.9). Under SRSWOR,

$$V_p\big(\bar{e}\big) = \Big(1 - \frac{n}{N}\Big)\frac{\sigma_e^2}{n} \quad \text{and} \quad v_p\big(\bar{e}\big) = \Big(1 - \frac{n}{N}\Big)\frac{s_{\hat{e}}^2}{n},$$

where σ_e^2 is the finite population variance defined over the variable $e_i = y_i - \mathbf{z}_i'\mathbf{B}$ and $s_{\hat{e}}^2$ is the sample variance computed over $\hat{e}_i = y_i - \mathbf{z}_i'\hat{\mathbf{B}}$.

5.4.2 GREG Under a General Sampling Design

Under a general sampling design with first order inclusion probabilities π_i and basic design weights $d_i = 1/\pi_i$, an approximately design-unbiased estimator for the census parameter $\mathbf{B}$ is given by

$$\hat{\mathbf{B}} = \Big(\sum_{i\in\mathbf{S}} d_i \mathbf{z}_i \mathbf{z}_i'\Big)^{-1}\Big(\sum_{i\in\mathbf{S}} d_i \mathbf{z}_i y_i\Big). \tag{5.17}$$

This is the so-called *survey weighted estimator* for the regression coefficients; see Chap. 7 for further discussion. Under the same partition $\hat{\mathbf{B}} = (\hat{B}_0, \hat{\mathbf{B}}_1')'$, we have $\hat{B}_0 = \hat{\mu}_{y\mathrm{H}} - \hat{\mathbf{B}}_1'\hat{\mu}_{\mathbf{x}\mathrm{H}}$, where $\hat{\mu}_{y\mathrm{H}} = \hat{N}^{-1}\sum_{i\in\mathbf{S}} d_i y_i$, $\hat{\mu}_{\mathbf{x}\mathrm{H}} = \hat{N}^{-1}\sum_{i\in\mathbf{S}} d_i \mathbf{x}_i$, $\hat{N} = \sum_{i\in\mathbf{S}} d_i$ and

$$\hat{\mathbf{B}}_1 = \Big\{\sum_{i\in\mathbf{S}} d_i\big(\mathbf{x}_i - \hat{\mu}_{\mathbf{x}\mathrm{H}}\big)\big(\mathbf{x}_i - \hat{\mu}_{\mathbf{x}\mathrm{H}}\big)'\Big\}^{-1}\Big\{\sum_{i\in\mathbf{S}} d_i\big(\mathbf{x}_i - \hat{\mu}_{\mathbf{x}\mathrm{H}}\big)y_i\Big\}. \tag{5.18}$$

The GREG estimator of μ_y under a general unequal probability sampling design is defined as

$$\hat{\mu}_{y\mathrm{GREG}} = \hat{\mathbf{B}}'\mu_{\mathbf{z}} = \hat{\mu}_{y\mathrm{H}} + \hat{\mathbf{B}}_1'\big(\mu_{\mathbf{x}} - \hat{\mu}_{\mathbf{x}\mathrm{H}}\big). \tag{5.19}$$

It can be shown that $E_\xi\big(\hat{\mathbf{B}}\big) = \beta$ under model (5.12), which further leads to $E_\xi\big(\hat{\mu}_{y\mathrm{GREG}} - \mu_y\big) = 0$. The GREG estimator defined by (5.19) under a general sampling design is model-unbiased prediction estimator of μ_y under model (5.12).

The development of design-based properties of $\hat{\mu}_{y\mathrm{GREG}}$ requires the same regularity conditions on the sampling design as well as the finite population previously stated in Sect. 5.3. Under those conditions, we have

$$\hat{\mu}_{y\mathrm{GREG}} = \hat{\mu}_{y\mathrm{H}} + \mathbf{B}_1'\big(\mu_{\mathbf{x}} - \hat{\mu}_{\mathbf{x}\mathrm{H}}\big) + o_p\Big(\frac{1}{\sqrt{n}}\Big),$$

where $\mathbf{B}_1$ is defined by (5.15), which can be estimated by $\hat{\mathbf{B}}_1$ defined by (5.18) under the general sampling design. Noting that $B_0 = \mu_y - \mathbf{B}_1'\mu_{\mathbf{x}}$, we have $E_p\big(\hat{\mu}_{y\mathrm{GREG}}\big) \doteq \mu_y$ and

$$V_p\big(\hat{\mu}_{y\mathrm{GREG}}\big) = V_p\big(\hat{\mu}_{y\mathrm{GREG}} - \mu_y\big) \doteq V_p\big(\hat{\mu}_{y\mathrm{H}} - B_0 - \mathbf{B}_1'\hat{\mu}_{\mathbf{x}\mathrm{H}}\big) = V_p\big(\hat{\mu}_{e\mathrm{H}}\big)\,,$$

where $\hat{\mu}_{e\mathrm{H}} = \hat{N}^{-1}\sum_{i\in\mathbf{S}} d_i e_i$ is the Hájek estimator defined through $e_i = y_i - B_0 - \mathbf{B}_1'\mathbf{x}_i = y_i - \mathbf{B}'\mathbf{z}_i$, which is the residual variable for the finite population and satisfies $\sum_{i=1}^N e_i = 0$. A design-based variance estimator for $\hat{\mu}_{y\mathrm{GREG}}$ can be computed using the variance estimator given in (5.9) with $\hat{e}_i = y_i - \hat{\mathbf{B}}'\mathbf{z}_i$.

5.4.3 The Generalized Difference Estimator

The generalized regression estimator is closely related to the generalized difference estimator (Cassel et al. 1976; Wu and Sitter 2001a). Let $c_1, c_2, \cdots, c_N$ be a sequence of known constants. Let

$$\hat{\mu}_{y\mathrm{GD}} = \frac{1}{N}\sum_{i\in\mathbf{S}} d_i y_i - \frac{1}{N}\sum_{i\in\mathbf{S}} d_i c_i + \frac{1}{N}\sum_{i=1}^{N} c_i\,, \tag{5.20}$$

where the subscript GD indicates "Generalized Difference" to reflect the difference structure of the estimator. It is straightforward to show that $E_p\big(\hat{\mu}_{y\mathrm{GD}}\big) = \mu_y$, i.e., the GD estimator is design-unbiased for any sequence of known constants.

Consider the following superpopulation model ξ:

$$E_\xi\big(y_i\big) = \mu_i \quad \text{and} \quad V_\xi\big(y_i\big) = v_i\sigma^2\,, \quad i = 1, 2, \cdots, N\,.$$

The y_i's are also assumed to be independent. The choice of $c_i = \mu_i = E_\xi\big(y_i\big)$ in the GD estimator turns out to be optimal in that it minimizes the anticipated variance $E_\xi\big\{V_p\big(\hat{\mu}_{y\mathrm{GD}}\big)\big\}$ among design-unbiased estimators if the sampling design satisfies $\pi_i \propto \sqrt{v_i}$ (Cassel et al. 1976).

Unfortunately, the μ_i's are typically unavailable and the optimal GD estimator cannot be constructed directly without suitable auxiliary information. Under the linear regression model (5.12) with known information $\mathbf{x}_i$, $i \in \mathbf{S}$ and $\mu_{\mathbf{x}}$, we have $\mu_i = \mathbf{z}_i'\beta$, where $\mathbf{z}_i = (1, \mathbf{x}_i')'$. The optimal GD estimator obtained from (5.20) with $c_i = \mu_i = \mathbf{z}_i'\beta$ is given by

$$\hat{\mu}_{y\mathrm{GD}} = \hat{\mu}_{y\mathrm{HT}} + \beta'\Big(\mu_{\mathbf{z}} - \hat{\mu}_{\mathbf{z}\mathrm{HT}}\Big)\,. \tag{5.21}$$

If we further replace the unknown superpopulation parameter β by the design-based estimator $\hat{\mathbf{B}}$ given by (5.17) and substitute N by $\hat{N}$ in $\hat{\mu}_{y\mathrm{HT}}$ and $\hat{\mu}_{\mathbf{z}\mathrm{HT}}$, the GD estimator from (5.21) becomes identical to the GREG estimator given in (5.19). In other words, the GREG estimator can be obtained as the optimal GD estimator under a linear regression model.

When the linear regression model (5.12) does not seem to be plausible, some commonly used nonparametric techniques such as local linear smoothing or spline fitting may be used to obtain fitted values for $\mu_i = E_\xi\big(y_i \mid \mathbf{x}_i\big) = \mu(\mathbf{x}_i)$ to construct "near-optimal" GD estimators. For linear models where $\mu(\mathbf{x}_i) = \mathbf{z}_i'\beta$, the GD estimator of μ_y given by (5.21) only requires $\mu_{\mathbf{x}}$ as the known auxiliary information. For nonparametric techniques, it often requires that complete auxiliary information $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N$ be available, due to the use of the term $N^{-1}\sum_{i=1}^N \mu_i = N^{-1}\sum_{i=1}^N \mu(\mathbf{x}_i)$ in the GD estimator; see Chap. 6 for further discussions on calibration weighting and estimation.

5.5 Problems

5.1 (Estimation of Population Covariance) Show that under SRSWOR the sample covariance s_{xy}^2 is a design-unbiased estimator for the population covariance σ_{xy}^2, where

$$s_{xy}^2 = \frac{1}{n-1}\sum_{i\in\mathbf{S}}\big(x_i - \bar{x}\big)\big(y_i - \bar{y}\big) \quad \text{and} \quad \sigma_{xy}^2 = \frac{1}{N-1}\sum_{i=1}^{N}\big(x_i - \mu_x\big)\big(y_i - \mu_y\big)\,.$$

Hint: Consider σ_z^2 and s_z^2 where $z_i = x_i + y_i$.

5.2 (Model-Based Prediction Under Model II) Suppose that the finite population follows the model (ξ): $y_i = \beta x_i + \varepsilon_i$, $i = 1, 2, \cdots, N$, where $x_i > 0$ and the ε_i's are independent random variables with $E_\xi(\varepsilon_i) = 0$ and $V_\xi(\varepsilon_i) = x_i\sigma^2$. The finite population mean μ_x for the auxiliary variable x is known. Let $\{(y_i, x_i), i \in \mathbf{S}\}$ be the sample data.

(a) Show that the weighted least square estimator of β under the model ξ is given by $\hat{\beta} = \bar{y}/\bar{x}$.
(b) Show that $\hat{\mu}_y = \hat{\beta}\mu_x$ is a model-unbiased prediction estimator for the finite population mean μ_y.
(c) Find the model-based variance $V_\xi(\hat{\mu}_y - \mu_y)$.
(d) Show that the alternative model-based prediction estimator for μ_y presented in Example 5.3 is identical to $\hat{\mu}_y = \hat{\beta}\mu_x$ with $\hat{\beta} = \bar{y}/\bar{x}$.

5.3 (Comparison of Two Prediction Estimators Under Model II) Consider the simple linear regression model without an intercept (Model II) with $v_i = 1$ (i.e., $V_\xi(\varepsilon_i) = \sigma^2$). The least square estimator $\hat{\beta}$ of β is given by (5.1) in Example 5.2. The predicted value for y_i is given by $\hat{y}_i = \hat{\beta}x_i$. Consider the two prediction estimators introduced in Examples 5.2 and 5.3:

$$\hat{\mu}_{y1} = \frac{1}{N}\sum_{i=1}^{N}\hat{y}_i = \hat{\beta}\mu_x \quad \text{and} \quad \hat{\mu}_{y2} = \frac{1}{N}\Big(\sum_{i\in\mathbf{S}} y_i + \sum_{i\notin\mathbf{S}}\hat{y}_i\Big).$$

Show that $V_\xi\big(\hat{\mu}_{y1} - \mu_y\big) \geq V_\xi\big(\hat{\mu}_{y2} - \mu_y\big)$, and the equality holds if and only if $s_x^2 = 0$.

5.4 (Model-Based Variance Estimation for the Regression Estimator) Consider Model III and the model-based prediction estimator $\hat{\mu}_y = \bar{y} + \hat{\beta}_1\big(\mu_x - \bar{x}\big)$ given in Example 5.4.

(a) Derive the theoretical variance $V_\xi\big(\hat{\mu}_y - \mu_y\big)$.
(b) Construct a variance estimator for $\hat{\mu}_y$ based on the result in (a).

5.5 (Model-Based Prediction Under Model II) Let $\hat{\mu}_y = \hat{\beta}\mu_x$ where $\hat{\beta} = \sum_{i\in\mathbf{S}} x_i y_i / \sum_{i\in\mathbf{S}} x_i^2$. Suppose that the sample $\mathbf{S}$ is selected by SRSWOR.

(a) Find the approximate design-based expectation of $\hat{\beta}$, assuming n is large.
(b) Argue that $\hat{\mu}_y$ is not a model-assisted estimator for μ_y.

5.6 (Balanced Sampling) Suppose that the complete auxiliary information $x_1, x_2, \cdots, x_N$ is available at the sample selection stage. We select a sample $\mathbf{S}$ of size n by SRSWOR. The sample is called *balanced* on x if $\bar{x} = \mu_x$, where $\bar{x}$ is the sample mean. A balanced sample can be selected through a rejective procedure under which we reject the selected sample if $|\bar{x} - \mu_x| > \delta$ for a pre-specified small tolerance $\delta > 0$. The rejection process continues until a sample satisfying $|\bar{x} - \mu_x| \leq \delta$ is selected.

(a) Show that under balanced sampling $\hat{\mu}_y = \bar{y}$ is an approximately model-unbiased prediction estimator for μ_y under Model II.
(b) Show that under balanced sampling $\hat{\mu}_y = \bar{y}$ is also an approximately model-unbiased prediction estimator for μ_y under Model III.

5.7 (Optimal Regression Estimator Under SRSWOR) Suppose that the finite population follows the model $y_i = \alpha + \beta x_i + \varepsilon_i$, $i = 1, 2, \cdots, N$, where the ε_i's are iid random variates with $E_\xi(\varepsilon_i) = 0$ and $V_\xi(\varepsilon_i) = \sigma^2$. We assume that β is a known constant. Let $\{(y_i, x_i), i \in \mathbf{S}\}$ be the sample data obtained from SRSWOR. The population mean μ_x is also known. Let $\hat{\mu}^*_{y\text{REG}} = \bar{y} + \beta(\mu_x - \bar{x})$.

(a) Show that $\hat{\mu}^*_{y\text{REG}}$ is both design-unbiased and model-unbiased for μ_y.
(b) Derive the design-based variance $V_p\big(\hat{\mu}^*_{y\text{REG}}\big)$ and identify conditions under which we have $V_p\big(\hat{\mu}^*_{y\text{REG}}\big) < V_p(\bar{y})$.
(c) Let $\hat{\mu}_c = \bar{y} + c(\mu_x - \bar{x})$, where c is some constant. Show that $c = B_1 = \sigma_{xy}/\sigma_x^2$ minimizes the design-based variance $V_p\big(\hat{\mu}_c\big)$ among estimators in the class $\{\hat{\mu}_c,\ -\infty < c < \infty\}$.
(d) Show that the choice of $c = \beta$ minimizes the anticipated design variance $E_\xi V_p\big(\hat{\mu}_c\big)$ under the model among estimators in the class $\{\hat{\mu}_c,\ -\infty < c < \infty\}$.

5.8 (Comparison Between Ratio and Regression Estimators) Suppose that the sample data $\{(y_i, x_i), i \in \mathbf{S}\}$ are collected by SRSWOR. Let

$$\hat{\mu}_{yR} = \frac{\bar{y}}{\bar{x}}\,\mu_x \quad \text{and} \quad \hat{\mu}_{y\text{REG}} = \bar{y} + \hat{\beta}_1(\mu_x - \bar{x})$$

be respectively the ratio estimator and the regression estimator of μ_y, where $\hat{\beta}_1 = s_{xy}/s_x^2$. Let $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$. We assume that the sample size n is large so the linearization variance for each of $\hat{\mu}_{yR}$ and $\hat{\mu}_{y\text{REG}}$ can be treated as the true variance.

(a) Show that $V_p\big(\hat{\mu}_{y\text{REG}}\big) \le V_p\big(\hat{\mu}_{yR}\big)$, and that equality holds if and only if $R = B_1$, where $R = \mu_y/\mu_x$ and $B_1 = \sigma_{xy}/\sigma_x^2$.
(b) Note that $R = B_1$ is equivalent to $B_0 = \mu_y - B_1\mu_x = 0$. Give an interpretation of the result in (a) from a superpopulation point of view.

5.9 (Optimality of GREG Under SRSWOR)

(a) Let $\mathbf{C}$ be a $k \times 1$ constant vector. Let $\hat{\mu}_{y\text{GREG}}(\mathbf{C}) = \bar{y} + \mathbf{C}'\big(\mu_{\mathbf{x}} - \bar{\mathbf{x}}\big)$. Show that $V_p\big(\hat{\mu}_{y\text{GREG}}(\mathbf{C})\big)$ is minimized when $\mathbf{C} = \mathbf{B}_1$, where $\mathbf{B}_1$ is given by (5.15).
(b) Use the result from (a) to argue that $V_p\big(\bar{y} - \mathbf{B}_1'\bar{\mathbf{x}}\big) \le V_p\big(\bar{y}\big)$.

5.10 (Separate Ratio Estimator Under Stratified Simple Random Sampling) Let $\{(y_{hi}, x_{hi}), i \in \mathbf{S}_h,\ h = 1, 2, \cdots, H\}$ be the survey data from stratified simple random sampling. Let μ_{xh}, $h = 1, 2, \cdots, H$ be the known stratum population means for the single auxiliary variable x. The separate ratio (SR) estimator of μ_y is defined as

$$\hat{\mu}_{y\text{SR}} = \sum_{h=1}^{H} W_h\,\frac{\bar{y}_h}{\bar{x}_h}\,\mu_{xh}\,,$$

where $\bar{y}_h$ and $\bar{x}_h$ are respectively the stratum sample means for y and x. See Sect. 3.1 for notational details.

(a) Describe a superpopulation model under which $\hat{\mu}_{y\text{SR}}$ is a model-unbiased prediction estimator of μ_y.
(b) Show that $E_p\big(\hat{\mu}_{y\text{SR}}\big) \doteq \mu_y$.
(c) Derive an approximate formula for $V_p\big(\hat{\mu}_{y\text{SR}}\big)$.
(d) Derive a design-based variance estimator based on (c).
(e) What conditions have you assumed in deriving the results in (b) and (c)?

5.11 (Combined Ratio Estimator Under Stratified Simple Random Sampling) Let $\{(y_{hi}, x_{hi}), i \in \mathbf{S}_h,\ h = 1, 2, \cdots, H\}$ be the survey data from stratified simple random sampling. Let μ_x be the known population mean for the single auxiliary variable x. The combined ratio (CR) estimator of μ_y is defined as

$$\hat{\mu}_{y\text{CR}} = \frac{\bar{y}_{st}}{\bar{x}_{st}}\mu_x \, ,$$

where $\bar{y}_{st} = \sum_{h=1}^{H} W_h \bar{y}_h$ and $\bar{x}_{st} = \sum_{h=1}^{H} W_h \bar{x}_h$.

(a) Describe a superpopulation model under which $\hat{\mu}_{y\text{CR}}$ is a model-unbiased prediction estimator of μ_y.
(b) Show that $E_p\big(\hat{\mu}_{y\text{CR}}\big) \doteq \mu_y$.
(c) Derive an approximate formula for $V_p\big(\hat{\mu}_{y\text{CR}}\big)$.
(d) Derive a design-based variance estimator based on (c).
(e) What conditions have you assumed in deriving the results in (b) and (c)?

5.12 (Separate Regression Estimator Under Stratified Simple Random Sampling) Let $\{(y_{hi}, x_{hi}), i \in \mathbf{S}_h, \;\; h = 1, 2, \cdots, H\}$ be the survey data from stratified simple random sampling. Let $\mu_{xh}, \; h = 1, 2, \cdots, H$ be the known stratum population means for the single auxiliary variable x. The separate regression ($SREG$) estimator of μ_y is defined as

$$\hat{\mu}_{y\text{SREG}} = \sum_{h=1}^{H} W_h \left\{\bar{y}_h + \hat{\beta}_{1h}\big(\mu_{xh} - \bar{x}_h\big)\right\} ,$$

where $\hat{\beta}_{1h} = s_{xyh}/s_{xh}^2$ is computed using data from stratum h.

(a) Describe a superpopulation model under which $\hat{\mu}_{y\text{SREG}}$ is model-unbiased prediction estimator of μ_y.
(b) Show that $E_p\big(\hat{\mu}_{y\text{SREG}}\big) \doteq \mu_y$.
(c) Derive an approximate formula for $V_p\big(\hat{\mu}_{y\text{SREG}}\big)$.
(d) Derive a design-based variance estimator based on (c).
(e) What conditions have you assumed in deriving the results in (b) and (c)?

5.13 (Ratio Estimator Under Two-Phase Sampling) The ratio estimator requires that the population mean μ_x of the x variable be known. When μ_x is not available but information on x can be collected with relatively low cost, a two-phase sampling strategy can be used: Select a large first phase sample $\mathbf{S}_1$ of size n_1 from the population using SRSWOR, and collect x_i for $i \in \mathbf{S}_1$; select a second phase sample $\mathbf{S}_2 \subset \mathbf{S}_1$ of size n_2 ($n_2 < n_1$) from units in the first phase sample, and collect y_i for $i \in \mathbf{S}_2$. Let

$$\hat{\mu}_{yR2} = \frac{\bar{y}_{[2]}}{\bar{x}_{[2]}}\bar{x}_{[1]} \, ,$$

where $\bar{x}_{[1]} = n_1^{-1}\sum_{i\in\mathbf{S}_1} x_i$, $\bar{x}_{[2]} = n_2^{-1}\sum_{i\in\mathbf{S}_2} x_i$ and $\bar{y}_{[2]} = n_2^{-1}\sum_{i\in\mathbf{S}_2} y_i$.

(a) Show that $E_p\big(\hat{\mu}_{yR2}\big) \doteq \mu_y$.
(b) Derive an approximate formula for $V_p\big(\hat{\mu}_{yR2}\big)$.
(c) Derive a design-based variance estimator using the two-phase sample data.

5.14 (Regression Estimator Under Two-Phase Sampling) Consider the same setting as for Problem 5.13. Let

$$\hat{\mu}_{y\text{REG2}} = \bar{y}_{[2]} + \hat{B}\big(\bar{x}_{[1]} - \bar{x}_{[2]}\big)\,,$$

where $\hat{B} = s_{xy2}/s_{x2}^2$, $s_{xy2} = (n_2-1)^{-1}\sum_{i\in\mathbf{S}_2}(y_i - \bar{y}_{[2]})(x_i - \bar{x}_{[2]})$ and $s_{x2}^2 = (n_2-1)^{-1}\sum_{i\in\mathbf{S}_2}(x_i - \bar{x}_{[2]})^2$.

(a) Show that $E_p\big(\hat{\mu}_{y\text{REG2}}\big) \doteq \mu_y$.
(b) Derive an approximate formula for $V_p\big(\hat{\mu}_{y\text{REG2}}\big)$.
(c) Derive a design-based variance estimator using the two-phase sample data.

5.15 (Variance Estimation for the Ratio Estimator Under SRSWOR) Consider three variance estimators for the ratio estimator $\hat{\mu}_{yR} = (\bar{y}/\bar{x})\mu_x$ in the form of

$$v_p^{(k)}\big(\hat{\mu}_{yR}\big) = \Big(\frac{\mu_x}{\bar{x}}\Big)^k\Big(1 - \frac{n}{N}\Big)\frac{1}{n}\big(s_y^2 + \hat{R}^2 s_x^2 - 2\,\hat{R}\,s_{xy}\big)\,, \qquad k = 0, 1, 2\,.$$

The variance estimator with $k = 0$ corresponds to the one given in Part (c) of Theorem 5.1. The variance estimator with $k = 2$ is given in (5.6). Design and conduct a simulation study to evaluate the performance of the three variance estimators.

(a) The finite population of size $N = 5000$ is generated from the model $y_i = 1 + x_i + \varepsilon_i$, where $x_i \sim \text{Exp}(1)$ and ε_i are iid $N(0,\sigma^2)$. Consider two values of σ^2 such that the finite population correlation coefficient between y and x is respectively 0.3 and 0.7.
(b) Consider sample sizes $n = 25$, 50 and 100.
(c) Evaluate the performance of a variance estimator by simulated bias and mean squared error. **Note**: The true variance of the ratio estimator under a given setting can be simulated through a set of independent samples.

Part II
Advanced Topics on Analysis of Probability Survey Samples

Chapter 6
Calibration Weighting and Estimation

One of the important features of survey data is the availability of auxiliary information from the survey population. Auxiliary population information may be available from census enumerations or administrative records. It could also be obtained from unusual sources such as satellite images for natural resource inventory surveys or a previous survey from the same population. Auxiliary information can be utilized at the survey design stage to facilitate frame construction, stratification, clustering and unequal probability selection of units. Calibration methods address issues related to the use of auxiliary information at the estimation stage of survey data analysis.

The original idea of calibration estimation in the presence of auxiliary information goes back to the raking ratio estimation method of Deming and Stephan (1940), where the objective was to estimate the cell proportions in a two-way contingency table with known marginal population totals and the survey sample is taken by simple random sampling; see Sect. 6.4 for further detail. Huang and Fuller (1978) were the first ones to introduce calibration weights using the name regression weights and then went further to consider range restricted regression weights; see Sect. 6.2.4 for further discussion. Calibration estimation and weighting methods for complex survey data were first formally used for household surveys. A considerable amount of research was conducted in the United States and Canada during the 1980s on weighting for household surveys. See, for instance, Alexander (1987), Copeland et al. (1987), Lemaitre and Dufour (1987), Luery (1980, 1986), Roman (1982), and Zieschang (1986, 1990). The paper by Deville and Särndal (1992) played a significant role in formalizing and popularizing the general concept and techniques of calibration weighting and estimation. Many different aspects of calibration techniques have subsequently been investigated and new directions have emerged.

There are two primary reasons why calibration weighting and estimation methods have become standard practice for complex surveys. The first is *efficiency*. Calibration over a set of carefully chosen auxiliary variables has proven to be an effective way of using known auxiliary population information. The calibrated

C. Wu, M. E. Thompson, *Sampling Theory and Practice*, ICSA Book
Series in Statistics, https://doi.org/10.1007/978-3-030-44246-0_6

weights can be applied to arbitrary study variables, and the gain of efficiency in estimation can be substantial if the study variable is highly correlated with the set of auxiliary variables. More importantly, design-based inferences remain valid even if the study variable is uncorrelated with the calibration variables. The second is *internal consistency*, i.e., to be consistent with known population aggregates. Large scale survey data are often analyzed for different aims and objectives and by different researchers. It is deemed important for survey organizations and statistical agencies to maintain consistency among survey results. In particular, estimates of certain population quantities obtained from the survey sample should agree with what are known from census or other external sources. The calibrated weights ensure that such internal consistency becomes an intrinsic part of any analysis.

This chapter presents calibration estimators and calibration weighting methods. Standard calibration methods are presented in Sects. 6.1 and 6.2 where population totals of the auxiliary variables are known. The model-calibration methods presented in Sect. 6.3 require so-called "complete auxiliary information" and focus primarily on efficiencies of estimators. The classic raking ratio estimation method and its extensions are discussed in Sect. 6.4. Connections between the calibration techniques and the empirical likelihood methods are discussed in Chap. 8.

6.1 Calibration Estimators

Suppose that auxiliary information is available in the form of known population totals $T_{\mathbf{x}}$ for the vector of k auxiliary variables $\mathbf{x} = (x_1, x_2, \ldots, x_k)'$. Let $\{(y_i, \mathbf{x}_i), i \in \mathbf{S}\}$ be the survey sample data set, where y_i and $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ik})'$ are the values of the response variable y and auxiliary variables $\mathbf{x}$, respectively, for unit i. The set of sampled units, $\mathbf{S}$, is selected by a probability sampling design with the first and second order inclusion probabilities π_i and π_{ij}. The survey data set may also include other response variables or covariates but they are omitted here from the notation without affecting the discussion on calibration estimators.

6.1.1 Definition of Calibration Estimators

The Horvitz-Thompson estimators of the population totals $T_{\mathbf{x}}$ are given by $\hat{T}_{\mathbf{x}\mathrm{HT}} = \sum_{i \in \mathbf{S}} d_i \mathbf{x}_i$. The estimated totals $\hat{T}_{\mathbf{x}\mathrm{HT}}$ usually do not equal the population quantities $T_{\mathbf{x}}$, leading to inconsistencies between survey-based estimates and the known auxiliary population information. For an arbitrary response variable y, a calibration estimator of the population total T_y is defined as

$$\hat{T}_{y\mathrm{C}} = \sum_{i \in \mathbf{S}} w_i y_i \,,$$

where the weights w_i are calibrated over the $\mathbf{x}$ variables. The weights w_i are calibrated if they satisfy two requirements. First, they are forced to satisfy the following equations:

$$\sum_{i \in \mathbf{S}} w_i \mathbf{x}_i = T_{\mathbf{x}} \,. \tag{6.1}$$

Equations (6.1) are called the *calibration equations* or *benchmark constraints*, and are the first requirement of the definition for calibration estimators. Note that the left hand side of equations (6.1) are the estimators $\hat{T}_{\mathbf{x}\mathrm{C}} = \sum_{i \in \mathbf{S}} w_i \mathbf{x}_i$, which thus match exactly the known population totals $T_{\mathbf{x}}$ on the right hand side of the equations. Unfortunately, the benchmark constraints (6.1) leave too much room for feasible solutions for w_i, and an arbitrarily chosen solution does not lead to an estimator $\hat{T}_{y\mathrm{C}}$ with desirable properties.

The second requirement of the definition is to make the w_i as close to the basic design weights d_i as possible. This is desirable due to the prominent role of the Horvitz-Thompson estimator in design-based inference. For the given sample $\mathbf{S}$ with sample size n, the closeness between the calibration weights $\mathbf{w} = (w_1, \ldots, w_n)'$ and the basic design weights $\mathbf{d} = (d_1, \ldots, d_n)'$ can be measured through a distance

$$D(\mathbf{w}, \mathbf{d}) = \sum_{i \in \mathbf{S}} G_i(w_i, d_i) \,,$$

where $G_i(w_i, d_i)$ depends on w_i and d_i and possibly other known quantities. The distance $D(\mathbf{w}, \mathbf{d})$ measures the amount of discrepancy between $\mathbf{w}$ and $\mathbf{d}$ and does not need to be a true Euclidean distance. For instance, the distance measure may not be symmetric in $\mathbf{w}$ and $\mathbf{d}$, and the usual triangle inequality is not relevant since there is no "third point" of interest other than $\mathbf{w}$ and $\mathbf{d}$. A minimum requirement for the distance measure is that $G_i(w_i, d_i) \geq 0$ for all i and $G_i(w_i, d_i) = 0$ if $w_i = d_i$.

The calibration weights w_i, $i \in \mathbf{S}$ are obtained by minimizing the distance $D(\mathbf{w}, \mathbf{d})$ subject to the benchmark constraints (6.1). Using the Lagrange multiplier method for the constrained minimization problem, it can be shown that w_i satisfies

$$g_i(w_i, d_i) - \mathbf{x}_i' \boldsymbol{\lambda} = 0 \,, \tag{6.2}$$

where $g_i(w, d) = \partial G_i(w, d) / \partial w$ and the $\boldsymbol{\lambda}$ denotes a vector of Lagrange multipliers. If (6.2) can be solved to obtain $w_i = h(d_i, \mathbf{x}_i' \lambda)$, then λ can be determined by the equations (6.1) through the revised form

$$\sum_{i \in \mathbf{S}} h(d_i, \mathbf{x}_i' \lambda) \, \mathbf{x}_i = T_{\mathbf{x}} \,. \tag{6.3}$$

For the chi-square distance to be discussed in the next section, the solution to (6.3) has a closed form. For other choices of $D(\mathbf{w}, \mathbf{d})$, finding the solution λ to (6.3) requires iterative procedures; see further details in Sect. 6.2.

The final set of calibration weights w_i depends on the set of auxiliary variables used in the calibration as well as the choice for the distance measure. There is no guarantee that a solution exists with an arbitrary distance measure. The existence of a solution typically depends on the sample configuration of the **x** variables; see Sect. 6.2 for further discussion. The uniqueness of the solution, if a solution exists, can be guaranteed under certain assumptions on $G_i(w_i, d_i)$ (Deville and Särndal 1992).

If the population size N is known, it can be incorporated into calibration estimation by imposing the constraint

$$\sum_{i \in \mathbf{S}} w_i = N . \tag{6.4}$$

From a computational point of view, it is more convenient to combine (6.1) and (6.4) into a single set of constraints

$$\sum_{i \in \mathbf{S}} w_i \mathbf{z}_i = T_{\mathbf{z}} , \tag{6.5}$$

where $\mathbf{z}_i = (1, \mathbf{x}_i')'$ and $T_{\mathbf{z}} = (N, T_{\mathbf{x}}')'$.

6.1.2 Calibration Estimators Under the Chi-Square Distance

The simplest form of the distance measure $D(\mathbf{w}, \mathbf{d})$ is the so-called chi-square distance given by

$$\chi^2(\mathbf{w}, \mathbf{d}) = \sum_{i \in \mathbf{S}} \frac{(w_i - d_i)^2}{q_i d_i} ,$$

where the q_i are pre-specified constants, independent of the sampling design, and are sometimes called the q-factor. Certain choices of q_i might lead to simplified forms of the estimator (Problem 6.2) or reduce the variance of the calibration estimator (Problem 6.3) but $q_i = 1$ are commonly used in practice. It is shown in Problem 6.1 that minimizing $\chi^2(\mathbf{w}, \mathbf{d})$ subject to constraints (6.1) leads to

$$w_i = d_i \left\{ 1 + \left(T_{\mathbf{x}} - \hat{T}_{\mathbf{x}\mathrm{HT}} \right)' \left(\sum_{j \in \mathbf{S}} d_j q_j \mathbf{x}_j \mathbf{x}_j' \right)^{-1} \mathbf{x}_i q_i \right\} , \tag{6.6}$$

where $\hat{T}_{\mathbf{x}\mathrm{HT}} = \sum_{i \in \mathbf{S}} d_i \mathbf{x}_i$. The resulting calibration estimator of T_y is given by

$$\hat{T}_{y\text{C}} = \sum_{i \in \mathbf{S}} w_i y_i = \hat{T}_{y\text{HT}} + \hat{\mathbf{B}}_{\text{C}}' \left(T_{\mathbf{x}} - \hat{T}_{\mathbf{x}\text{HT}} \right) , \tag{6.7}$$

where $\hat{T}_{y\text{HT}} = \sum_{i \in \mathbf{S}} d_i y_i$ and

$$\hat{\mathbf{B}}_{\text{C}} = \left(\sum_{i \in \mathbf{S}} d_i q_i \mathbf{x}_i \mathbf{x}_i' \right)^{-1} \sum_{i \in \mathbf{S}} d_i q_i \mathbf{x}_i y_i \, .$$

The calibration estimator $\hat{T}_{y\text{C}}$ given by (6.7) for the population total T_y has the same structure as the generalized regression estimator $\hat{\mu}_{y\text{GREG}}$ given by (5.19) for the population mean μ_y. The $\hat{\mathbf{B}}_{\text{C}}$ becomes identical to $\hat{\mathbf{B}}$ given in (5.17) if we take $q_i = 1$ and include (6.4) as part of the calibration constraints. The definition of calibration estimators does not involve any models. However, it will become clear from discussions in Sect. 6.3 that calibration directly over auxiliary variables through constraints (6.1) can be motivated by a linear regression superpopulation model, which justifies the connection to the generalized regression estimator.

The calibration estimator $\hat{T}_{y\text{C}}$ is a nonlinear estimator of T_y, since the second term in (6.7) is a product of two random quantities, and therefore is no longer exactly design-unbiased. Note that $\hat{\mathbf{B}}_{\text{C}}$ is an estimator for the following population parameter

$$\mathbf{B}_{\text{C}} = \left(\sum_{i=1}^{N} q_i \mathbf{x}_i \mathbf{x}_i' \right)^{-1} \sum_{i=1}^{N} q_i \mathbf{x}_i y_i \, .$$

Under certain regularity conditions on the survey design and the variables y_i, $\mathbf{x}_i$ and the factor q_i as discussed in Chap. 7 on survey weighted estimating equations, we have

$$\hat{\mathbf{B}}_{\text{C}} - \mathbf{B}_{\text{C}} = O_p\left(n^{-1/2} \right) .$$

This leads to the following linearized theoretical expression for the calibration estimator,

$$\hat{T}_{y\text{C}} = \hat{T}_{y\text{HT}} + \mathbf{B}_{\text{C}}' \left(T_{\mathbf{x}} - \hat{T}_{\mathbf{x}\text{HT}} \right) + o_p\left(N n^{-1/2} \right) .$$

It follows that $\hat{T}_{y\text{C}}$ is approximately design-unbiased and the approximate design-based variance is given by

$$V_p\left(\hat{T}_{y\text{C}} \right) \doteq V_p\left(\hat{T}_{y\text{HT}} - \mathbf{B}_{\text{C}}' \hat{T}_{\mathbf{x}\text{HT}} \right) = V_p\left(\hat{T}_{e\text{HT}} \right) ,$$

where $\hat{T}_{e\mathrm{HT}} = \sum_{i\in\mathbf{S}} d_i e_i$ and $e_i = y_i - \mathbf{x}_i'\mathbf{B}_{\mathrm{C}}$. The theoretical variance of $\hat{T}_{y\mathrm{C}}$ is asymptotically equivalent to the variance of the Horvitz-Thompson estimator of $\hat{T}_{e\mathrm{HT}}$ defined over the residual variable e_i. An estimated variance of $\hat{T}_{y\mathrm{C}}$ can be obtained from $v_p\big(\hat{T}_{e\mathrm{HT}}\big)$, with e_i replaced by $\hat{e}_i = y_i - \mathbf{x}_i'\hat{\mathbf{B}}_{\mathrm{C}}$.

It is important to notice that the calibration estimator of T_y is design-consistent regardless of the auxiliary variables used in calibration and their relation to the y variable. Calibration estimators are generally more efficient than the Horvitz-Thompson estimator. Biases are usually negligible with moderate sample sizes and variances are typically smaller than the variance of the HT estimator. Suppose that the population size N is known and both (6.4) and (6.1) are used for calibration. It can be seen that under simple random sampling without replacement and the choice $q_i = 1$ in $\chi^2(\mathbf{w}, \mathbf{d})$, the calibration estimator of T_y is given by

$$\hat{T}_{y\mathrm{C}} = N\,\hat{\mu}_{y\mathrm{GREG}}\,,$$

where $\hat{\mu}_{y\mathrm{GREG}}$ is the GREG estimator of the population mean μ_y given in Sect. 5.4.1, which is the optimal regression estimator under SRSWOR (Problem 5.9).

The q-factor in the chi-square distance measure has negligible effect on the bias of the calibration estimator since it appears only in $\hat{\mathbf{B}}_{\mathrm{C}}$. Different choices of the q-factor, however, might affect the variance of the estimator through the residual variable $e_i = y_i - \mathbf{x}_i'\mathbf{B}_{\mathrm{C}}$. One of the research problems is to identify specific forms of the q_i for more efficient calibration estimation. Under Poisson sampling, it is shown in Problem 6.3 that setting $q_i = \pi_i^{-1} - 1$ in $\chi^2(\mathbf{w}, \mathbf{d})$ leads to the optimal calibration estimator.

The choice of $q_i = \pi_i^{-1} - 1$ has been discussed in several research papers, including Brewer (1999) on cosmetic calibration, Berger et al. (2003) on optimal regression estimation, Tan (2013) on efficient calibration estimators under rejective and high-entropy sampling, and Park and Kim (2014) on instrumental-variable calibration. Tan and Wu (2015) provided an intuitive interpretation for the choice of $q_i = \pi_i^{-1} - 1$. Recall that $A_i = I(i \in \mathbf{S})$ is the indicator variable for unit i being selected in the sample. We have $E_p(A_i/\pi_i) = 1$ and $V_p(A_i/\pi_i) = \pi_i^{-1} - 1$. The value of $q_i = \pi_i^{-1} - 1$ reflects the variation of selecting unit i into the sample under the survey design. In the extreme case of $\pi_i = 1$, there is no uncertainty associated with selecting unit i, and it is desirable to force $w_i = d_i = 1$. With the choice of $q_i = \pi_i^{-1} - 1$, large values of π_i are associated with large values of q_i^{-1}, which forces w_i to be closer to d_i for the constrained minimization of $\chi^2(\mathbf{w}, \mathbf{d})$. Section 6.2 provides some empirical results on the impact of the q-factor on the distributions of calibration weights.

For other distance measures $D(\mathbf{w}, \mathbf{d})$ belonging to the class discussed by Deville and Särndal (1992), the resulting calibration estimators are asymptotically equivalent to the GREG-type estimator given by (6.7) using the chi-square distance. Two of those distance measures are discussed in the next section.

6.2 Calibration Weighting

The calibration weights w_i given by (6.6), obtained under the chi-square distance $\chi^2(\mathbf{w}, \mathbf{d})$, are also called the *generalized regression* weights since the calibration estimator $\hat{T}_{y\mathrm{C}} = \sum_{i\in\mathbf{S}} w_i y_i$ is a GREG-type estimator given by (6.7). We can rewrite (6.6) as $w_i = d_i g_i$, where the so-called g-weights (not to be confused with the q-factor in the distance measure) are given by

$$g_i = 1 + \left(T_{\mathbf{x}} - \hat{T}_{\mathbf{x}\mathrm{HT}}\right)' \left(\sum_{j\in\mathbf{S}} d_j q_j \mathbf{x}_j \mathbf{x}_j'\right)^{-1} \mathbf{x}_i q_i = 1 + \boldsymbol{\lambda}_1' \mathbf{x}_i q_i \tag{6.8}$$

and

$$\boldsymbol{\lambda}_1 = \left(\sum_{j\in\mathbf{S}} d_j q_j \mathbf{x}_j \mathbf{x}_j'\right)^{-1} \left(T_{\mathbf{x}} - \hat{T}_{\mathbf{x}\mathrm{HT}}\right).$$

The g-weights satisfy $g_i = 1 + o_p(1)$ under certain regularity conditions; see Sect. 6.2.3 for further details. For finite samples, however, the generalized regression g-weights given by (6.8) may take values which are very different from 1 and sometimes even negative values, resulting in negative calibration weights.

There are several alternative distance measures available for obtaining positive calibration weights. Computational algorithms and finite sample behaviour of the final calibration weights are major considerations for choosing a particular distance measure. We focus on two popular ones related to the Kullback-Leibler information distance (Kullback and Leibler 1951). Both distance measures belong to the general class of Deville and Särndal (1992) and their connection to the Kullback-Leibler distance was discussed by Wu and Lu (2016). More importantly, there exist efficient and reliable computational procedures for the constrained minimization problem with both distance measures. In Sect. 6.2.4, we also provide discussion on the practically important problem of range-restricted calibration weighting methods.

6.2.1 Generalized Exponential Tilting

The modified backward Kullback-Leibler distance between $\mathbf{w} = (w_1, \ldots, w_n)$ and $\mathbf{d} = (d_1, \ldots, d_n)$, weighted by an arbitrary q-factor, is given by

$$ET(\mathbf{w}, \mathbf{d}) = \sum_{i\in\mathbf{S}} q_i^{-1} \left\{ w_i \log\left(\frac{w_i}{d_i}\right) - w_i + d_i \right\}.$$

It can be shown that $G_i(w_i, d_i) = q_i^{-1}\{w_i \log(w_i/d_i) - w_i + d_i\} \geq 0$ for all i and $G_i(w_i, d_i) = 0$ if $w_i = d_i$. The notation ET stands for "Exponential Tilting", due to the form of the final weights w_i given below.

The calibration weights obtained by minimizing the distance $ET(\mathbf{w}, \mathbf{d})$ subject to constraint (6.1) are given by $w_i = d_i g_i$, where

$$g_i = \exp(\lambda_2' \mathbf{x}_i q_i) \tag{6.9}$$

and λ_2 is the solution to

$$\Delta_1(\lambda) = \sum_{i \in \mathbf{S}} d_i \mathbf{x}_i \exp\left(\lambda' \mathbf{x}_i q_i\right) - t_{\mathbf{x}} = \mathbf{0}\,. \tag{6.10}$$

The final weights $w_i = d_i \exp(\lambda_2' \mathbf{x}_i q_i)$ are an exponentially tilted version of the design weights d_i. It is apparent that $w_i > 0$ for all i. Folsom (1991) provides an early example of exponential weight adjustment. Kim (2010) contains further discussions of calibration estimation using exponential tilting.

Equations (6.10) are a nonlinear system and finding the solution λ_2 requires iterative procedures. The following standard Newton-Raphson procedure can be used to find the solution λ_2 to (6.10):

$$\lambda^{(m+1)} = \lambda^{(m)} - \left\{\Delta_2\left(\lambda^{(m)}\right)\right\}^{-1} \Delta_1\left(\lambda^{(m)}\right), \quad m = 0, 1, 2, \ldots\,, \tag{6.11}$$

where

$$\Delta_2(\lambda) = \frac{\partial}{\partial \lambda} \Delta_1(\lambda) = \sum_{i \in \mathbf{S}} q_i d_i \mathbf{x}_i \mathbf{x}_i' \exp\left(\lambda' \mathbf{x}_i q_i\right).$$

The initial value of λ can simply be chosen as $\lambda^{(0)} = \mathbf{0}$. If $\sum_{i \in \mathbf{S}} w_i = N$ is part of the constraints for calibration, then a necessary and sufficient condition for the existence of a solution and the convergence of the Newton-Raphson algorithm is that the vector of population means $\mu_{\mathbf{x}} = T_{\mathbf{x}}/N$ is an inner point of the convex hull formed by $\{\mathbf{x}_i, i \in \mathbf{S}\}$. This is due to a duality property of the constrained minimization problem as discussed in Wu and Lu (2016): minimizing $ET(\mathbf{w}, \mathbf{d})$ with respect to the w_i's under the constraints (6.1) is equivalent to the problem of unconstrained convex minimization of $K(\lambda) = \sum_{i \in \mathbf{S}} q_i^{-1} d_i \exp(\lambda' \mathbf{x}_i q_i) - \lambda' T_{\mathbf{x}}$ with respect to λ, since $\partial K(\lambda)/\partial \lambda = \Delta_1(\lambda)$ and $\partial^2 K(\lambda)/\partial \lambda \partial \lambda' = \sum_{i \in \mathbf{S}} q_i d_i \mathbf{x}_i \mathbf{x}_i' \exp(\lambda' \mathbf{x}_i q_i)$ is positive definite.

The iterative procedure (6.11) can be easily implemented in popular statistical software packages; see Appendix A.3.2 for computing codes in R for the pseudo empirical likelihood methods where $\Delta_1(\lambda)$ and $\Delta_2(\lambda)$ have different forms but the Newton-Raphson iterative procedures are essentially the same.

6.2.2 Generalized Pseudo Empirical Likelihood

The modified forward Kullback-Leibler distance between $\mathbf{w}$ and $\mathbf{d}$ is obtained by switching the positions of w_i and d_i in $ET(\mathbf{w}, \mathbf{d})$ and is given by

$$EL(\mathbf{w}, \mathbf{d}) = \sum_{i \in \mathbf{S}} q_i^{-1} \left\{ d_i \log\left(\frac{d_i}{w_i}\right) - d_i + w_i \right\},$$

where EL stands for "Empirical Likelihood". The distance is related to the pseudo empirical likelihood function, to be discussed in Chap. 8, with a general q-factor.

Using the standard Lagrange multiplier method (Problem 6.4), the solution to minimizing $EL(\mathbf{w}, \mathbf{d})$ with respect to w_i subject to calibration equations (6.1) is given by $w_i = d_i g_i$, where

$$g_i = \frac{1}{1 - \lambda_3' \mathbf{x}_i q_i} \tag{6.12}$$

and the Lagrange multiplier λ_3 is the solution to

$$\Delta_1(\lambda) = \sum_{i \in \mathbf{S}} \frac{d_i \mathbf{x}_i}{1 - \lambda' \mathbf{x}_i q_i} - t_{\mathbf{x}} = \mathbf{0}\,. \tag{6.13}$$

The standard Newton-Raphson iterative procedure (6.11) might be used to find the solution to (6.13). It turns out that the computational problem with $EL(\mathbf{w}, \mathbf{d})$ has a specific feature which can be handled more efficiently by the following modified procedures.

The problem of constrained minimization of $EL(\mathbf{w}, \mathbf{d})$ with respect to the w_i's subject to (6.1) is equivalent to the unconstrained minimization of

$$K(\lambda) = -\sum_{i \in \mathbf{S}} q_i^{-1} d_i \log\left(1 - \lambda' \mathbf{x}_i q_i\right) - \lambda' T_{\mathbf{x}}$$

with respect to λ within the restricted region $\Lambda = \{\lambda \mid 1 - \lambda' \mathbf{x}_i q_i > 0,\ i \in \mathbf{S}\}$. The set Λ is convex, since for any $\alpha \in (0, 1)$, $\alpha\lambda_1 + (1 - \alpha)\lambda_2 \in \Lambda$ if the two points λ_1, $\lambda_2 \in \Lambda$. The function $K(\lambda)$ is also convex, since the Hessian matrix

$$\Delta_2(\lambda) = \frac{\partial^2}{\partial\lambda\partial\lambda'} K(\lambda) = \sum_{i \in \mathbf{S}} \frac{q_i d_i \mathbf{x}_i \mathbf{x}_i'}{(1 - \lambda' \mathbf{x}_i q_i)^2} \tag{6.14}$$

is positive definite. The minimum point of $K(\lambda)$ is the solution to $\partial K(\lambda)/\partial\lambda = \mathbf{0}$, which are the same equations $\Delta_1(\lambda) = \mathbf{0}$ given by (6.13). The following modified Newton-Raphson procedure, originating from the algorithm of Chen et al. (2002) for the pseudo empirical likelihood methods (Sect. 8.4.1), can be used to find the solution to (6.13).

Step 0 Let $\lambda^{(0)} = \mathbf{0}$. Set $m = 0$ and $\varepsilon = 10^{-8}$.

Step 1 Calculate $\delta^{(m)} = \big\{\Delta_2(\lambda^{(m)})\big\}^{-1}\Delta_1(\lambda^{(m)})$, where $\Delta_1(\lambda)$ is given in (6.13) and $\Delta_2(\lambda)$ is given by (6.14). If $\|\delta^{(m)}\| \le \varepsilon$, stop the iteration and report $\lambda = \lambda^{(m)}$; otherwise go to Step 2.

Step 2 If $1 - \big(\lambda^{(m)} - \delta^{(m)}\big)'\mathbf{x}_i q_i \le 0$ for any $i \in \mathbf{S}$, let $\delta^{(m)} = \delta^{(m)}/2$. Repeat the step if necessary.

Step 3 Set $\lambda^{(m+1)} = \lambda^{(m)} - \delta^{(m)}$, $m = m + 1$. Go to Step 1.

The above procedures use the standard Newton-Raphson method plus a checking step at each iteration to ensure that the solution is a point inside the restricted convex set Λ. See Appendix A.3.2 for implementation of the algorithm in R.

6.2.3 Comparisons Among Three Calibration Methods

The generalized regression g-weights have the simple linear form given by (6.8). The Lagrange multiplier satisfies $\lambda_1 = O_p(n^{-1/2})$ under the commonly assumed conditions $N^{-1}(T_{\mathbf{x}} - \hat{T}_{\mathbf{x}\mathrm{HT}}) = O_p(n^{-1/2})$ and $N^{-1}\sum_{i\in\mathbf{S}} d_i q_i \mathbf{x}_i\mathbf{x}_i' = O_p(1)$. The conditions can be stated more specifically on the survey design, the auxiliary variables and the q-factor as detailed in Wu and Rao (2006) and Tan and Wu (2015). Under the same regularity conditions, we have $\max_{i\in\mathbf{S}}\|q_i\mathbf{x}_i\| = o_p(n^{1/2})$. This leads to the conclusion that $\max_{i\in\mathbf{S}}|g_i| = 1 + o_p(1)$ for the generalized regression method for calibration weighting.

The g-weights for generalized exponential tilting and generalized pseudo empirical likelihood have nonlinear forms and are given by (6.9) and (6.12), respectively. The expressions for the g_i involve Lagrange multipliers λ_2 and λ_3 which require solving nonlinear systems using iterative procedures. It might come as a surprise that both λ_2 and λ_3 are asymptotically equivalent to λ_1. It is shown by Deville and Särndal (1992, Result 3 on page 379) that $\lambda_k = \lambda_1 + O_p(n^{-1})$ for $k = 2, 3$. Let $g_i^{(k)}$, $k = 1, 2, 3$ be respectively the g-weights for the three calibration weighting methods: (i) generalized regression; (ii) generalized exponential tilting; and (iii) generalized pseudo empirical likelihood. We have

(i) $g_i^{(1)} = 1 + \lambda_1'\mathbf{x}_i q_i$;

(ii) $g_i^{(2)} = \exp(\lambda_2'\mathbf{x}_i q_i) = 1 + \lambda_2'\mathbf{x}_i q_i + o_p(n^{-1/2}) = 1 + \lambda_1'\mathbf{x}_i q_i + o_p(n^{-1/2})$;

(iii) $g_i^{(3)} = 1/(1 - \lambda_3'\mathbf{x}_i q_i) = 1 + \lambda_3'\mathbf{x}_i q_i + o_p(n^{-1/2}) = 1 + \lambda_1'\mathbf{x}_i q_i + o_p(n^{-1/2})$.

The three methods produce similar calibration weights up to the order $o_p(n^{-1/2})$. For finite samples, however, there are noticeable differences among the weights. Wu and Lu (2016) examined the behavior of g-weights through a simulation study. They considered a population of $N = 4000$ with four auxiliary variables: $\mathbf{x}_i = (x_{i1}, x_{i2}, x_{i3}, x_{i4})'$. Samples of size $n = 100$ were selected by the Rao-Sampford method (Rao 1965; Sampford 1967) with inclusion probabilities π_i proportional to x_{i2}. The variation of the basic design weights $d_i = 1/\pi_i$ was controlled by setting

Table 6.1 Simulated lower and upper quantiles of the g-weights ($\max \pi_i / \min \pi_i = 100$)

q-factor	$D(\mathbf{w}, \mathbf{d})$	0.0%	0.1%	0.5%	1.0%	99%	99.5%	99.9%	100%
$q_i = 1$	χ^2	−1.15	−0.76	−0.09	0.19	2.11	2.32	3.11	6.52
	ET	0.02	0.12	0.26	0.36	2.27	2.59	3.89	9.27
	EL	0.19	0.31	0.41	0.48	2.40	3.10	6.09	11.73
$q_i = \pi_i^{-1} - 1$	χ^2	−0.68	−0.29	0.17	0.34	1.94	2.24	3.09	4.53
	ET	0.03	0.15	0.34	0.44	1.99	2.40	3.57	6.71
	EL	0.18	0.28	0.44	0.51	1.96	2.59	4.74	9.72

$\max \pi_i / \min \pi_i = 100$. Two choices of the q-factor, $q_i = 1$ and $q_i = \pi_i^{-1} - 1$, were considered and all four auxiliary variables were used for calibration. It is of primary interest to see the patterns of the g-weights at the tail regions (i.e., small or large values). The lower and upper quantiles of the g-weights based on 1000 simulated samples of size $n = 100$ are presented in Table 6.1. The 0.0% and 100% quantiles represent the minimum and maximum values of the weights.

Under the simulation settings, small or large values of g-weights do occur but the frequencies for $g_i < 0.2$ or $g_i > 5$ are less than one percent. The generalized regression g-weights take a few negative values for most simulated samples. The ET and EL methods both produce positive g-weights, with smaller weights occurring more often for the ET method and larger weights more frequently for the EL methods. The use of $q_i = \pi_i^{-1} - 1$ in the distance measure has an impact of reducing the variation of the q-weights for all three methods.

6.2.4 Range-Restricted Calibration

Calibration weights are routinely supplied by survey organizations as part of the public-use data files for design-based analysis. The data files might be used by different researchers and for different objectives. When the population size N is known and $\sum_{i \in \mathbf{S}} w_i = N$ is included as a constraint for calibration weighting, the calibration estimator $\hat{T}_{y\text{C}} = \sum_{i \in \mathbf{S}} w_i y_i$ could be interpreted as an expansion estimator. Extremely large or small or even negative g-weights make such an interpretation unacceptable to users. In addition, high variations in calibration weights reduce the efficiency of analysis when main study variables are weakly correlated with the calibration variables or the research problems are beyond the estimation of descriptive population parameters.

The issue of negative values in regression weights has been a known problem for many years. Huang and Fuller (1978) made the first effort to address the issue. Further discussions can be found in Park and Fuller (2005). The generalized exponential tilting and the generalized pseudo empirical likelihood methods produce calibration weights which are guaranteed positive, but the problem of extremely small or very large weights remains.

A general research problem is on range-restricted methods for calibration weighting. Let $c_1 \in (0, 1)$ and $c_2 > 1$ be two pre-specified constants. The problem is how to obtain calibration weights $w_i = d_i g_i$ which satisfy

$$c_1 d_i \le w_i \le c_2 d_i \quad \text{or} \quad c_1 \le g_i \le c_2 \,, \quad i \in \mathbf{S} \,. \tag{6.15}$$

The value of $g_i = w_i/d_i$ represents the relative amount of departure of the calibration weight w_i from the basic design weight d_i, and it is desirable to restrict the range of g_i so that the calibration weights are not dramatically different from the design weights. From a computational point of view, the problem of range-restricted calibration amounts to minimizing a distance measure subject to the benchmark constraints (6.1) as well as the range restrictions (6.15). Unfortunately, this seemingly simple problem does not seem to have a simple solution. In fact, there might be no solutions to the problem if the restrictions on the g_i's are too tight, since $\sum_{i \in \mathbf{S}} w_i \mathbf{x}_i \to \hat{T}_{\mathbf{x}\mathrm{HT}}$ as $c_1 \to 1$ and $c_2 \to 1$. If $\hat{T}_{\mathbf{x}\mathrm{HT}} \neq T_{\mathbf{x}}$ for the given sample, the constraints (6.1) and (6.15) will become impossible to satisfy simultaneously at certain points as we move both c_1 and c_2 towards 1.

Deville and Särndal (1992) suggested to use $w_i = d_i g_i$ where

$$g_i = \frac{c_1(c_2 - 1) + c_2(1 - c_1)\exp(b\lambda' \mathbf{x}_i q_i)}{(c_2 - 1) + (1 - c_1)\exp(b\lambda' \mathbf{x}_i q_i)} \,,$$

$b = (c_2 - c_1)/\{(1 - c_1)(c_2 - 1)\}$. It is apparent that $c_1 \le g_i \le c_2$ for all i. The value of λ is determined by the calibration equations $\sum_{i \in \mathbf{S}} d_i g_i \mathbf{x}_i = T_{\mathbf{x}}$. Kim (2010) proposed to use $g_i = \exp(\lambda_0 + \lambda \mathbf{u}_i)$, where $\mathbf{u}_i = \mathbf{u}(\mathbf{x}_i)$ is a vector of instrumental variables derived from $\mathbf{x}_i$. The values of λ_0 and λ are the solution to $\sum_{i \in \mathbf{S}} d_i g_i \mathbf{x}_i = T_{\mathbf{x}}$, and the form of $\mathbf{u}(\mathbf{x}_i)$ is chosen to satisfy (6.15). Kim (2010) suggested to use a truncated form of $\mathbf{x}_i$ for $\mathbf{u}_i$. The major issues with this type of approach, however, are the lack of efficient computational procedures and the nonexistence of a solution for the method.

There exist methods to obtain range-restricted calibration weights that allow the benchmark constraints (6.1) to be nonbinding (i.e., not hold exactly). Rao and Singh (1997) proposed a ridge-shrinkage method when the exact calibration equations are replaced by $|\sum_{i \in \mathbf{S}} w_i \mathbf{x}_i - T_{\mathbf{x}}| \le \delta |T_{\mathbf{x}}|$ for some pre-specified tolerance δ. Chen et al. (2002) presented an algorithm to search for range-restricted calibration weights with the minimum amount of relaxation of the benchmark constraints. See Problem 6.5 for a simplified version of the method.

The observation from Sect. 6.2.3 that only a small fraction of g-weights might take extreme values motivates the following simple weight trimming method for range-restricted calibration (Wu and Lu 2016). Let $w_i = d_i g_i$ be the calibration weights obtained from a well established method. Let

$$w_i^* = \begin{cases} c_1 d_i & \text{if} \quad g_i < c_1 \\ k w_i & \text{if} \quad c_1 \le g_i \le c_2 \\ c_2 d_i & \text{if} \quad g_i > c_2 \end{cases}$$

be the trimmed weights. The re-scaling constant k is chosen such that $\sum_{i \in \mathbf{S}} w_i = \sum_{i \in \mathbf{S}} w_i^*$. If N is known and $\sum_{i \in \mathbf{S}} w_i = N$ is part of the calibration equations, we could simply set k to satisfy $\sum_{i \in \mathbf{S}} w_i^* = N$. The benchmark constraints might not hold exactly with the trimmed weights w_i^* but the departure is typically small when the restrictions are not too tight. More importantly, the trimmed weights can be used for valid design-based inferences.

Consider the commonly used "symmetric" range-restriction where $c_2 = \eta > 1$ and $c_1 = 1/\eta$. Let $w_i^* = w_i^*(\eta)$ be the final trimmed weights. Let $\hat{T}_y(\eta) = \sum_{i \in \mathbf{S}} w_i^*(\eta) y_i$. If the initial calibration weights w_i are all positive, we have

$$\hat{T}_y(1) = \hat{T}_{\text{HT}} = \sum_{i \in \mathbf{S}} d_i y_i \qquad \text{and} \qquad \hat{T}_y(\infty) = \hat{T}_{y\text{C}} = \sum_{i \in \mathbf{S}} w_i y_i \, .$$

The calibration estimator $\hat{T}_y(\eta)$ based on the trimmed weights lies between the two valid design-based estimators $\hat{T}_{\text{HT}}$ and $\hat{T}_{y\text{C}}$. Wu and Lu (2016) demonstrated through a simulation study that $\hat{T}_y(\eta)$ is always more efficient than $\hat{T}_{\text{HT}}$ in terms of mean squared errors and is almost identical to $\hat{T}_{y\text{C}}$ with $\eta \ge 4$ for the settings used in the simulation.

6.3 Model-Calibration Estimators

Conventional calibration weighting and estimation uses auxiliary information in the forms of known population totals or means. For certain survey populations, auxiliary information might be available for every unit in the population. This is the so-called complete auxiliary information where $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$ is known for the entire survey population. Complete auxiliary information is typically available from census data or administrative data such as business and industry registrations or government records of individual level data files. There are confidentiality issues and practical constraints and procedures for accessing such detailed information. Benchmarking over the known population totals might remain as a point of interest, but the focus of this section is on how to use the information, if it is available, for more efficient estimation of population parameters. Note that the population size N is always known with complete auxiliary information. Let $\{(y_i, \mathbf{x}_i), i \in \mathbf{S}\}$ be the survey sample data set, selected by a general probability sampling design with inclusion probabilities π_i and π_{ij}.

6.3.1 Model-Calibration Estimators of the Population Total

Calibration estimators $\hat{T}_{y\text{C}}$ of the population total under the benchmark constraints (6.1) are asymptotically equivalent to a GREG-type estimator, which is motivated by a linear regression model as discussed in Sects. 5.2–5.4. The relation between conventional calibration estimators and a linear model can also be justified by the extreme case where the response variable y and the auxiliary variables $\mathbf{x}$ have a deterministic linear relationship: $y_i = \mathbf{x}_i'\beta$, $i = 1, \ldots, N$. In this case $T_y = T_{\mathbf{x}}'\beta$, and the calibrated weights w_i which satisfy the benchmark constraints $\sum_{i \in \mathbf{S}} w_i \mathbf{x}_i = T_{\mathbf{x}}$ would lead to the "perfect" calibration estimator of T_y: $\hat{T}_{y\text{C}} = \sum_{i \in \mathbf{S}} w_i y_i = \sum_{i \in \mathbf{S}} w_i \mathbf{x}_i'\beta = T_{\mathbf{x}}'\beta = T_y$. If the linear model contains an intercept, then the constraint (6.4) should also be included.

Consider a semiparametric superpopulation model ξ specified by the first two moments of the conditional distribution of the response given the auxiliary variables:

$$E_\xi\big(y_i \mid \mathbf{x}_i\big) = \mu_i = \mu(\mathbf{x}_i, \theta) \quad \text{and} \quad V_\xi\big(y_i \mid \mathbf{x}_i\big) = v_i = \sigma^2 v(\mu_i)\,, \quad i = 1, \ldots, N\,, \tag{6.16}$$

where $\mu(\mathbf{x}, \theta)$ is a known mean function and $v(\mu)$ is a known variance function. The y_i's are assumed to be conditionally independent given the $\mathbf{x}_i$'s. The θ is a vector of superpopulation parameters. Let θ_N be the census parameter and $\hat{\theta}$ be the design-based estimator of θ_N. We assume that $\hat{\theta} = \theta_N + O_p(n^{-1/2})$. See Chap. 7 for detailed discussions on estimation of model parameters using survey data. Wu and Sitter (2001a) defined the model-calibration estimator of T_y as $\hat{T}_{y\text{MC}} = \sum_{i \in \mathbf{S}} w_i y_i$, where the calibrated weights w_i minimize a distance measure $D(\mathbf{w}, \mathbf{d})$ subject to

$$\sum_{i \in \mathbf{S}} w_i = N\,, \tag{6.17}$$

$$\sum_{i \in \mathbf{S}} w_i \mu(\mathbf{x}_i, \hat{\theta}) = \sum_{i=1}^{N} \mu(\mathbf{x}_i, \hat{\theta})\,. \tag{6.18}$$

The constraint (6.17) is re-listed here from (6.4) to emphasize its role in model-calibration since the constraint (6.18) cannot be assumed to include (6.17). The fitted values $\hat{\mu}_i = \mu(\mathbf{x}_i, \hat{\theta})$, not the values $\mathbf{x}_i$ of the auxiliary variables, are used in forming the calibration constraint (6.18). The quantity $\sum_{i=1}^{N} \mu(\mathbf{x}_i, \hat{\theta})$ typically requires complete auxiliary information $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ and as a consequence, the population size N is assumed to be known.

Under the chi-square distance $\chi^2(\mathbf{w}, \mathbf{d})$ with a general q-factor, the model-calibration estimator is given by

$$\hat{T}_{y\text{MC}} = \hat{T}_{y\text{HT}} + \Big(N - \sum_{i \in \mathbf{S}} d_i\Big)\hat{B}_0 + \Big(\sum_{i=1}^{N} \hat{\mu}_i - \sum_{i \in \mathbf{S}} d_i \hat{\mu}_i\Big)\hat{B}_1\,,$$

where

$$(\hat{B}_0, \hat{B}_1)' = \hat{\mathbf{B}} = \Big(\sum_{i\in\mathbf{S}} d_i q_i \mathbf{u}_i \mathbf{u}_i'\Big)^{-1} \sum_{i\in\mathbf{S}} d_i q_i \mathbf{u}_i y_i$$

and $\mathbf{u}_i = (1, \hat{\mu}_i)'$. The second term $\big(N - \sum_{i\in\mathbf{S}} d_i\big)\hat{B}_0$ in $\hat{T}_{y\text{MC}}$ was incidentally omitted in equation (9) of Wu and Sitter (2001a, page 187). Let $\tilde{\mu} = \sum_{i\in\mathbf{S}} d_i q_i \hat{\mu}_i / \sum_{i\in\mathbf{S}} d_i q_i$ and $\tilde{y} = \sum_{i\in\mathbf{S}} d_i q_i y_i / \sum_{i\in\mathbf{S}} d_i q_i$. We have

$$\hat{B}_1 = \frac{\sum_{i\in\mathbf{S}} d_i q_i (\hat{\mu}_i - \tilde{\mu})(y_i - \tilde{y})}{\sum_{i\in\mathbf{S}} d_i q_i (\hat{\mu}_i - \tilde{\mu})^2} \quad \text{and} \quad \hat{B}_0 = \tilde{y} - \hat{B}_1 \tilde{\mu}\,.$$

If constraint (6.17) is dropped, the single calibration equation (6.18) with the chi-square distance yields

$$\hat{T}^*_{y\text{MC}} = \sum_{i\in\mathbf{S}} w_i y_i = \hat{T}_{y\text{HT}} + \Big(\sum_{i=1}^{N} \hat{\mu}_i - \sum_{i\in\mathbf{S}} d_i \hat{\mu}_i\Big)\hat{B}^*_1\,,$$

where $\hat{B}^*_1 = \big\{\sum_{i\in\mathbf{S}} d_i q_i \hat{\mu}_i y_i\big\} / \big\{\sum_{i\in\mathbf{S}} d_i q_i \hat{\mu}_i^2\big\}$. It should be noted that the weights w_i for the model-calibration estimators $\hat{T}_{y\text{MC}}$ or $\hat{T}^*_{y\text{MC}}$ depend on the response variable y through the estimator $\hat{\theta}$ for the model parameters. The weights are not suitable for public-use survey data files for general design-based inferences.

Model-calibration estimators have several attractive features. The first important property is that the effect of estimating the model parameters is asymptotically negligible. Under certain regularity conditions, it is shown by Wu and Sitter (2001a) that

$$\sum_{i=1}^{N} \hat{\mu}_i - \sum_{i\in\mathbf{S}} d_i \hat{\mu}_i = \sum_{i=1}^{N} \mu_i - \sum_{i\in\mathbf{S}} d_i \mu_i + O_p\big(Nn^{-1}\big)\,,$$

where $\hat{\mu}_i = \mu(\mathbf{x}_i, \hat{\theta})$, $\mu_i = \mu(\mathbf{x}_i, \theta)$ and $\hat{\theta} = \theta + O_p(n^{-1/2})$. One can simply treat $\hat{\theta}$ as if it is the true value of the parameter for asymptotic development. It is straightforward to show that the model-calibration estimators of T_y are asymptotically design-unbiased. They are also approximately model-unbiased under the assumed model.

The second feature of the model-calibration estimator is its connection to the generalized regression estimator. Under a linear regression model with an intercept, we have $\mu_i = \mu(\mathbf{x}_i, \theta) = \mathbf{z}_i'\theta$, where $\mathbf{z}_i = (1, \mathbf{x}_i')'$. The constraint (6.18) becomes $\sum_{i\in\mathbf{S}} w_i \mathbf{z}_i' \hat{\theta} = T_{\mathbf{z}}' \hat{\theta}$, which only requires information on the population totals $T_{\mathbf{z}} = (N, T_{\mathbf{x}}')'$. Consider the survey-weighted estimator of θ given by

$$\hat{\theta} = \Big(\sum_{i\in\mathbf{S}} d_i \mathbf{z}_i \mathbf{z}_i'\Big)^{-1} \sum_{i\in\mathbf{S}} d_i \mathbf{z}_i y_i \, . \tag{6.19}$$

It is shown in Problem 6.6 that $\hat{B}_1 = \hat{B}_1^* = 1$ and $\hat{B}_0 = 0$ if we set $q_i = 1$ in the distance measure. Consequently, both $\hat{T}_{y\text{MC}}$ and $\hat{T}_{y\text{MC}}^*$ reduce to the generalized regression estimator $\hat{T}_{y\text{GREG}} = \hat{T}_{y\text{HT}} + \hat{\theta}'\big(T_{\mathbf{z}} - \hat{T}_{\mathbf{z}\text{HT}}\big)$.

The third major property of the model-calibration estimator is its optimality among a class of calibration estimators. Consider calibration estimators $\hat{T}_{y\text{C}}(m)$ under the constraint

$$\sum_{i\in\mathbf{S}} w_i m(\mathbf{x}_i) = \sum_{i=1}^{N} m(\mathbf{x}_i) \, ,$$

where $m(\mathbf{x})$ is a real-valued function of $\mathbf{x}$. It is shown in Wu (2003) that the anticipated asymptotic design-based variance $E_\xi\big[AV_p\big\{\hat{T}_{y\text{C}}(m)\big\}\big]$ under model (6.16) is minimized when we choose $m(\mathbf{x}) = \mu(\mathbf{x}, \theta)$.

6.3.2 Model-Calibration Estimators of Other Parameters

The model-calibration approach to the estimation of population totals provides a general framework for constructing efficient calibration estimators for other finite population parameters. We discuss two examples.

The finite population distribution function $F_y(t) = N^{-1}\sum_{i=1}^{N} I(y_i \le t)$ at a fixed t can be viewed as a population mean defined over the indicator variable $I(y \le t)$. The conventional calibration estimator with constraints on the auxiliary variables $\mathbf{x}$ is not efficient due to the weak correlation between $I(y \le t)$ and $\mathbf{x}$. Let $u_i = E_\xi\big\{I(y_i \le t) \mid \mathbf{x}_i\big\} = P(y_i \le t \mid \mathbf{x}_i)$. The optimal calibration estimator of $F_y(t)$ is given by $\hat{F}_{y\text{MC}}(t) = N^{-1}\sum_{i\in\mathbf{S}} w_i I(y_i \le t)$, where the weights w_i minimize a distance measure subject to

$$\sum_{i\in\mathbf{S}} w_i = N \quad \text{and} \quad \sum_{i\in\mathbf{S}} w_i \hat{u}_i = \sum_{i=1}^{N} \hat{u}_i \, .$$

The $\hat{u}_i$'s are the fitted values for the binary variable $I(y_i \le t)$. Chen and Wu (2002) discussed three alternative ways of obtaining the u_i's. It should be noted that the model-calibrated weights w_i depend on the value of t. If the objective is to estimate $F_y(t)$ as a distribution function, a practical approach is to impose multiple constraints using, for instance, five values of t at the 10%, 25%, 50%, 75% and 90% quantiles of the sampled y_i's.

The finite population variance $\sigma_y^2 = (N-1)^{-1}\sum_{i=1}^N (y_i - \mu_y)^2$ could be a parameter of interest. The Laplace form of σ_y^2 (Problem 2.1) is given by

$$\sigma_y^2 = \frac{1}{N(N-1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \left(y_i - y_j\right)^2 .$$

This is a parameter of quadratic forms (Liu and Thompson 1983; Sitter and Wu 2002). Let $d_{ij} = 1/\pi_{ij}$. The model-calibration estimator can be constructed as

$$\hat{\sigma}_{y\mathrm{MC}}^2 = \frac{1}{N(N-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} w_{ij}\left(y_i - y_j\right)^2 ,$$

where the weights w_{ij} minimize the chi-square distance $\chi^2(\mathbf{w}, \mathbf{d}) = \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} (w_{ij} - d_{ij})^2/d_{ij}$ subject to the constraint

$$\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} w_{ij}u_{ij} = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} u_{ij} . \tag{6.20}$$

Under model (6.16), the optimal calibration variable is given by $u_{ij} = E_\xi\left\{(y_i - y_j)^2 \mid \mathbf{x}_i, \mathbf{x}_j\right\} = \left(\mu_i - \mu_j\right)^2 + \sigma^2\left\{v(\mu_i) + v(\mu_j)\right\}$. Let $\hat{\mu}_i$ be the fitted value for y_i. It is convenient to replace (6.20) by two constraints

$$\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} w_{ij}\left(\hat{\mu}_i - \hat{\mu}_j\right)^2 = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \left(\hat{\mu}_i - \hat{\mu}_j\right)^2 . \tag{6.21}$$

$$\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} w_{ij}\left\{v(\hat{\mu}_i) + v(\hat{\mu}_j)\right\} = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \left\{v(\hat{\mu}_i) + v(\hat{\mu}_j)\right\} . \tag{6.22}$$

The right-hand side of (6.21) can be simplified as $N\sum_{i=1}^N \left(\hat{\mu}_i - \bar{\mu}_N\right)^2$, where $\bar{\mu}_N = N^{-1}\sum_{i=1}^N \hat{\mu}_i$, and the right-hand side of (6.22) can be simplified as $(N-1)\sum_{i=1}^N v(\hat{\mu}_i)$. If the model (6.16) has a homogeneous variance structure, i.e., $v(\mu_i) = 1$, the constraint (6.22) becomes

$$\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} w_{ij} = N(N-1)/2 .$$

Sitter and Wu (2002) and Wu (2003) contain further discussions and simulation results on optimal model-calibration estimators for different population parameters.

6.4 Raking Ratio Estimation

Raking ratio estimation was first discussed by Deming and Stephan (1940) where the auxiliary information is in the form of known marginal totals of a contingency table. The data set displayed in Table 6.2 is a five percent synthetic sample derived from the New England age-by-state table of the *Fifteenth Census of the U.S.*, 1930 (Deming and Stephan 1940). The first entry (among the three) in the (ij)th cell is the observed count n_{ij} for state i and age group j. The sample is taken by SRSWOR, and the goal is to estimate the population count N_{ij} for the (ij)th cell, $i = 1, \ldots, r$ and $j = 1, \ldots, c$. For this particular example, we have $r = 6$ and $c = 4$. The marginal totals $N_{i\cdot} = \sum_{j=1}^{c} N_{ij}$, $i = 1, \ldots, r$ for the states and $N_{\cdot j} = \sum_{i=1}^{r} N_{ij}$, $j = 1, \ldots, c$ for the age groups are known, and so is the population total $N = \sum_{i=1}^{r} \sum_{j=1}^{c} N_{ij}$.

The problem is equivalent to estimating the cell proportions $P_{ij} = N_{ij}/N$. The design unbiased estimator of P_{ij} is given by $Q_{ij} = n_{ij}/n$, where $n = \sum_{i=1}^{r} \sum_{j=1}^{c} n_{ij}$ is the total sample size. The estimators Q_{ij} usually do not match the known marginal proportions, i.e., $\sum_{j=1}^{c} Q_{ij}$ may differ from $N_{i\cdot}/N$ and $\sum_{i=1}^{r} Q_{ij}$ may not be the same as $N_{\cdot j}/N$. The classical raking ratio estimator is obtained through the so-called iterative proportional fitting procedure (IPFP) (Deming and

Table 6.2 Classical raking ratio estimation versus empirical likelihood method

Age		7–13	14 and 15	16 and 17	18–20	
j		1	2	3	4	
State	i					$nN_{i\cdot}/N$
Maine	1	3623	781	557	313	5252
		3613	781	550	308	
		3613	781	550	309	
New Hampshire	2	1570	395	251	155	2395
		1588	401	251	155	
		1588	401	251	155	
Vermont	3	1553	419	264	116	2432
		1608	435	270	119	
		1608	435	270	119	
Massachusetts	4	10,538	2455	1706	1160	15,766
		10,492	2452	1680	1141	
		10,492	2451	1681	1142	
Rhode Island	5	1681	353	171	154	2330
		1662	350	167	150	
		1662	350	167	151	
Connecticut	6	3882	857	544	339	5662
		3915	867	543	338	
		3915	867	543	338	
$nN_{\cdot j}/N$		22,877	5285	3462	2213	$n = 33,837$

Stephan 1940; Stephan 1942). The procedure starts from $Q_{ij}^{(0)} = Q_{ij}$ and obtains the results from the first iteration as

$$Q_{ij}^{(1)} = Q_{ij}^{(0)} \frac{N_{i\cdot}/N}{\sum_{k=1}^{c} Q_{ik}^{(0)}}\,, \quad i = 1, \ldots, r\,; \; j = 1, \ldots, c\,.$$

The $Q_{ij}^{(1)}$'s match the known marginal proportions $N_{i\cdot}/N$ for all the rows in the table. The next iteration $Q_{ij}^{(2)}$ is obtained by similar ratio adjustment to $Q_{ij}^{(1)}$ to match the known marginal proportions for all the columns. The process of raking ratio adjustment among the rows and the columns continues until convergence is achieved. Let $\hat{P}_{ij}$ be the final raking ratio estimator of P_{ij}.

Ireland and Kullback (1968) showed that the raking ratio estimators $\hat{P}_{ij}$ minimize the Kullback-Leibler information distance (Kullback and Leibler 1951)

$$I(\mathbf{P}, \mathbf{Q}) = \sum_{i=1}^{r} \sum_{j=1}^{c} P_{ij} \log\left(\frac{P_{ij}}{Q_{ij}}\right)$$

with respect to the P_{ij}'s subject to the constraints on matching the marginal population proportions. Since $\sum_{i=1}^{r} \sum_{j=1}^{c} P_{ij} = \sum_{i=1}^{r} \sum_{j=1}^{c} Q_{ij} = 1$, we can re-write $I(\mathbf{P}, \mathbf{Q})$ as

$$ET(\mathbf{P}, \mathbf{Q}) = \sum_{i=1}^{r} \sum_{j=1}^{c} \left\{ P_{ij} \log\left(\frac{P_{ij}}{Q_{ij}}\right) - P_{ij} + Q_{ij} \right\},$$

which is the same as the distance measure for the generalized exponential tilting with the q-factor as 1. The estimators $\hat{P}_{ij}$ can be obtained by minimizing $ET(\mathbf{P}, \mathbf{Q})$ subject to

$$\sum_{i=1}^{r} \sum_{j=1}^{c} P_{ij} = 1\,, \tag{6.23}$$

$$\sum_{j=1}^{c} P_{ij} = \frac{N_{i\cdot}}{N}\,, \quad i = 1, \ldots, r-1\,, \tag{6.24}$$

$$\sum_{i=1}^{r} P_{ij} = \frac{N_{\cdot j}}{N}\,, \quad j = 1, \ldots, c-1\,. \tag{6.25}$$

The computational algorithm presented in Sect. 6.2.1 can be used directly to find $\hat{P}_{ij}$ if we combine constraints (6.24) and (6.25) into

$$\sum_{i=1}^{r}\sum_{j=1}^{c} P_{ij}\mathbf{x}_{ij} = \mu_{\mathbf{x}}\,, \tag{6.26}$$

where $\mathbf{x}_{ij}$ is the $[(r-1)+(c-1)]\times 1$ vector of the first $r-1$ row and the first $c-1$ column indicator variables, and

$$\mu_{\mathbf{x}} = \big(N_{1\cdot},\ \ldots,\ N_{(r-1)\cdot}\,, N_{\cdot 1}\,, \ldots, N_{\cdot(c-1)}\big)'/N$$

is the vector of marginal population proportions.

The raking ratio estimators may be replaced by a calibration estimator obtained by minimizing the pseudo empirical likelihood distance

$$EL(\mathbf{P},\mathbf{Q}) = \sum_{i=1}^{r}\sum_{j=1}^{c}\Big\{Q_{ij}\log\Big(\frac{Q_{ij}}{P_{ij}}\Big) - Q_{ij} + P_{ij}\Big\} = \sum_{i=1}^{r}\sum_{j=1}^{c} Q_{ij}\log\Big(\frac{Q_{ij}}{P_{ij}}\Big)$$

under the constraints (6.23) and (6.26). The algorithm discussed in Sect. 6.2.2 is suitable for computing the final estimators of P_{ij}. The second entries in each of cells in Table 6.2 are the raking ratio estimates of nP_{ij} and the third entries are the calibration estimates of nP_{ij} using the EL distance. The results are almost identical for the two methods.

The method of raking ratio adjustment of Deming and Stephan (1940) was the first formal approach to using auxiliary population information, with the primary objective of achieving internal consistency. Brackstone and Rao (1979) contained a study on classic raking ratio estimation. The discussions in this section show that raking ratio estimation can be treated as a special case of calibration methods. Raking ratio estimation with high dimensional contingency tables with known marginal population totals can be handled in a similar manner. Generalized raking estimation under unequal probability sampling designs can also be achieved through calibration. See Deville et al. (1993) and Rao and Wu (2009) for further discussions.

6.5 Additional Remarks

Calibration weighting and estimation has been a routine practice at many survey organizations. Several statistical agencies have developed standard softwares to carry out the computational tasks. However, “Model-free calibration approach can lead to erroneous inferences for some of the response variables, even in fairly large samples” (Rao 2011). When there are a large number of auxiliary variables which could be used for calibration, blindly including all of them for calibration weighting can result in weights which are highly variable. Extremely high variation in survey weights is one of the main causes for the inefficiency of design-based inferences for some of the response variables. A judiciously selected set of auxiliary variables to be used for calibration is always the first step towards a set of less variable calibration weights.

The calibrated weights are built based on the initial set of basic design weights. The validity of inferences using calibrated weights relies on the validity of the initial weights. Calibration weighting is a useful tool to achieve internal consistency and estimation efficiency but calibration techniques should not be applied to scenarios where the initial weights do not provide valid inferences.

Range-restricted calibration is an effective way to reduce the variation in the final weights and to remove extremely large and small weights. The weight trimming method discussed in Sect. 6.2.4 is a simple and robust approach to produce a final set of weights that can be used for general design-based inferences. The basic design weights could be a major contributor to the variation in the calibration weights. Some research has been carried out on how to trim survey weights and the impact of weight trimming on survey data analysis. See, for instance, Potter (1990) and Chowdhury et al. (2007). Any attempts to directly trim or alter the basic design weights, however, need to be done with extra caution: design-consistency of survey-weighted estimators using the trimmed or altered weights could be lost. Variance estimation and confidence intervals using range-restricted calibration weights is an important topic that requires further research. Modified bootstrap methods similar to those described in Chap. 10 which take into account of range-restrictions are possible approaches for computing valid variance estimates.

The model-calibration approach to using complete auxiliary information (Sect. 6.3) focuses mainly on estimation efficiency. Nonparametric model-calibration methods were discussed by Montanari and Ranalli (2005). Other related work on calibration includes the local polynomial regression estimators (Breidt and Opsomer 2000) and estimators using penalized splines (Breidt et al. 2005). Under the semiparametric model specified by (6.16), the optimality results on model-calibration estimators indicate that optimal estimation of the population total only involves the mean function $\mu(\mathbf{x}, \theta)$ but optimal estimation of the population distribution function and quantiles as well as other quadratic parameters such as population variance requires both the mean function $\mu(\mathbf{x}, \theta)$ and the variance function $v(\mu)$.

There have been developments on using calibration as a tool for handling missing data, a topic to be discussed in Chap. 9. See, for instance, Lundström and Särndal (1999), Särndal and Lundström (2005), Kott (2003), Chang and Kott (2008) and Kott and Chang (2010). There have also been developments on using model-calibration for doubly robust and multiply robust inference for missing data problems. The auxiliary information is used under a two-phase setting for model-calibration (Wu and Luan 2003), and multiple models for the response variables and the missing data mechanism can be incorporated through model-calibrated constraints. Han and Wang (2013), Han (2014) and Wu and Zhang (2019) provide detailed discussions and additional references on the topic. The review papers by Särndal (2007), Kim and Park (2010), Lumley et al. (2011) and Wu and Lu (2016) contain further materials and references on calibration methods.

6.6 Problems

6.1 (Calibration Estimators Under the Chi-Square Distance) Use the Lagrange multiplier method to find the solution of minimizing $\chi^2(\mathbf{w}, \mathbf{d})$ subject to constraints (6.1):

(a) Write down the corresponding equations (6.2) and (6.3) under the current setting.
(b) Show that the calibration weights w_i are given by (6.6).

6.2 (Ratio Estimator as a Calibration Estimator) Consider a single auxiliary variable $x > 0$ with known population total $T_x = \sum_{i=1}^N x_i$. Show that the calibration estimator $\hat{T}_{y\mathrm{C}}$ of T_y under the chi-square distance $\chi^2(\mathbf{w}, \mathbf{d})$ with $q_i = x_i^{-1}$ reduces to the generalized ratio estimator discussed in Sect. 5.3.

6.3 (Optimal Calibration Estimator Under Poisson Sampling) Suppose that the sample $\mathbf{S}$ is selected by Poisson sampling with first order inclusion probabilities π_i. Consider the GREG-type estimators of T_y in the form of

$$\hat{T}_y = \hat{T}_{y\mathrm{HT}} + \mathbf{B}'\left(T_{\mathbf{x}} - \hat{T}_{\mathbf{x}\mathrm{HT}}\right),$$

where $\mathbf{B}$ is a constant vector with the same dimension as the vector $\mathbf{x}$ of auxiliary variables.

(a) Find the optimal $\mathbf{B}_{opt}$ such that the design-based variance $V_p\left(\hat{T}_y\right)$ is minimized among all choices of $\mathbf{B}$.
(b) Show that the calibration estimator $\hat{T}_{y\mathrm{C}}$ given by (6.7) under the chi-square distance with $q_i = \pi_i^{-1} - 1$ is asymptotically optimal among the class of GREG-type estimators.

6.4 (Calibration Weighting with Kullback-Leibler Distance)

(a) Show that the calibration weights obtained by minimizing $ET(\mathbf{w}, \mathbf{d})$ subject to constraints (6.1) are given by $w_i = d_i g_i$, where $g_i = \exp(\lambda_2' \mathbf{x}_i q_i)$ and λ_2 satisfies (6.10).
(b) Show that the calibration weights obtained by minimizing $EL(\mathbf{w}, \mathbf{d})$ subject to constraints (6.1) are given by $w_i = d_i g_i$, where $g_i = (1 - \lambda_3' \mathbf{x}_i q_i)^{-1}$ and λ_3 satisfies (6.13).
(c) Develop computing programs for finding the solutions to (6.10) and (6.13), and test the programs through a simulation study.

6.5 (Range-Restricted Calibration) Consider a calibration weighting method which produces positive calibration weights w_i. Suppose that we replace the standard benchmark constraints (6.1) by a relaxed form

$$\sum_{i \in \mathbf{S}} w_i \mathbf{x}_i = T_{\mathbf{x}} + \delta\left(\hat{T}_{\mathbf{x}\mathrm{HT}} - T_{\mathbf{x}}\right) \tag{6.27}$$

for some $\delta \in [0, 1]$.

(a) Show that, for any pre-specified $c_1 \in (0, 1)$ and $c_2 > 1$, there exists a δ such that the calibration weights w_i obtained by minimizing the distance measure subject to the relaxed constraints (6.27) satisfy the range-restriction (6.15).

(b) Describe a computational method for finding the smallest possible value δ for the given c_1 and c_2.

6.6 (Model-Calibration Under a Linear Regression Model) Consider a linear regression model with an intercept where $\mu_i = \mu(\mathbf{x}_i, \theta) = \mathbf{z}_i'\theta$ and $\mathbf{z}_i = (1, \mathbf{x}_i')'$. The survey-weighted estimator $\hat{\theta}$ of θ is given in (6.19). The q-factor in the chi-square distance is set as $q_i = 1$. The definitions of $\hat{B}_0$, $\hat{B}_1$ and $\hat{B}_1^*$ are given in Sect. 6.3.1. Let $\hat{\mu}_i = \mathbf{z}_i'\hat{\theta}$.

(a) Show that $\sum_{i\in\mathbf{S}} d_i\hat{\mu}_i y_i = \sum_{i\in\mathbf{S}} d_i\hat{\mu}_i^2$, which implies $\hat{B}_1^* = 1$.

(b) Show that $\sum_{i\in\mathbf{S}} d_i y_i = \sum_{i\in\mathbf{S}} d_i\hat{\mu}_i$, which further leads to $\hat{B}_1 = 1$ and $\hat{B}_0 = 0$.

6.7 (Model-Calibration Estimator of the Population Variance) Consider the linear regression model where $\mu_i = E_\xi\big(y_i \mid \mathbf{x}_i\big) = \mathbf{x}_i'\beta$ and consider the model-calibration estimator of the finite population variance σ_y^2 discussed in Sect. 6.3.2.

(a) Derive the final simplified form of $\hat{\sigma}_{y\mathrm{MC}}^2$ under the chi-square distance and the single constraint (6.21) for a general survey sample with inclusion probabilities π_i and π_{ij}.

(b) Derive the final simplified form of $\hat{\sigma}_{y\mathrm{MC}}^2$ under the chi-square distance and the single constraint (6.21) when the sample is selected by SRSWOR.

(c) Conduct a simulation study to compare the performance of $\hat{\sigma}_{y\mathrm{MC}}^2$ under SRSWOR to the simple estimator of σ_y^2 using the sample variance s^2.

Chapter 7
Regression Analysis and Estimating Equations

Survey samples are selected from a finite population. The survey population is typically a well-defined real world population, and main objectives of surveys often focus on descriptive characteristics of the survey population. Design-based inferences are the dominant approach for survey data analysis under such scenarios.

Statistical models and inferences on model parameters using complex survey data have become an important topic for survey research. It has become increasingly common for researchers in social, health and medical sciences to collect data through complex surveys, and the inferential objectives are to explore general or causal relationships among variables and to estimate model parameters or effect of treatment. The terms "*analytical surveys*" and "*analytic uses of survey data*" are often used in this context (Thompson 1997).

Estimation of model parameters using a survey sample is not as simple as applying standard techniques. Consider the simple linear regression model without an intercept,

$$y = x\beta + \varepsilon\,, \quad E_\xi(\varepsilon) = 0 \quad \text{and} \quad V_\xi(\varepsilon) = x\sigma^2\,, \tag{7.1}$$

where ξ denotes the model and $x > 0$. The parameters of interest are β and σ^2. Suppose that $\{(y_i, x_i), i = 1, \ldots, n\}$ is a random sample from the model,

$$y_i = x_i\beta + \varepsilon_i\,, \quad i = 1, \ldots, n\,, \tag{7.2}$$

where $E_\xi(\varepsilon_i) = 0$, $V_\xi(\varepsilon_i) = x_i\sigma^2$ and $\varepsilon_1, \ldots, \varepsilon_n$ are independent. It can be shown that the weighted least square estimator of β is given by $\hat{\beta} = \bar{y}/\bar{x}$, the ratio of two sample means, with basic properties $E_\xi(\hat{\beta}) = \beta$ and $V_\xi(\hat{\beta}) = n^{-1}\sigma^2/\bar{x}$ under model (7.1). An unbiased estimator for σ^2 is given by $\hat{\sigma}^2 = (n-1)^{-1}\sum_{i=1}^n (y_i - x_i\hat{\beta})^2/x_i$, which leads to an unbiased variance estimator for $\hat{\beta}$ given by $v(\hat{\beta}) = n^{-1}\hat{\sigma}^2/\bar{x}$.

C. Wu, M. E. Thompson, *Sampling Theory and Practice*, ICSA Book
Series in Statistics, https://doi.org/10.1007/978-3-030-44246-0_7

Suppose that $\{(y_i, x_i), i \in \mathbf{S}\}$ is obtained through a survey and the survey sample $\mathbf{S}$ is selected by a PPS sampling design (without replacement) with inclusion probabilities $\pi_i \propto x_i$. To make inferences on the regression coefficient β, one approach used in practice is to assume that the *sampling design is ignorable*, i.e., the survey data $\{(y_i, x_i), i \in \mathbf{S}\}$ follow the same model (7.2), and use the same weighted least square estimator $\hat{\beta} = \bar{y}/\bar{x}$.

Ignorability of survey design features for inferences on model parameters is difficult to verify in practice. In the simple example of regression modelling discussed above, one might argue that the conditional moment structure of the model (7.2) still holds for the survey sample data, i.e., $E_\xi(y_i \mid x_i) = x_i\beta$ and $V_\xi(y_i \mid x_i) = x_i\sigma^2$, $i \in \mathbf{S}$. If this is the case, the estimator $\hat{\beta} = \bar{y}/\bar{x}$ ignoring the survey design is still unbiased for β. However, it is easy to argue that the assumption "$\varepsilon_1, \ldots, \varepsilon_n$ are independent" could be violated, which implies that the estimator $\hat{\sigma}^2$ might be biased for σ^2. Furthermore, units with large values in x are more likely to be included in the survey sample under the PPS sampling design, and the simple sample mean $\bar{x}$ is likely to have bigger values as compared to random samples from model (7.1). It follows that the variance estimator $v(\hat{\beta}) = n^{-1}\hat{\sigma}^2/\bar{x}$ becomes problematic since the ignorability assumption on the design is questionable.

There exist several alternative approaches to inferences on statistical models using complex survey data. This chapter discusses a commonly used approach which combines model-based inferences and design-based estimation under a two-stage joint randomization framework. The formulation of problems starts with a general statistical model but the estimation procedures are mostly design-based. Both the model structure and the survey design features are handled through the use of survey weighted estimating equations.

7.1 Parameters of Superpopulations and Survey Populations

Let y be the study variable and $\mathbf{x}$ be the vector of covariates. Suppose that $(y, \mathbf{x})$ follows a superpopulation model ξ, represented by a class of distributions indexed by the parameter $\theta \in \Theta$, where Θ denotes the parameter space. We consider y to be one-dimensional in Sects. 7.1–7.3 of this chapter. Sect. 7.4 contains discussions on longitudinal and clustered data where the study variables are vector-valued. The model parameter θ is a $k \times 1$ vector with $k \geq 1$.

7.1.1 Model Parameters and Estimating Functions

We consider a $r \times 1$ vector-valued *estimating function* $\mathbf{g}(y, \mathbf{x}; \theta)$ which involves random variables y and $\mathbf{x}$ as well as the parameter θ. It is called an *unbiased estimating function* if

$$E_\xi\big\{\mathbf{g}(y,\mathbf{x};\theta)\big\} = \mathbf{0}\,, \quad \theta \in \Theta\,. \tag{7.3}$$

Under certain models such as those for regression analysis, the estimating function $\mathbf{g}(y, \mathbf{x}; \theta)$ might be conditionally unbiased:

$$E_\xi\big\{\mathbf{g}(y,\mathbf{x};\theta) \mid \mathbf{x}\big\} = \mathbf{0}\,, \quad \theta \in \Theta\,. \tag{7.4}$$

Conditionally unbiased estimating functions are also (unconditionally) unbiased, since equations (7.4) imply (7.3). In this section and the next section, we consider scenarios where $k = r$, i.e., the number of unknown parameters equals to the number of components in the estimating function. Scenarios with $k < r$ are discussed in Chap. 8. It is of little practical interest to consider scenarios with $k > r$ since the parameters are unidentifiable for these cases.

Some commonly encountered parameters in statistical inferences can be defined through unbiased estimating functions.

1. The estimating function $g(y, \theta) = y - \theta$ defines the population mean $\theta = E_\xi(y)$ for the study variable y.
2. The estimating function $\mathbf{g}(y, \theta) = \big(y - \theta_1, (y - \theta_1)^2 - \theta_2\big)'$ for $\theta = (\theta_1, \theta_2)'$ defines the population mean $\theta_1 = E_\xi(y)$ and the population variance $\theta_2 = V_\xi(y)$ for the study variable y.
3. The distribution function of y for a given t, i.e., $\theta = F(t) = P(y \le t)$, is defined through $g(y, \theta) = I(y \le t) - \theta$, where $I(\cdot)$ is the indicator function.
4. The 100γth quantile θ of y is defined through the estimating function $g(y, \theta) = I(y \le \theta) - \gamma$, where $\gamma \in (0, 1)$.
5. For a univariate x variable, the estimating function $g(y, x; \theta) = y - \theta x$ defines the ratio $\theta = E_\xi(y)/E_\xi(x)$, assuming $E_\xi(x) \neq 0$.
6. The estimating function $\mathbf{g}(y, \mathbf{x}; \theta) = \mathbf{x}(y - \mathbf{x}'\theta)$ defines the unweighted version of regression coefficients θ. If the regression model contains an intercept, the $\mathbf{x}$ variables are extended to have 1 as the first component.
7. Regression coefficients θ for a semiparametric model under the framework of generalized linear models can be specified by

$$\mathbf{g}(y,\mathbf{x};\theta) = \mathbf{D}(\mathbf{x};\theta)\big\{V(\mu(\mathbf{x},\theta))\big\}^{-1}\big\{y - \mu(\mathbf{x},\theta)\big\}\,, \tag{7.5}$$

where $\mu(\cdot,\cdot)$ and $V(\cdot)$ are the mean function and the variance function, and $\mathbf{D}(\mathbf{x}; \theta)$ is typically specified as $\partial\mu(\mathbf{x}, \theta)/\partial\theta$.

Other advanced examples of estimating functions include the extended quasi-score function (Godambe and Thompson 1989) and estimation of parameters for quantile regression models (Koenker and Bassett 1978).

Suppose that $\{(y_i, \mathbf{x}_i), i = 1, \ldots, n\}$ is a random sample from the model ξ. The model parameter θ can be estimated by $\hat{\theta}$ which is the solution (assuming it uniquely exists) to the following *estimating equations*:

$$\mathbf{G}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{g}(y_i, \mathbf{x}_i; \theta) = \mathbf{0}\,. \tag{7.6}$$

Note that the number of equations in (7.6) is the same as the number of unknown parameters in θ, and the "redundant" factor $1/n$ is for facilitating asymptotic development.

The general theory of estimating functions provides a unification of two of the principal estimation methods in statistics, namely, the least squares method and the maximum likelihood method. Minimizing the sum of squares of errors or maximizing the likelihood function often leads to solving a system of equations. Some of the early research on estimating functions followed Godambe's (1960) paper on the properties of the score function, and explored optimality properties of specific forms of estimating functions. For a scalar θ and without the presence of the covariates $\mathbf{x}$, an unbiased estimating function $g^* = g^*(y, \theta)$ is said to be optimal if

$$E_\xi\{(g^*)^2\}/\{E_\xi(\partial g^*/\partial\theta)\}^2 \le E_\xi\{(g)^2\}/\{E_\xi(\partial g/\partial\theta)\}^2\,, \quad \theta \in \Theta$$

holds for any unbiased estimating function $g = g(y, \theta)$. The book edited by Godambe (1991a) contains a rich collection of papers on many theoretical and applied aspects of estimating functions.

An important development in estimating functions is the theory of m-estimators. Let $\mathbf{g}(y, \mathbf{x}; \theta)$ be an estimating function and let $\mathbf{G}(\theta) = E_\xi\{\mathbf{g}(y, \mathbf{x}; \theta)\}$. Let θ_0 be the true value of the parameters defined as the solution to $\mathbf{G}(\theta) = \mathbf{0}$. The so-called m-estimator $\hat{\theta}$ of θ_0 based on the random sample $\{(y_i, \mathbf{x}_i), i = 1, \ldots, n\}$ is simply the solution to $\mathbf{G}_n(\theta) = \mathbf{0}$ as specified by the estimating equations (7.6). Consistency of the m-estimators and the required regularity conditions are covered by Theorem 2.1 and Lemma 2.4 in Newey and McFadden (1994). Asymptotic normality of the m-estimators and the additional required regularity conditions can be found from Theorem 5.41 in van der Vaart (2000). Section 3.2 of Tsiatis (2006) also contains a useful review on m-estimators.

7.1.2 Survey Population Parameters

The finite population $\mathbf{U} = \{1, 2, \ldots, N\}$ of size N presents a "census data set" $\{(y_i, \mathbf{x}_i), i = 1, 2, \ldots, N\}$ which can be viewed as a random sample of size N from a superpopulation model ξ. If the model parameter θ is defined through the unbiased estimating function $\mathbf{g}(y, \mathbf{x}; \theta)$, we may define, at least conceptually, an "estimator" θ_N for θ based on the census data set as the solution to the *census estimating equations*:

$$\mathbf{G}_N(\theta) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{g}(y_i, \mathbf{x}_i; \theta) = \mathbf{0}\,. \tag{7.7}$$

In other words, the "estimator" θ_N satisfies $\mathbf{G}_N(\theta_N) = \mathbf{0}$ and possesses two properties simultaneously (Godambe and Thompson 1986):

(i) If the census data set is known, the θ_N is an estimate of the model parameter θ for the superpopulation.
(ii) If the census data set is unknown, the θ_N defines a parameter for the survey population.

It turns out that many parameters of the survey population, either descriptive measures or for analytic uses, can be defined as the solution to a set of census estimating equations. For instance, the estimating function $g(y, \theta) = y - \theta$ and $\mathbf{G}_N(\theta_N) = \mathbf{0}$ defines the finite population mean $\theta_N = \mu_y = N^{-1}\sum_{i=1}^{N} y_i$ for the study variable y. The estimating function $\mathbf{g}(y, \mathbf{z}; \theta) = \mathbf{z}(y - \mathbf{z}'\theta)$ and $\mathbf{G}_N(\theta_N) = \mathbf{0}$ defines θ_N as the *population regression coefficients*; see Sect. 7.3.1 for further detail.

The definition of a survey population parameter θ_N through census estimating equations is mainly for the estimation theory to be presented in the rest of the chapter. The population total $T_y = \sum_{i=1}^{N} y_i$, one of the primary descriptive parameters, does not fit into the current framework and plays no crucial role in analytic uses of surveys. From a computational point of view, the solutions to the estimating equations $\mathbf{G}_N(\theta_N) = \mathbf{0}$ may have (i) an explicit closed form expression; (ii) multiple roots or no feasible answers; or (iii) an implicit form which requires iterative procedures for computation. This is explained further in Sect. 7.2.1.

Analytic uses of survey data generally refer to inferences on parameters of a superpopulation model or survey population parameters whose form is motivated by such a model (Godambe and Thompson 1986, 2009). The census estimating equations formulation makes it possible to focus on the survey population parameters θ_N using design-based estimation methods. When the superpopulation model ξ is valid and the survey population follows the model, the survey population parameter θ_N would be very close to the model parameter θ, especially for large N. Under this scenario estimating θ_N is effectively estimating θ. When the superpopulation model fails, rendering the model parameter meaningless, the survey population parameter θ_N might still be a meaningful target for design-based inference. For instance, the simple linear regression model (7.1) might be misspecified, but the estimating function for the weighted least squares estimator of β remains as $g(y, x; \beta) = x(y - x\beta)/v = y - x\beta$ for $v = x$. The census estimating equation $N^{-1}\sum_{i=1}^{N}(y_i - x_i\beta) = 0$ leads to $\beta_N = \mu_y/\mu_x$, the ratio of two population means for the variables y and x.

7.2 Survey Weighted Estimating Equations

Let $\{(y_i, \mathbf{x}_i), i \in \mathbf{S}\}$ be a survey sample dataset where the sample $\mathbf{S}$ is collected through a probability sampling design with the first and the second order inclusion probabilities π_i and π_{ij}. The objective of this section is to develop point estimators and confidence intervals/regions for the survey population parameter θ_N as well as the superpopulation model parameter θ.

The function $\mathbf{G}_N(\theta)$ in the census estimating equations (7.7) can be viewed as a population mean defined over the variable $\mathbf{g}_i = \mathbf{g}(y_i, \mathbf{x}_i; \theta)$ with the given θ. The design-based Horvitz-Thompson estimator of $\mathbf{G}_N(\theta)$ is given by $\mathbf{G}_n(\theta) = N^{-1}\sum_{i\in\mathbf{S}} d_i\mathbf{g}_i$ where $d_i = 1/\pi_i$ are the basic design weights. This leads naturally to the following *survey weighted estimating equations*:

$$\mathbf{G}_n(\theta) = \frac{1}{N}\sum_{i\in\mathbf{S}} d_i\,\mathbf{g}(y_i, \mathbf{x}_i; \theta) = \mathbf{0}\,. \tag{7.8}$$

For simplicity of notation, we used the same notation $\mathbf{G}_n(\theta)$ as in (7.6), which corresponds to the special case of $d_i = N/n$. It is apparent that the factor $1/N$ is not needed for computation and is included for theoretical considerations. An important result on survey weighted estimating equations is given by Theorem 1 of Godambe and Thompson (1986) with the notation $g(y_i, \theta)$, which states that, if $\sum_{i=1}^N g(y_i, \theta)$ is the linear optimal estimating function under the superpopulation model, then $\sum_{i\in\mathbf{S}} d_i g(y_i, \theta)$ is linear optimal under the joint randomization of both the model ξ and the probability survey design p.

7.2.1 *Point Estimators*

Let $\hat{\theta}$ be a solution to $\mathbf{G}_n(\theta) = \mathbf{0}$ given by (7.8). Molina and Skinner (1992) called $\hat{\theta}$ the maximum pseudo-likelihood estimator when fitting a generalized linear model with the estimating function given by (7.5). We first focus on design-based properties of $\hat{\theta}$ as an estimator of θ_N. Under suitable regularity conditions given below on the $\mathbf{g}(y_i, \mathbf{x}_i; \theta)$ and the survey design, we show that $\hat{\theta}$ is design-consistent and asymptotically normally distributed. For design-based asymptotic framework, see discussions in Sect. 2.4.

C8.1 The estimating function $\mathbf{g}(y_i, \mathbf{x}_i; \theta)$ is twice differentiable with respect to θ.

C8.2a The survey design and the finite population satisfy $\mathbf{G}_n(\theta) - \mathbf{G}_N(\theta) = O_p\big(n^{-1/2}\big)$ for any fixed $\theta \in \Theta$.

C8.2b The survey design and the finite population ensure that $\mathbf{G}_n(\theta)$ is asymptotically normally distributed with mean $\mathbf{G}_N(\theta)$ and entries of the variance-covariance matrix at the order $O\big(n^{-1}\big)$ for any fixed $\theta \in \Theta$.

C8.3 The survey design, the finite population and the estimating function satisfy $\partial\mathbf{G}_n(\theta)/\partial\theta = O_p(1)$ and $\partial^2\mathbf{G}_n(\theta)/\partial\theta\partial\theta' = O_p(1)$ for any fixed $\theta \in \Theta$.

Conditions C8.2a and C8.2b are the usual anticipated properties of the Horvitz-Thompson estimator $N^{-1}\sum_{i\in\mathbf{S}} d_i y_i$, with the variable y_i replaced by $\mathbf{g}_i = \mathbf{g}(y_i, \mathbf{x}_i; \theta)$. Condition C8.2b implies C8.2a. Condition C8.3 is also a typical property assumed for the Horvitz-Thompson estimator, i.e., $N^{-1}\sum_{i\in\mathbf{S}} d_i y_i = O_p(1)$, with the variable y_i replaced by $\partial\mathbf{g}_i/\partial\theta$ and $\partial^2\mathbf{g}_i/\partial\theta\partial\theta'$. For vector forms of θ and $\mathbf{g}_i$, all conditions are component-wise.

Theorem 7.1 *Under a suitable asymptotic framework and the regularity conditions C8.1 and C8.3 for the estimating function, the sampling design and the finite population:*

(a) With the additional condition C8.2a, the survey weighted estimating equation estimator $\hat{\theta}$ satisfies $\hat{\theta} - \theta_N = O_p\big(n^{-1/2}\big)$.

(b) With the additional condition C8.2b, the survey weighted estimating equation estimator $\hat{\theta}$ is asymptotically normally distributed with mean θ_N and variance-covariance matrix

$$\big\{\mathbf{H}_N(\theta_N)\big\}^{-1} V_p\big\{\mathbf{G}_n(\theta_N)\big\}\big\{\mathbf{H}'_N(\theta_N)\big\}^{-1}, \tag{7.9}$$

where $V_p\{\mathbf{G}_n(\theta_N)\}$ is the design-based variance-covariance matrix of the Horvitz-Thompson estimator $\mathbf{G}_n(\theta_N)$ and $\mathbf{H}_N(\theta) = N^{-1}\sum_{i=1}^N \partial\mathbf{g}(y_i, \mathbf{x}_i; \theta)/\partial\theta$.

Proof We sketch the proof of Part (a) for a scalar θ. The vector form involves the multivariate version of Taylor series expansions with similar technical arguments. Consider the following Taylor series expansion,

$$G_n(\hat{\theta}) - G_n(\theta_N) = \Big[\frac{\partial}{\partial\theta}G_n(\theta)\Big]_{\theta=\theta_N}\big(\hat{\theta} - \theta_N\big) + \frac{1}{2}\Big[\frac{\partial^2}{\partial\theta^2}G_n(\theta)\Big]_{\theta=\theta^*}\big(\hat{\theta} - \theta_N\big)^2,$$

where θ^* is a point between $\hat{\theta}$ and θ_N. By definition, we have $G_n(\hat{\theta}) = 0$. Condition C8.2a plus the definition of θ_N through $G_N(\theta_N) = 0$ implies that $G_n(\theta_N) = O_p\big(n^{-1/2}\big)$. It follows from Condition C8.3 that we must have $\hat{\theta} - \theta_N = O_p\big(n^{-1/2}\big)$.

The result of Part (a) and the Taylor series expansion lead to the asymptotic approximation of the survey weighted estimating equation estimator,

$$\hat{\theta} = \theta_N - \big\{\mathbf{H}_N(\theta_N)\big\}^{-1}\mathbf{G}_n(\theta_N) + O_p\big(n^{-1}\big), \tag{7.10}$$

where $\mathbf{H}_N(\theta) = \mathbf{H}_n(\theta) + O_p\big(n^{-1/2}\big)$ and

$$\mathbf{H}_n(\theta) = \partial\mathbf{G}_n(\theta)/\partial\theta = N^{-1}\sum_{i\in\mathbf{S}} d_i\big\{\partial\mathbf{g}(y_i, \mathbf{x}_i; \theta)/\partial\theta\big\}.$$

Asymptotic normality of $\hat{\theta}$ follows from the asymptotic normality of $\mathbf{G}_n(\theta_N)$ under Condition C8.2b and the expansion (7.10), with the asymptotic design-based variance-covariance matrix of $\hat{\theta}$ given by (7.9). Note that $\mathbf{H}'_N(\theta_N)$ in (7.9) represents the transpose of $\mathbf{H}_N(\theta_N)$. □

Thompson (1997) defines *design-consistency* of the estimator $\hat{\theta}$ for the parameter θ_N as "$(\hat{\theta} - \theta_N)/\theta_N$ converges to zero in probability". This is necessary since certain finite population parameters such as population totals may have different order of magnitude. For population parameters defined through census estimating equations, we usually have $\theta_N = O(1)$, and consequently design-consistency is equivalent to the traditional definition "$\hat{\theta} - \theta_N$ converges to zero in probability" or "$\hat{\theta} - \theta_N = o_p(1)$". The result from Part (a) of the theorem is stronger than $\hat{\theta}$ being design-consistent. Section 4.2 of Thompson (1997) contains materials on population quantities defined as functions of population totals.

There are three possible scenarios for $\hat{\theta}$ from solving $\mathbf{G}_n(\theta) = \mathbf{0}$: (i) an explicit closed form expression for $\hat{\theta}$; (ii) an implicit expression defined through iterative procedures; and (iii) no solution or multiple roots to $\mathbf{G}_n(\theta) = \mathbf{0}$, requiring further refinement to define a unique solution. Regression coefficients for a linear model are an example for scenario (i); see Sect. 7.3.1 for further detail. Scenario (ii) requires the estimating function $\mathbf{g}(y_i, \mathbf{x}_i; \theta)$ to be smooth over θ. The approximation formula (7.10) can be used to build the Newton-Raphson iterative procedures:

$$\theta^{(t+1)} = \theta^{(t)} - \left\{\mathbf{H}_n(\theta^{(t)})\right\}^{-1}\mathbf{G}_n(\theta^{(t)}) , \quad t = 0, 1, 2, \ldots . \tag{7.11}$$

It is the computable sample-based $\mathbf{H}_n(\theta)$, not the conceptual population-based $\mathbf{H}_N(\theta)$, that is used in the iteration. The initial value $\theta^{(0)}$ needs to be chosen for the particular problem at hand. See Sect. 7.3.2 for further discussion. If the population size N is unknown, the factor N^{-1} should be dropped from both $\mathbf{H}_n(\theta)$ and $\mathbf{G}_n(\theta)$.

The technical developments with Theorem 7.1 do not apply to scenarios where the estimating function $\mathbf{g}(y_i, \mathbf{x}_i; \theta)$ is non-differentiable with respect to θ. The population quantiles are a primary example. The estimating function for defining the 100γth quantile is $g(y, \theta) = I(y \le \theta) - \gamma$, where $\gamma \in (0, 1)$. The census estimating equation for defining the 100γth finite population quantile θ_N is given by

$$G_N(\theta) = \frac{1}{N}\sum_{i=1}^{N}\left\{I(y_i \le \theta) - \gamma\right\} = F_y(\theta) - \gamma = 0 ,$$

where $F_y(t)$ is the finite population distribution function defined in Sect. 4.2.4. The equation $F_y(\theta) - \gamma = 0$ for a given $\gamma \in (0, 1)$ has either no solution or an infinite number of solutions since $F_y(t)$ is a nondecreasing step function. The 100γth population quantile is defined (Sect. 4.2.4) as $\theta_N = t_\gamma = \inf\{\theta \mid F_y(\theta) - \gamma \ge 0\}$. It can be seen that $G_N(\theta_N) = O(N^{-1})$ holds for all scenarios. The survey weighted estimating equation is given by

$$G_n(\theta) = \frac{1}{N}\sum_{i\in\mathbf{S}} d_i\big\{I(y_i \le \theta) - \gamma\big\} = 0\,,$$

which is equivalent to $\hat{F}_{y\mathrm{H}}(\theta)-\gamma = 0$, where $\hat{F}_{y\mathrm{H}}(t) = \sum_{i\in\mathbf{S}} d_i I(y_i \le t)/\sum_{i\in\mathbf{S}} d_i$ is the Hájek estimator of $F_y(t)$. The equation $\hat{F}_{y\mathrm{H}}(\theta)-\gamma = 0$ has either no solution or an infinite number of solutions. The estimating equation based estimator of $\theta_N = t_\gamma$ for the given γ can be uniquely defined as $\hat{\theta} = \hat{t}_\gamma = \inf\{\theta \mid \hat{F}_{y\mathrm{H}}(\theta) - \gamma \ge 0\}$.

Asymptotic theories for non-differentiable estimating functions are more challenging. For the population quantiles, the estimator is an inversion of the estimated distribution function. The estimating function for the distribution function $F_y(t)$ at a fixed t has the form $g(y,\theta) = I(y \le t)-\theta$ and is differentiable with respect to θ. A natural approach is to establish a Bahadur representation for the "survey weighted" quantile estimator $\hat{\theta} = \hat{t}_\gamma$ to link with the "differentiable" $g(y,\theta) = I(y \le t) - \theta$ for defining $\hat{F}_{y\mathrm{H}}(t)$:

$$\hat{\theta} - \theta_N = \frac{1}{f(\theta_N)}\Big\{\hat{F}_{y\mathrm{H}}(\theta_N) - F_y(\theta_N)\Big\} + o_p\big(n^{-1/2}\big)\,, \tag{7.12}$$

where $f(t)$ is the density function for the limiting distribution function of $F_y(t)$ as $N \to \infty$. Francisco and Fuller (1991) have given sufficient conditions for the representation (7.12) to hold. Further discussions can be found in Thompson (1997, Sect. 4.1.4) and Chen and Wu (2002). Wang and Opsomer (2011) contains materials on other non-differentiable survey estimators and their asymptotic properties.

The survey weighted estimating equation estimator $\hat{\theta}$ can also be used to estimate the superpopulation model parameter θ. For theoretical development, it is necessary to consider the joint randomization framework under both the model ξ and the probability sampling design p (Thompson 1984). In addition to the linear optimality property of $\mathbf{G}_n(\theta)$ as demonstrated by Godambe and Thompson (1986), we also have $\hat{\theta} - \theta = O_p(n^{-1/2})$ under the joint ξp framework with suitable regularity conditions. This is clearly the case if the "census estimator" θ_N for the model parameter θ satisfies $\theta_N - \theta = O_p(N^{-1/2})$ under the model and the survey weighted estimating equation estimator $\hat{\theta}$ for the finite population parameter θ_N satisfies $\hat{\theta} - \theta_N = O_p(n^{-1/2})$ under the sampling design, due to the decomposition of the bias $\hat{\theta} - \theta = (\hat{\theta} - \theta_N) + (\theta_N - \theta)$.

7.2.2 Design-Based Variance Estimation

The Horvitz-Thompson estimator $\mathbf{G}_n(\theta_N) = N^{-1}\sum_{i\in\mathbf{S}} d_i\mathbf{g}_i$ for the fixed θ_N, where $\mathbf{g}_i = \mathbf{g}(y_i, \mathbf{x}_i; \theta_N)$ is an $r \times 1$ vector, has the $r \times r$ design-based variance-covariance matrix

$$V_p\{\mathbf{G}_n(\theta_N)\} = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}(\pi_{ij} - \pi_i\pi_j)\,\frac{\mathbf{g}_i}{\pi_i}\,\frac{\mathbf{g}_j}{\pi_j}' \,,$$

where $\pi_{ii} = \pi_i$ as introduced in Sect. 4.1. An unbiased design-based variance estimator is given by

$$v_p\{\mathbf{G}_n(\theta_N)\} = \frac{1}{N^2}\sum_{i\in\mathbf{S}}\sum_{j\in\mathbf{S}}\frac{\pi_{ij} - \pi_i\pi_j}{\pi_{ij}}\,\frac{\mathbf{g}_i}{\pi_i}\,\frac{\mathbf{g}_j}{\pi_j}' \,.$$

For the estimator $\hat{\theta}$ defined through a differentiable estimating function, the design-based estimator for the variance-covariance matrix can be constructed from (7.9) through a plug-in method,

$$v_p(\hat{\theta}) = \{\mathbf{H}_n(\hat{\theta})\}^{-1} v_p\{\mathbf{G}_n(\hat{\theta})\}\{\mathbf{H}_n'(\hat{\theta})\}^{-1} \,,$$

where $\mathbf{H}_n(\hat{\theta})$ and $v_p\{\mathbf{G}_n(\hat{\theta})\}$ are obtained respectively from $\mathbf{H}_n(\theta_N)$ and $v_p\{\mathbf{G}_n(\theta_N)\}$, with the unknown θ_N replaced by $\hat{\theta}$. This is also called the "sandwich variance estimator". It should be noted that the factor $1/N$ in $\mathbf{H}_n(\hat{\theta})$ and the factor $1/N^2$ in $v_p\{\mathbf{G}_n(\hat{\theta})\}$ can be dropped in calculating $v_p(\hat{\theta})$.

The validity of the plug-in variance estimator $v_p(\hat{\theta})$ also requires the two quantities $\mathbf{H}_n(\theta)$ and $v_p\{\mathbf{G}_n(\theta)\}$ as functions of θ to be smooth at $\theta = \theta_N$. This will ensure that replacing θ_N by $\hat{\theta}$ does not change the values of the two functions asymptotically, i.e., $\mathbf{H}_n(\hat{\theta})/\mathbf{H}_n(\theta_N) = 1 + o_p(1)$ and $v_p\{\mathbf{G}_n(\hat{\theta})\}/v_p\{\mathbf{G}_n(\theta_N)\} = 1 + o_p(1)$. Binder (1983) has given sufficient conditions for the sandwich variance estimator to be valid.

Binder and Roberts (2009) have discussed scenarios where the design-based variance estimator $v_p(\hat{\theta})$ is also a valid variance estimator for the estimation of the model parameter θ. Under the joint ξp randomization framework,

$$V(\hat{\theta}) = E_\xi\{V_p(\hat{\theta})\} + V_\xi\{E_p(\hat{\theta})\} \,.$$

The term $E_\xi\{V_p(\hat{\theta})\}$ is typically of order $O(n^{-1})$ while the term $V_\xi\{E_p(\hat{\theta})\}$ is usually of order $O(N^{-1})$. When the sampling fraction n/N is small, the second term becomes negligible. The design-based variance estimator $v_p(\hat{\theta})$ is approximately unbiased for $V(\hat{\theta})$ since

$$E\{v_p(\hat{\theta})\} = E_\xi[E_p\{v_p(\hat{\theta})\}] \doteq E_\xi\{V_p(\hat{\theta})\} \doteq V(\hat{\theta}) \,.$$

Binder and Roberts (2009) have further discussed scenarios where the second term $V_\xi\{E_p(\hat{\theta})\}$ cannot be dropped and needs to be estimated as part of variance estimation for inferences on model parameters. See Sect. 14.3.3 for an example where the sampling fraction n/N is not small and the two variance components are of the same order.

7.2.3 Confidence Intervals and Regions

Confidence intervals or regions for the survey population parameter θ_N can be constructed in two different ways. The conventional approach is to use a Wald-type statistic based on the point estimator $\hat{\theta}$ and the variance estimator $v_p(\hat{\theta})$ and treat $\{v_p(\hat{\theta})\}^{-1/2}(\hat{\theta} - \theta_N)$ approximately as $N(\mathbf{0}, I_r)$, the standard multivariate normal distribution. If θ_N is one dimensional, the $(1 - \alpha)$-level confidence interval for θ_N based on the asymptotic normality of $\hat{\theta}$ is given by

$$\left(\hat{\theta} - Z_{\alpha/2}\big\{v_p(\hat{\theta})\big\}^{1/2},\ \ \hat{\theta} + Z_{\alpha/2}\big\{v_p(\hat{\theta})\big\}^{1/2}\right),$$

where $Z_{\alpha/2}$ is the upper $(\alpha/2)$-quantile of the standard normal distribution. When θ_N is an $r \times 1$ vector, the $(1-\alpha)$-level confidence region for θ_N may be constructed as

$$\left\{ \theta \;\middle|\; \big(\hat{\theta} - \theta\big)'\big\{v_p(\hat{\theta})\big\}^{-1}\big(\hat{\theta} - \theta\big) \le \chi_r^2(\alpha)\right\}, \tag{7.13}$$

where $\chi_r^2(\alpha)$ is the upper α-quantile of the χ^2 distribution with r degrees of freedom.

The confidence region (7.13) provides an α-level significance test for H_0: $\theta_N = \theta_0$ versus H_1: $\theta_N \neq \theta_0$ with a pre-specified θ_0. The χ^2 test statistic is computed as

$$\chi^2 = \big(\hat{\theta} - \theta_0\big)'\big\{v_p(\hat{\theta})\big\}^{-1}\big(\hat{\theta} - \theta_0\big).$$

We reject H_0 if $\chi^2 > \chi_r^2(\alpha)$. For a scalar θ_N, the test is equivalent to rejecting H_0 if $|z| > Z_{\alpha/2}$, where $z = (\hat{\theta} - \theta_0)/\{v_p(\hat{\theta})\}^{1/2}$.

An alternative approach is to use the asymptotic normality of $\mathbf{G}_n(\theta)$ at $\theta = \theta_N$, which has mean $\mathbf{G}_N(\theta_N) = \mathbf{0}$ and estimated variance-covariance matrix $v_p\{\mathbf{G}_n(\hat{\theta})\}$. We can consider a region defined as

$$\left\{ \theta \;\middle|\; \big\{\mathbf{G}_n'(\theta)\big\}\big[v_p\{\mathbf{G}_n(\hat{\theta})\}\big]^{-1}\big\{\mathbf{G}_n(\theta)\big\} \le \chi_r^2(\alpha)\right\}. \tag{7.14}$$

It should be noted that the population size N is not required for computing the region given by (7.14). For a vector-valued θ_N, the region defined by (7.14) might be difficult to use. However, it provides an α-level significance test for H_0: $\theta_N = \theta_0$ versus H_1: $\theta_N \neq \theta_0$ similar to the method from the region (7.13). The χ^2 test statistic is computed as

$$\chi^2 = \big\{\mathbf{G}_n'(\theta_0)\big\}\big[v_p\{\mathbf{G}_n(\hat{\theta})\}\big]^{-1}\big\{\mathbf{G}_n(\theta_0)\big\},$$

and we reject H_0 if $\chi^2 > \chi_r^2(\alpha)$.

When θ_N is one-dimensional and $G_n(\theta)$ is a monotone function of θ, the form given by (7.14) defines a $(1-\alpha)$-level confidence interval for θ_N. The lower and upper limits of the interval can be obtained as values of θ satisfying

$$G_n(\theta) = \pm\, Z_{\alpha/2}\big[v_p\{G_n(\hat{\theta})\}\big]^{1/2} ,$$

which amounts to treating $z = G_n(\theta)/[v_p\{G_n(\hat{\theta})\}]^{1/2}$ as asymptotically distributed as $N(0,1)$. Thompson (1997, Sect. 4.1.1) and Godambe and Thompson (1999) argued that it is more appealing to use $v_p\{G_n(\theta)\}$, which is the design-based variance estimator of $G_n(\theta)$ for the given θ. The pivotal quantity

$$z^* = G_n(\theta)/[v_p\{G_n(\theta)\}]^{1/2}$$

might be better approximated by $N(0,1)$ than the conventional z statistic involving $v_p\{G_n(\hat{\theta})\}$.

Suppose that the sample $\mathbf{S}$ of size n is selected by a single-stage PPS sampling design with negligible sampling fraction n/N. It is shown in Problem 7.1 that an approximate variance estimator for $G_n(\theta) = N^{-1}\sum_{i\in\mathbf{S}} d_i g_i(\theta)$ at $\theta = \theta_N$ is given by

$$v_p\{G_n(\theta)\} = \frac{1}{N^2}\sum_{i\in\mathbf{S}} d_i^2\{g_i(\theta)\}^2 .$$

The lower and upper limits of the confidence interval may be obtained by solving

$$\frac{\{G_n(\theta)\}^2}{v_p\{G_n(\theta)\}} = \frac{\{\sum_{i\in\mathbf{S}} d_i g_i(\theta)\}^2}{\sum_{i\in\mathbf{S}} d_i^2\{g_i(\theta)\}^2} = \chi_1^2(\alpha) . \tag{7.15}$$

The method, however, can be difficult to use in practice since the above equation may have more than two solutions and the solutions may not have closed form expressions. See Problems 7.2 and 7.3 for applications of the method.

An important problem in statistical inference is estimation and testing in the presence of nuisance parameters. Let $\theta_N = (\theta_1', \theta_2')'$, where θ_1 is $k_1 \times 1$ and θ_2 is $k_2 \times 1$, and θ_N is $k \times 1$ with $k = k_1 + k_2$. The parameter of interest is θ_1 and the other parameter θ_2 is treated as nuisance. Godambe (1991b) and Binder and Patak (1994) discussed the special case where θ_1 is a scalar. Section 8.6 presents general results on inferences with nuisance parameters.

7.3 Regression Analysis with Survey Data

Regression analysis is a commonly used statistical tool for exploring relationships among variables and for establishing associations between responses and covariates.

It is well known that association does not imply causation, but regression analysis is also one of the major techniques required for causal inferences.

Let y be the response variable and $\mathbf{x}$ be a vector of covariates. The primary objective of regression analysis is to identify the form of the conditional mean of the response given the covariates, i.e., $E_\xi(y \mid \mathbf{x}) = \mu(\mathbf{x})$, where ξ indicates the underlying superpopulation model for $(y, \mathbf{x})$. For scenarios where there are many covariates observed along with the response variable, another major objective of regression analysis is to identify the set of covariates with or without interactions that are "significant" to the response. Although parametric methods such as maximum likelihood and nonparametric methods such as local polynomial smoothing (Fan and Gijbels 1996) have been used for regression analysis, semiparametric methods are more widely used in practice, especially for complex survey data. Semiparametric models are specified through the first and the second moments: a known functional form for the conditional mean $\mu(\mathbf{x}) = \mu(\mathbf{x}, \theta)$ with a set of unknown parameters θ plus a known form for the variance structure. The approach fits naturally to the general theory of estimating equations.

There has been research on the use or misuse of survey weights for regression analysis and statistical modelling with complex survey data. See, for instance, the paper by Pfeffermann (1993) and the discussion paper by Gelman (2007). This section presents design-based regression analysis for complex surveys, with the major focus on outlining the details for point and variance estimation for the regression coefficients and the related computational procedures. Design-based regression modelling, variable selection and general hypothesis testing are further discussed in Chap. 8. The survey dataset is represented by $\{(y_i, \mathbf{x}_i), i \in \mathbf{S}\}$ and the survey design is characterized by the inclusion probabilities π_i and π_{ij}.

7.3.1 Linear Regression Analysis

Suppose that the finite population follows a linear regression superpopulation model (ξ),

$$y_i = \mathbf{z}_i'\beta + \varepsilon_i\,, \quad i = 1, 2, \ldots, N\,, \tag{7.16}$$

where $\beta = (\beta_0, \beta_1, \cdots, \beta_k)'$ are the regression coefficients, $\mathbf{z}_i = (1, x_{i1}, x_{i2}, \cdots, x_{ik})'$ with the first component as 1 for the intercept, and the ε_i's are independent with $E_\xi(\varepsilon_i) = 0$ and $V_\xi(\varepsilon_i) = v_i\sigma^2$. This is the same model as (5.12) introduced in Sect. 5.4 where the focus is on model-assisted estimation for the population mean μ_y. In this section the main goal is on design-based inference for the regression coefficients β. We consider a homogeneous variance structure where $v_i = 1$ and $V_\xi(\varepsilon_i) = \sigma^2$. Regression analysis under heteroscedasticity is left as an exercise (Problem 7.4).

The conventional least square estimator (LSE) β_N of the regression coefficients β using the census data $\{(y_i, \mathbf{z}_i), i = 1, \cdots, N\}$ is obtained by minimizing $Q(\beta) =$

$\sum_{i=1}^{N}(y_i - \mathbf{z}_i'\beta)^2$ with respect to β. It amounts to solving $\partial Q(\beta)/\partial\beta = \mathbf{0}$, or equivalently, finding the solution to the census estimating equations

$$\sum_{i=1}^{N}\mathbf{g}(y_i, \mathbf{z}_i, \beta) = \sum_{i=1}^{N}\mathbf{z}_i(y_i - \mathbf{z}_i'\beta) = \mathbf{0}\,, \tag{7.17}$$

where $\mathbf{g}(y_i, \mathbf{z}_i, \beta) = \mathbf{z}_i(y_i - \mathbf{z}_i'\beta)$. The solution to (7.17) is the so-called *population regression coefficients* and has a closed form expression given by

$$\beta_N = \Bigl(\sum_{i=1}^{N}\mathbf{z}_i\mathbf{z}_i'\Bigr)^{-1}\sum_{i=1}^{N}\mathbf{z}_i y_i\,.$$

The quantity defined by β_N is a finite population parameter and can be of interest by itself without referring to the superpopulation model (7.16).

The estimator $\hat{\beta}$ of β_N or β is obtained by solving the survey weighted estimating equations

$$\sum_{i\in\mathbf{S}}d_i\mathbf{g}(y_i, \mathbf{z}_i, \beta) = \sum_{i\in\mathbf{S}}d_i\mathbf{z}_i(y_i - \mathbf{z}_i'\beta) = \mathbf{0}\,,$$

where $d_i = 1/\pi_i$ are the survey design weights. The estimator has a closed form expression given by

$$\hat{\beta} = \Bigl(\sum_{i\in\mathbf{S}}d_i\mathbf{z}_i\mathbf{z}_i'\Bigr)^{-1}\sum_{i\in\mathbf{S}}d_i\mathbf{z}_i y_i\,. \tag{7.18}$$

Design-based regression analysis focuses on the population regression coefficients β_N, which characterizes the relationship between the response variable y and the covariates $(x_1, \cdots, x_k)$ for the given finite population. Design-based properties of the estimator $\hat{\beta}$ given by (7.18) differ from those of the conventional model-based LSE in both the bias and the variance. First, the estimator $\hat{\beta}$ is not exactly design-unbiased for β_N, since it is a nonlinear function of two Horvitz-Thompson estimators $\sum_{i\in\mathbf{S}}d_i\mathbf{z}_i\mathbf{z}_i'$ and $\sum_{i\in\mathbf{S}}d_i\mathbf{z}_i y_i$. However, the relation $\hat{\beta} - \beta_N = O_p(n^{-1/2})$ still holds under certain regularity conditions as demonstrated by the general results presented in Part (a) of Theorem 7.1. Second, the design-based asymptotic variance-covariance matrix of $\hat{\beta}$ is different from the model-based result. Following Part (b) of Theorem 7.1, we have

$$AV_p(\hat{\beta}) = \Bigl(\sum_{i=1}^{N}\mathbf{z}_i\mathbf{z}_i'\Bigr)^{-1}V_p\Bigl(\sum_{i\in\mathbf{S}}d_i\mathbf{g}_i(\beta_N)\Bigr)\Bigl(\sum_{i=1}^{N}\mathbf{z}_i\mathbf{z}_i'\Bigr)^{-1}\,,$$

where $AV_p(\cdot)$ denotes the asymptotic design-based variance and $\mathbf{g}_i(\beta_N) = \mathbf{z}_i(y_i - \mathbf{z}_i'\beta_N)$. The design-based asymptotic variance $AV_p(\hat{\beta})$ does not reduce to the familiar expression $\sigma^2(\sum_{i=1}^N \mathbf{z}_i\mathbf{z}_i')^{-1}$ as seen in conventional regression analysis. The estimated design-based variance of $\hat{\beta}$ is given by

$$v_p(\hat{\beta}) = \Big(\sum_{i\in\mathbf{S}} d_i\mathbf{z}_i\mathbf{z}_i'\Big)^{-1} v_p\Big(\sum_{i\in\mathbf{S}} d_i\mathbf{g}_i(\beta_N)\Big)\Big(\sum_{i\in\mathbf{S}} d_i\mathbf{z}_i\mathbf{z}_i'\Big)^{-1},$$

where $v_p\left(\sum_{i\in\mathbf{S}} d_i\mathbf{g}_i(\beta_N)\right)$ is the design-based variance estimator for the "Horvitz-Thompson estimator" $\sum_{i\in\mathbf{S}} d_i\mathbf{g}_i(\beta_N)$, with β_N replaced by $\hat{\beta}$ for computing the final variance estimate.

Let $\mathbf{Z}_n = (\mathbf{z}_1', \cdots, \mathbf{z}_n')'$ be the $n \times (k+1)$ matrix for the observed values of the covariates. Let $\mathbf{W}_n = \text{diag}(d_1, \cdots, d_n)$ be the diagonal matrix for the survey design weights. Let $\mathbf{Y}_n = (y_1, \cdots, y_n)'$. The survey weighted estimator $\hat{\beta}$ given by (7.18) can be written as

$$\hat{\beta} = \Big(\mathbf{Z}_n'\mathbf{W}_n\mathbf{Z}_n\Big)^{-1}\mathbf{Z}_n'\mathbf{W}_n\mathbf{Y}_n .$$

The form of $\hat{\beta}$ is identical to the weighted least square estimator (WLSE) for the regression coefficients. If one fits a regression model using standard (non-survey based) statistical software with the specification of the weight matrix $\mathbf{W}_n$, one would obtain the correct point estimate for β_N. However, the reported standard errors and test results on β_N would be incorrect since the model-based variance estimate is incorrect under the design-based framework.

Appendix A.4.1 provides an R function for survey weighted linear regression analysis for single-stage unequal probability sampling designs. The design-based variance estimator uses the result from Part (c) of Problem 7.1.

7.3.2 Logistic Regression Analysis

When the response variable is binary, i.e., $y = 1$ or 0, linear regression models can no longer be used. The relationship between the y and a set of covariates $\mathbf{x}$ is modelled through $E_\xi(y \mid \mathbf{x}) = P(y = 1 \mid \mathbf{x})$, where ξ indicates the model. We spell out the details for design-based inferences under the logistic regression model. Design-based inferences under the probit model or the complementary log-log model can also be developed using results from survey weighted estimating equation methods.

Let $\{(y_i, \mathbf{x}_i),\ i = 1, \cdots, N\}$ be the census data for the finite population, where $y_i = 1$ or 0 and $\mathbf{x}_i = (x_{i1}, \cdots, x_{ik})'$. Let $p_i = E_\xi(y_i \mid \mathbf{x}_i) = P(y_i = 1 \mid \mathbf{x}_i)$. The full log-likelihood function based on the census data is given by

$$\ell(\beta) = \log\left\{\prod_{i=1}^{N} p_i^{y_i}(1-p_i)^{1-y_i}\right\} = \sum_{i=1}^{N}\Big\{y_i \log p_i + (1-y_i)\log(1-p_i)\Big\},$$

where β is the vector of model parameters to be specified below. Under the logistic regression model (ξ),

$$\log\left(\frac{p_i}{1-p_i}\right) = \mathbf{z}_i'\beta\,,$$

where $\mathbf{z}_i = (1, x_{i1}, \cdots, x_{ik})'$ are the covariates (with the first component as 1) and $\beta = (\beta_0, \beta_1, \cdots, \beta_k)'$ are the regression coefficients (with an intercept). The census estimating equations can be obtained as the score equations $\partial\ell(\beta)/\partial\beta = \mathbf{0}$ and are given by

$$\sum_{i=1}^{N}\mathbf{g}(y_i, \mathbf{z}_i, \beta) = \sum_{i=1}^{N}\mathbf{z}_i(y_i - p_i) = \mathbf{0}\,, \qquad (7.19)$$

where $\mathbf{g}(y_i, \mathbf{z}_i, \beta) = \mathbf{z}_i(y_i - p_i)$ and $p_i = \exp(\mathbf{z}_i'\beta)/\{1 + \exp(\mathbf{z}_i'\beta)\}$. The solution to (7.19) defines the population regression coefficients β_N for the logistic regression model. Unfortunately, there is no closed form expression for β_N.

The design-based estimator $\hat{\beta}$ of the population regression coefficients β_N is the solution to

$$\mathbf{G}_n(\beta) = \sum_{i\in\mathbf{S}} d_i\mathbf{g}(y_i, \mathbf{z}_i, \beta) = \sum_{i\in\mathbf{S}} d_i\mathbf{z}_i(y_i - p_i) = \mathbf{0}\,. \qquad (7.20)$$

Note that the redundant factor N^{-1} used in (7.8) is dropped here. Once again, there is no closed form expression for $\hat{\beta}$. Let $\mathbf{p}_n = (p_1, \cdots, p_n)'$. The equations (7.20) can be written as

$$\mathbf{G}_n(\beta) = \mathbf{Z}_n'\mathbf{W}_n(\mathbf{Y}_n - \mathbf{p}_n) = \mathbf{0}\,,$$

where $\mathbf{Z}_n$, $\mathbf{W}_n$ and $\mathbf{Y}_n$ are defined in Sect. 7.3.1. It can be shown that

$$\mathbf{H}_n(\beta) = \frac{\partial\mathbf{G}_n(\beta)}{\partial\beta} = -\mathbf{Z}_n'\mathbf{M}_n\mathbf{Z}_n\,,$$

where $\mathbf{M}_n = \text{diag}(d_1p_1(1-p_1), \cdots, d_np_n(1-p_n))$ is a diagonal matrix. The estimator $\hat{\beta}$ can be computed using the iterative procedure (7.11) and the specific forms of $\mathbf{G}_n(\beta)$ and $\mathbf{H}_n(\beta)$ derived for the logistic regression model.

Appendix A.4.2 provides an R function for survey weighted logistic regression analysis for single-stage unequal probability sampling designs. Once again, the design-based variance estimator uses the result from Part (c) of Problem 7.1.

7.4 Longitudinal Surveys and Generalized Estimating Equations

There are two major types of survey designs, namely, cross-sectional surveys and longitudinal surveys. Cross-sectional surveys are a "one-time" study where interest lies in the characteristics of the population at a particular time point. All discussions prior to this section as well as all remaining chapters in Part II of the book deal with cross-sectional surveys. Longitudinal surveys, also called *panel studies*, measure the variables of interest on units in a randomly selected sample at several time points. Each time point is referred to as a *cycle* or *wave*. The same set of units, i.e., the *panel*, remains in the sample during a reference time period. Longitudinal surveys allow measurements of the responses over time and measurements of time-varying explanatory variables for the same unit and therefore are equipped for the exploration of population changes over time at the individual level.

One of the major challenges in analyzing longitudinal data is the necessity of dealing with correlated multivariate responses measured over time on the same unit. The most popular method for analyzing non-survey-based longitudinal data is the generalized estimating equations (GEE) approach (Liang and Zeger 1986). In this section we describe the so-called Pseudo-GEE method for analyzing longitudinal survey data (Rao 1998; Carrillo et al. 2010). Another major challenge for longitudinal surveys is the missing data problem, which is not discussed further in the book. Some relevant references on handling missing data in longitudinal studies can be found in Carrillo et al. (2011).

Let $(y_{ij}; x_{ij1}, \cdots, x_{ijk})$ be the values of the response variable y and the vector of k covariates $(x_1, \cdots, x_k)$ for unit i at the time of the jth cycle of the survey, $j = 1, \cdots, t_i$. The number of cycles t_i can be different for different units but in most studies $t_i = t$ is common for all units. This is typically the case for large scale surveys and is considered in this section for notational simplicity. Let $\mathbf{x}_{ij} = (1, x_{ij1}, \cdots, x_{ijk})'$ and $\mathbf{x}_i = (\mathbf{x}_{i1}, \cdots, \mathbf{x}_{it})'$. Note that $\mathbf{x}_{ij}$ is a $(k+1) \times 1$ vector and $\mathbf{x}_i$ is a $t \times (k+1)$ matrix representing all the values of the covariates for unit i over the t cycles.

We assume that the superpopulation model ξ can be characterized by the following three components:

1. The conditional mean response $\mu_{ij} = E_\xi(y_{ij} \mid \mathbf{x}_{ij})$ is related to the linear predictor $\eta_{ij} = \mathbf{x}_{ij}'\beta$ through a monotone link function $h(\cdot)$: $\mu_{ij} = h^{-1}(\eta_{ij}) = h^{-1}(\mathbf{x}_{ij}'\beta)$, where $\beta = (\beta_0, \beta_1, \cdots, \beta_k)'$ are the model parameters and $h^{-1}(\cdot)$ denotes the inverse function of $h(\cdot)$.
2. The conditional variance of y_{ij} given $\mathbf{x}_{ij}$ is specified by $V_\xi(y_{ij} \mid \mathbf{x}_{ij}) = \phi v(\mu_{ij})$, where $v(\cdot)$ is the variance function with a known form and $\phi > 0$ is called a dispersion parameter.
3. The conditional variance-covariance matrix of $\mathbf{y}_i = (y_{i1}, \cdots, y_{it})'$ is given by $V_\xi(\mathbf{y}_i \mid \mathbf{x}_i) = \mathbf{A}_i^{1/2}\mathbf{R}_i\,\mathbf{A}_i^{1/2}$, where $\mathbf{A}_i = \text{diag}\{\phi v(\mu_{i1}), \cdots, \phi v(\mu_{it})\}$ and $\mathbf{R}_i$ is

the correlation matrix for the t measurements from unit i with either a specified structure involving additional parameters or an unspecified structure.

We also assume that the finite population is a random sample from the superpopulation, which implies that

4. The response vectors $\mathbf{y}_i$ and $\mathbf{y}_l$ given $\mathbf{x}_i$ and $\mathbf{x}_l$ are independent for two different units $i \neq l$.

Among the four model components described above, items 1, 2, and 4 are similar to those for the generalized linear models (GLM). However, there are two important and unique features in the model specifications for longitudinal data which are not part of GLM. Firstly, it is possible to include time-dependent covariates in ξ to explore changes over time. Such variables can be as simple as age or variables by specific design features of the study. This allows the examination of the effectiveness of, say, population interventions before and after certain time points while controlling other factors in the study. An example of such designs is described in Chap. 12 with reference to the International Tobacco Control (ITC) Policy Evaluation Project. Secondly, "it is the (third) component, the incorporation of the within-subject association among the repeated responses from the same individual, that represents the main extension of GLM to longitudinal data" (Fitzmaurice et al. 2004).

Following the GEE approach as described in Liang and Zeger (1986), the finite population parameters β_N are defined as the solution to the census generalized estimating equations

$$\mathbf{G}_N(\beta) = \sum_{i=1}^{N} \mathbf{g}(\mathbf{y}_i, \mathbf{x}_i, \beta) = \sum_{i=1}^{N} \left(\frac{\partial \mu_i}{\partial \beta}\right)' \mathbf{V}_i^{-1}(\mathbf{y}_i - \mu_i) = \mathbf{0}\,, \tag{7.21}$$

where $\mu_i = (\mu_{i1}, \cdots, \mu_{it})' = E_\xi(\mathbf{y}_i \mid \mathbf{x}_i)$ is the $t \times 1$ conditional mean responses and $\mathbf{V}_i = \mathbf{A}_i^{1/2}\mathbf{R}_i\,\mathbf{A}_i^{1/2}$ is the $t \times t$ working variance-covariance structure with the chosen correlation matrix $\mathbf{R}_i$. Note that μ_i and β are linked through $\mu_{ij} = h^{-1}(\mathbf{x}_{ij}'\beta)$ and $\mathbf{g}(\mathbf{y}_i, \mathbf{x}_i, \beta) = (\partial\mu_i/\partial\beta)'\mathbf{V}_i^{-1}(\mathbf{y}_i - \mu_i)$. The dimensions of $\partial\mu_i/\partial\beta$ is $t \times (k+1)$ and $\mathbf{g}(\mathbf{y}_i, \mathbf{x}_i, \beta)$ is $(k+1) \times 1$.

The design-based estimator $\hat{\beta}$ of the finite population parameters β_N is the solution to the survey weighted generalized estimating equations (Rao 1998; Carrillo et al. 2010)

$$\mathbf{G}_n(\beta) = \sum_{i \in \mathbf{S}} d_i \mathbf{g}(\mathbf{y}_i, \mathbf{x}_i, \beta) = \sum_{i \in \mathbf{S}} d_i \left(\frac{\partial \mu_i}{\partial \beta}\right)' \mathbf{V}_i^{-1}(\mathbf{y}_i - \mu_i) = \mathbf{0}\,, \tag{7.22}$$

where d_i is the design weight for selecting unit i into the longitudinal sample $\mathbf{S}$. The estimator $\hat{\beta}$ is termed as the Pseudo-GEE estimator for the model parameters β and is consistent under suitable regularity conditions and the joint randomization of the GEE model ξ and the survey design (Carrillo et al. 2010).

The Newton-Raphson procedure (7.11) can be used to find the Pseudo-GEE estimator $\hat{\beta}$. Unfortunately, the matrix $\mathbf{H}_n(\beta) = \{\partial \mathbf{G}_n(\beta)/\partial \beta\}$ used in (7.11) involves $\partial \mathbf{g}(\mathbf{y}_i, \mathbf{x}_i, \beta)/\partial \beta$, which turns out to be extremely complicated. Fortunately, it can be shown (Problem 7.6) that $\{\partial \mathbf{G}_n(\beta)/\partial \beta\}/\{\mathbf{H}_n^*(\beta)\} = 1 + o_p(1)$ when β is the true value of the model ξ, where

$$\mathbf{H}_n^*(\beta) = -\sum_{i \in \mathbf{S}} d_i \left(\frac{\partial \mu_i}{\partial \beta}\right)' \mathbf{V}_i^{-1} \left(\frac{\partial \mu_i}{\partial \beta}\right) . \tag{7.23}$$

The procedure (7.11) can be modified to solve (7.22) with $\mathbf{H}_n(\beta)$ replaced by $\mathbf{H}_n^*(\beta)$.

The dispersion parameter ϕ is not needed for finding the point estimator $\hat{\beta}$ since it appears as a multiplying constant in $\mathbf{V}_i$ and can be dropped from the equation system (7.22). However, the dispersion parameter ϕ is required for variance estimation. The parameter ϕ can be estimated by

$$\hat{\phi} = \frac{\sum_{i \in \mathbf{S}} d_i \sum_{j=1}^{t} r_{ij}^2}{\sum_{i \in \mathbf{S}} d_i t - (k+1)} ,$$

where $r_{ij} = (y_{ij} - \hat{\mu}_{ij})/\{v(\hat{\mu}_{ij})\}^{1/2}$ is the standardized residual and $\hat{\mu}_{ij} = h^{-1}(\mathbf{x}_{ij}'\hat{\beta})$ is the fitted value for y_{ij}.

Specification of the correlation matrix $\mathbf{R}_i$ requires certain prior knowledge about the measurements. For instance, it might be reasonable to assume that two measurements y_{ij} and y_{il} from the same unit i are independent given the covariates if $|j - l| > 1$, i.e., if y_{ij} and y_{il} are not measured over two adjacent time points. A practically important scenario is that $\mathbf{R}_i$ has an unspecified structure but remains the same for all units, i.e., $\mathbf{R}_i = \mathbf{R} = (\rho_{jl})$. The unknown correlation ρ_{jl} can then be estimated by

$$\hat{\rho}_{jl} = \frac{\sum_{i \in \mathbf{S}} d_i r_{ij} r_{il}/\hat{\phi}}{\sum_{i \in \mathbf{S}} d_i - (k+1)} ,$$

where r_{ij} is the standardized residual defined earlier. It should be noted that the estimation of the dispersion parameter ϕ requires the chosen correlation matrix $\mathbf{R}$ but the specification and estimation of $\mathbf{R}$ also requires the estimated value of ϕ. One strategy to overcome this difficulty is to first use an independent variance-covariance structure with $\mathbf{R}$ being the identity matrix to obtain the fitted values $\hat{\mu}_{ij}$ and to compute an initial estimate for ϕ. The correlation matrix can then be estimated using the initially estimated ϕ. The process can be iterated back and forth a couple times if needed.

The design-based variance for the Pseudo-GEE estimator $\hat{\beta}$ can be estimated using the sandwich-type variance estimator and is given by

$$v_p(\hat{\beta}) = \left\{\mathbf{H}_n^*(\hat{\beta})\right\}^{-1} v_p\left\{\mathbf{G}_n(\hat{\beta})\right\}\left\{\mathbf{H}_n^*(\hat{\beta})\right\}^{-1} ,$$

where $\mathbf{H}_n^*(\beta)$ is defined in (7.23) and $v_p\{\mathbf{G}_n(\hat{\beta})\}$ is the estimated design-based variance of $\mathbf{G}_n(\beta)$ given in (7.22), all evaluated at $\hat{\beta}$.

We conclude this section by reiterating that the most important feature of longitudinal surveys is that the initial set of sampled units are followed and measured on subsequent cycles (or waves). Changes of the population at individual level usually cannot be detected through different cross-sectional surveys at different time points. For instance, two independent surveys at two different time points may report the same estimate of smoking prevalence in a population. It would be difficult to ascertain whether all smokers and non-smokers possess the same status at the two time points or whether some individuals in both groups have changes in smoking status. For the latter case it would be even more difficult to determine whether such changes are due to, say, a population intervention or some other uncontrolled factors. Longitudinal surveys and the pseudo-GEE approach to inference provide a powerful tool for scientific investigations on population changes over time at individual level.

7.5 Problems

7.1 (Variance Estimation Under PPS Sampling with Replacement) Let z_i, $i = 1, \ldots, N$ be a size variable such that $z_i > 0$ and $\sum_{i=1}^N z_i = 1$. The parameter θ_N is a $k \times 1$ vector and is defined as the solution to the equations $\mathbf{K}_N(\theta) = \sum_{i=1}^N \mathbf{g}_i(\theta) = \mathbf{0}$. A sample of n units is selected by PPS sampling with replacement, with probabilities z_i for each selection. Let $\mathbf{K}_n(\theta) = n^{-1}\sum_{i=1}^n z_i^{-1}\mathbf{g}_i(\theta)$, which is the Hansen-Hurwitz estimator of $\mathbf{K}_N(\theta) = \sum_{i=1}^N \mathbf{g}_i(\theta)$. It follows that $E_p\{\mathbf{K}_n(\theta)\} = \mathbf{K}_N(\theta)$ for any given θ.

(a) Argue that an unbiased estimator for the variance-covariance matrix of the Hansen-Hurwitz estimator is given by

$$v_p\{\mathbf{K}_n(\theta)\} = \frac{1}{n}\frac{1}{n-1}\Big[\sum_{i=1}^n \{z_i^{-1}\mathbf{g}_i(\theta)\}\{z_i^{-1}\mathbf{g}_i(\theta)\}' - n\{\mathbf{K}_n(\theta)\}\{\mathbf{K}_n(\theta)\}'\Big].$$

(b) Argue that, under suitable regularity conditions, a design-consistent estimator of the variance-covariance matrix of $\mathbf{K}_n(\theta)$ at $\theta = \theta_N$ is given by

$$v_p\{\mathbf{K}_n(\theta_N)\} = \frac{1}{n^2}\sum_{i=1}^n \{z_i^{-1}\mathbf{g}_i(\theta_N)\}\{z_i^{-1}\mathbf{g}_i(\theta_N)\}'.$$

(c) For single-stage PPS sampling with negligible sampling fraction n/N and inclusion probabilities $\pi_i = nz_i$, the estimator of part (b) can be used as an approximation for variance estimation. Show that the variance-covariance estimator for $\mathbf{G}_n(\theta_N) = N^{-1}\sum_{i\in\mathbf{S}} d_i\mathbf{g}_i(\theta_N)$ under this setting reduces to

$$v_p\{\mathbf{G}_n(\theta_N)\} = \frac{1}{N^2}\sum_{i\in\mathbf{S}} d_i^2\{\mathbf{g}_i(\theta_N)\}\{\mathbf{g}_i(\theta_N)\}'\,.$$

7.2 (Alternative Confidence Interval for $\theta_N = \mu_y$ Under SRSWOR) The estimating function for defining $\theta_N = \mu_y$ is $g_i(\theta) = y_i - \theta$. Under simple random sampling without replacement with small sampling fraction n/N, the lower and upper limits of the $(1-\alpha)$-level confidence interval specified by (7.15) are obtained by solving

$$\frac{\{\sum_{i\in\mathbf{S}}(y_i-\theta)\}^2}{\sum_{i\in\mathbf{S}}(y_i-\theta)^2} = \chi_1^2(\alpha)\,.$$

(a) Derive the closed form expressions for the lower and upper limits for the interval.

(b) Run a simulation study to compare the performance of the interval with the conventional interval. Consider scenarios for $n/N \le 1\%$ with $n = 10$, 100 and 1000.

7.3 (Alternative Confidence Interval for $\theta_N = P$ Under SRSWOR) Suppose that $y_i = 1$ or 0 and $\theta_N = P = M/N$ is the population proportion of units with $y_i = 1$, which is the solution to $\sum_{i=1}^N(y_i-\theta) = 0$. Under simple random sampling without replacement, the sample-based estimator $\hat{\theta}$ is the solution to $\sum_{i\in\mathbf{S}}(y_i-\theta) = 0$. We have $V_p\{\sum_{i\in\mathbf{S}}(y_i-\theta_N)\} = V_p\{n(\bar{y}-\theta_N)\} = n(1-n/N)(1+1/(N-1))\theta_N(1-\theta_N)$, which can be approximated by $n\theta_N(1-\theta_N)$ for large N and small n/N. The lower and upper limits of the confidence interval for $\theta_N = P$ can be obtained as values of θ satisfying

$$\frac{(\bar{y}-\theta)^2}{n^{-1}\theta(1-\theta)} = \chi_1^2(\alpha)\,.$$

The resulting confidence interval is also referred to as the Wilson Interval (Wilson 1927).

(a) Derive closed form expressions for the lower and upper limits for the interval.

(b) Run a simulation study to compare the performance of the interval with the conventional interval. Consider scenarios for $n/N \le 1\%$ with $n = 10$, 100 and 1000.

7.4 (Linear Regression Models with Heteroscedasticity) Consider the linear regression model (7.16) with heterogeneous variance structure $V_\xi(\varepsilon_i) = v_i\sigma^2$, where $v_i = v(\mathbf{x}_i)$ is a known positive function of $\mathbf{x}_i$. The weighted least square estimator (WLSE) β_N of the regression coefficients β using the census data $\{(y_i, \mathbf{z}_i), i = 1, \cdots, N\}$ is obtained by minimizing $Q(\beta) = \sum_{i=1}^N(y_i - \mathbf{z}_i'\beta)^2/v_i$ with respect to β.

(a) Find the explicit expression for β_N.
(b) Find the explicit expression for the design-based estimator $\hat{\beta}$.
(c) Give an expression for the design-based variance estimator $v_p(\hat{\beta})$.

7.5 (The Complementary Log-Log Regression Model for Binary Responses) Let $\{(y_i, \mathbf{x}_i),\ i = 1, \cdots, N\}$ be the census data for the finite population, where $y_i = 1$ or 0 and $\mathbf{x}_i = (x_{i1}, \cdots, x_{ik})'$. Let $p_i = E_\xi(y_i \mid \mathbf{x}_i) = P(y_i = 1 \mid \mathbf{x}_i)$. The complementary log-log regression model specifies that $\log(-\log(p_i)) = \mathbf{z}_i'\beta$, where $\mathbf{z}_i = (1, x_{i1}, \cdots, x_{ik})'$ are the covariates and $\beta = (\beta_0, \beta_1, \cdots, \beta_k)'$ are the regression coefficients.

(a) Develop an iterative procedure for finding the design-based estimator $\hat{\beta}$.
(b) Write an R function for computing $\hat{\beta}$ and the estimated variance $v_p(\hat{\beta})$ for single-stage unequal probability sampling designs using the result from Part (c) of Problem 7.1.

7.6 (The Pseudo-GEE Estimator with Longitudinal Surveys) Let $\mathbf{G}_n(\beta)$ be defined by (7.22) for the GEE model and a longitudinal survey. Argue that $\{\partial \mathbf{G}_n(\beta)/\partial\beta\}/\{\mathbf{H}_n^*(\beta)\} = 1 + o_p(1)$ when β is the true value of the GEE model ξ, where

$$\mathbf{H}_n^*(\beta) = -\sum_{i \in \mathbf{S}} d_i \left(\frac{\partial \mu_i}{\partial \beta}\right)' \mathbf{V}_i^{-1} \left(\frac{\partial \mu_i}{\partial \beta}\right).$$

Chapter 8
Empirical Likelihood Methods

Likelihood-based approaches are one of the major pillars for statistical inferences. Parametric likelihood theory presents many attractive tools such as maximum likelihood estimators and generalized likelihood ratio tests. Unfortunately, those widely applicable techniques do not lend themselves naturally to design-based inferences in survey sampling. Under the design-based framework, the finite population is viewed as fixed, and randomization is induced by the probability sampling design, $\mathscr{P}$. The probability measure $\mathscr{P}$ is defined over subsets of the finite population units and is not associated with any particular survey variables and their distributions.

Godambe (1966) defined the design-based likelihood function for a particular variable y as follows. Let $\tilde{\mathbf{y}} = (\tilde{y}_1, \tilde{y}_2, \cdots, \tilde{y}_{\mathrm{N}})'$ be the population vector of parameters, where the tilde indicates that each y_i is treated as an unknown parameter. The likelihood function of $\tilde{\mathbf{y}}$ is the probability of observing the sample data $\{y_i, i \in \mathbf{S}\}$ for the given $\tilde{\mathbf{y}}$, which is given by

$$L(\tilde{\mathbf{y}}) = P(y_i, i \in \mathbf{S} \mid \tilde{\mathbf{y}}) = \begin{cases} \mathscr{P}(\mathbf{S}) & \text{if } \; y_i = \tilde{y}_i \; \text{ for } \; i \in \mathbf{S}\,, \\ 0 & \text{otherwise}\,, \end{cases}$$

where $\mathscr{P}(\mathbf{S})$ denotes the probability of selecting the sample $\mathbf{S}$ under the design. Although the likelihood function $L(\tilde{\mathbf{y}})$ is well defined, it is uninformative in the sense that all possible non-observed values $y_i, i \notin \mathbf{S}$ lead to the same "flat" likelihood function. This difficulty arises because of the distinct labels i associated with the units in the sample data that make the sample unique.

Hartley and Rao (1968) proposed a *scale-load* approach under the assumption that the variable y is measured on a scale with a finite set of unknown scale points $y_l^*, l = 1, \ldots, L$. If the sample of size n is taken by SRSWOR with n_l units in the sample having the value y_l^*, the observed scale-loads $(n_1, \ldots, n_{\mathrm{L}})$ follow the multi-hypergeometric distribution. If the sampling fraction n/N is small, the likelihood function may be approximated by the multinomial distribution. Hartley and Rao (1969) generalized the scale-load approach to PPS sampling with replacement. The

C. Wu, M. E. Thompson, *Sampling Theory and Practice*, ICSA Book Series in Statistics, https://doi.org/10.1007/978-3-030-44246-0_8

basic concept behind the scale-load approach is similar to the empirical likelihood method discussed in this chapter but the papers by Hartley and Rao (1968, 1969) focused primarily on point estimation. Hartley and Rao (1968) also considered a constrained maximization problem with the known population mean of the auxiliary variable used in a calibration equation and showed that the resulting maximum scale-load estimator is asymptotically equivalent to the regression estimator. This result was later "re-discovered" by Chen and Qin (1993) using the empirical likelihood formulation of Owen (1988).

This chapter presents empirical likelihood methods for complex surveys. We first focus on point estimation and confidence intervals for the single descriptive finite population parameter $\theta = \mu_y$ using the pseudo empirical likelihood method. We then develop general inferential procedures for parameters defined through estimating equations using either the pseudo empirical likelihood approach or the sample empirical likelihood approach. When the population size N is unknown and the parameter of interest is the population total T_y, a generalized pseudo empirical likelihood method can be used.

8.1 Pseudo Empirical Likelihood and Sample Empirical Likelihood

The empirical likelihood method was first proposed by Owen (1988) for independent samples and has been developed into a general inference tool for many areas of statistics. Suppose that $\{y_1, \ldots, y_n\}$ is an independent and identically distributed sample from Y with an unknown distribution function $F(y)$ and mean $\mu = E(Y)$. Owen's empirical likelihood formulation consists of three basic components. The first is the log empirical likelihood (EL) function,

$$\ell(\mathbf{p}) = \log\Big(\prod_{i=1}^{n} p_i\Big) = \sum_{i=1}^{n} \log(p_i)\,, \tag{8.1}$$

where $\mathbf{p} = (p_1, \ldots, p_n)$ is a discrete probability measure over the sample $\{y_1, \ldots, y_n\}$. The second is the normalization constraint

$$\sum_{i=1}^{n} p_i = 1 \tag{8.2}$$

with $p_i > 0$ for all i. The third component is the constraint induced by the parameter of interest, $\mu = E(Y)$, and is given by

$$\sum_{i=1}^{n} p_i y_i = \mu\,. \tag{8.3}$$

The parameter constraint (8.3) is a sample version of the moment condition $E(Y) = \mu$. The EL function $\ell(\mathbf{p})$ under the normalization constraint (8.2) is maximized at $\hat{p}_1 = \ldots = \hat{p}_n = n^{-1}$. The maximum EL estimator of μ is given by $\hat{\mu} = \sum_{i=1}^{n} \hat{p}_i y_i = \bar{y}$, which is the sample mean, and the maximum EL estimator of $F(y)$ is given by $\hat{F}(y) = \sum_{i=1}^{n} \hat{p}_i I(y_i \le y) = n^{-1} \sum_{i=1}^{n} I(y_i \le y)$, which is the empirical distribution function. The most important result from Owen (1988) is that the empirical likelihood ratio statistic for μ follows asymptotically a standard χ^2 distribution under some mild moment conditions, which established a nonparametric version of the Wilks' Theorem.

The empirical likelihood method is nonparametric and has many attractive features including data-driven and range-respecting likelihood ratio confidence intervals and effective use of auxiliary information through constrained maximization of the EL function. However, the method cannot be used directly for complex survey data since the resulting estimators are not consistent under the design-based framework unless the sample is taken by simple random sampling without replacement with small sampling fractions (Chen and Qin 1993).

There are two possible formulations of a suitable version of empirical likelihood for complex survey data. The first is the *pseudo empirical likelihood* approach which modifies the EL function (8.1) to take into account the survey design features while keeping the normalization constraint (8.2) and the parameter constraint (8.3) unchanged. Let π_i be the first order inclusion probabilities and $d_i = 1/\pi_i$ be the design weights. Chen and Sitter (1999) proposed to replace (8.1) by the pseudo empirical likelihood function

$$\ell_{\text{CS}}(\mathbf{p}) = \sum_{i \in \mathbf{S}} d_i \log(p_i) ,$$

where the subscript CS indicates Chen and Sitter. The pseudo EL approach is motivated by viewing the survey population as an independent and identically distributed sample of N units from a superpopulation with the census EL function $\sum_{i=1}^{N} \log(p_i)$. The pseudo EL function $\ell_{\text{CS}}(\mathbf{p})$ is the Horvitz-Thompson estimator of the census EL function. Maximizing $\ell_{\text{CS}}(\mathbf{p})$ subject to the normalization constraint (8.2) leads to $\hat{p}_i = \tilde{d}_i(\mathbf{S}) = d_i / \sum_{j \in \mathbf{S}} d_j$ and the maximum pseudo EL estimator of μ_y is given by $\hat{\mu}_{y\text{PEL}} = \sum_{i \in \mathbf{S}} \hat{p}_i y_i = \hat{\mu}_{y\text{H}}$, the Hájek estimator.

The second formulation is the so-called the *sample empirical likelihood*. The formulation uses the same EL function (8.1) from independent samples and the same normalization constraint (8.2) but modifies the parameter constraint to incorporate the survey design weights. We may rewrite the constraint (8.3) as $\sum_{i=1}^{n} p_i(y_i - \mu) = 0$, which corresponds to $E(Y - \mu) = 0$. The sample empirical likelihood approach uses the following modified constraint for the parameter,

$$\sum_{i \in \mathbf{S}} p_i \{d_i(y_i - \mu)\} = 0 . \qquad (8.4)$$

The inclusion of the design weights d_i in constraint (8.4) attempts to alleviate the effect of dependence among sampled units under a complex survey design. Maximizing $\ell(\mathbf{p})$ given by (8.1) subject to the normalization constraint (8.2) gives $\hat{p}_i = n^{-1}$, $i \in \mathbf{S}$, and the maximum sample empirical likelihood estimator of μ is obtained by solving $\sum_{i=1}^{n} \hat{p}_i\{d_i(y_i - \mu)\} = 0$ and is given by $\hat{\mu}_{y\text{SEL}} = \sum_{i=1}^{n} \hat{p}_i y_i = \hat{\mu}_{y\text{H}}$, the Hájek estimator.

Neither the pseudo empirical likelihood nor the sample empirical likelihood involves second order inclusion probabilities which are required for design-based variance estimation. It will become clear in subsequent sections that empirical likelihood ratio confidence intervals or general hypothesis tests using either approach require additional information on the survey design.

8.2 Pseudo Empirical Likelihood for Non-stratified Sampling

The pseudo empirical likelihood function can be formulated differently for non-stratified and stratified sampling designs, which facilitates more flexible use of auxiliary population information. For non-stratified sampling designs, Wu and Rao (2006) proposed to use the following modified version of the pseudo EL function

$$\ell_{\text{WR}}(\mathbf{p}) = n \sum_{i \in \mathbf{S}} \tilde{d}_i(\mathbf{S}) \log(p_i)\,, \tag{8.5}$$

where the subscript WR indicates Wu and Rao and $\tilde{d}_i(\mathbf{S}) = d_i / \sum_{j \in \mathbf{S}} d_j$ are the normalized design weights. The modified version $\ell_{\text{WR}}(\mathbf{p})$ reduces to $\ell(\mathbf{p})$ under simple random sampling and is more convenient for constructing pseudo EL ratio confidence intervals.

We consider two scenarios for pseudo EL inferences, depending on whether auxiliary population information is available.

8.2.1 Pseudo EL Without Auxiliary Information

Without any auxiliary information, the solution to maximizing $\ell_{\text{WR}}(\mathbf{p})$ under the normalization constraint (8.2) is given by $\hat{p}_i = \tilde{d}_i(\mathbf{S})$. The maximum pseudo EL estimator of μ_y is given by $\hat{\mu}_{y\text{PEL}} = \sum_{i \in \mathbf{S}} \hat{p}_i y_i = \sum_{i \in \mathbf{S}} d_i y_i / \sum_{i \in \mathbf{S}} d_i = \hat{\mu}_{y\text{H}}$ as shown in Sect. 8.1.

Construction of pseudo EL ratio confidence intervals for μ_y follows the same route as for parametric likelihood inferences. Let $\ell_{\text{WR}}(\hat{\mathbf{p}}) = n \sum_{i \in \mathbf{S}} \tilde{d}_i(\mathbf{S}) \log(\hat{p}_i)$ be the "global" maximum of the pseudo EL function under the normalization constraint (8.2); Let $\ell_{\text{WR}}(\hat{\mathbf{p}}(\theta)) = n \sum_{i \in \mathbf{S}} \tilde{d}_i(\mathbf{S}) \log(\hat{p}_i(\theta))$ be the "restricted" maximum of the pseudo EL function, where $\hat{\mathbf{p}}(\theta) = (\hat{p}_1(\theta), \ldots, \hat{p}_n(\theta))$ maximize

$\ell_{\rm WR}(\mathbf{p})$ subject to (8.2) and the additional constraint induced by the parameter $\theta_0 = \mu_y$,

$$\sum_{i \in \mathbf{S}} p_i y_i = \theta \,, \tag{8.6}$$

for a fixed value of θ. The notation θ_0 indicates the true value of the unknown parameter θ. The pseudo empirical log-likelihood ratio function for the given θ is computed as

$$r_{\rm WR}(\theta) = \ell_{\rm WR}\big(\hat{\mathbf{p}}(\theta)\big) - \ell_{\rm WR}\big(\hat{\mathbf{p}}\big) \,. \tag{8.7}$$

The asymptotic distribution of $r_{\rm WR}(\theta)$ requires a suitable asymptotic framework as discussed in Sect. 2.4 and the following regularity conditions on the survey design and the study variable y.

C1 The sampling design $\mathscr{P}(\mathbf{S})$ and values of the study variable $\{y_1, \ldots, y_{\rm N}\}$ satisfy $\max_{i \in \mathbf{S}} |y_i| = o_p(n^{1/2})$.
C2 The sampling design $\mathscr{P}(\mathbf{S})$ satisfies $N^{-1} \sum_{i \in \mathbf{S}} d_i - 1 = O_p(n^{-1/2})$.
C3 The Horvitz-Thompson estimator $N^{-1} \sum_{i \in \mathbf{S}} d_i y_i$ of μ_y is asymptotically normally distributed.

Conditions C2 and C3 are expected to hold for commonly encountered sampling designs and survey populations. It follows from Lemma 11.2 in Owen (2001, page 218) that $\max\{|y_1|, \ldots, |y_{\rm N}|\} = o_p(N^{1/2})$ if $N^{-1} \sum_{i=1}^N y_i^2 = O_p(1)$, and condition C1 holds for any sampling designs if $n/N \to c \neq 0$. Condition C3 is part of the foundation for conventional design-based inferences.

Let $V_p(\hat{\mu}_{y\rm H})$ be the design-based variance of the Hájek estimator of μ_y and σ_y^2 be the finite population variance of the y variable.

Theorem 8.1 *Under a suitable asymptotic framework and the regularity conditions C1–C3 on the sampling design and the finite population,*

(a) The adjusted pseudo EL statistic $-2r_{\rm WR}(\theta)/\mathit{deff}_{\rm H}$ *converges in distribution to a* χ^2 *random variable with one degree of freedom when* $\theta = \mu_y$, *where*

$$\mathit{deff}_{\rm H} = V_p(\hat{\mu}_{y\rm H})/\big(\sigma_y^2/n\big)$$

is the design effect for the Hájek estimator.

(b) Under simple random sampling with replacement or simple random sampling without replacement with negligible sampling fractions, $-2r_{\rm WR}(\theta)$ *converges in distribution to a* χ^2 *random variable with one degree of freedom when* $\theta = \mu_y$.

Proof It can be shown that the $\hat{p}_i(\theta)$'s, which maximize $\ell_{\mathrm{WR}}(\mathbf{p})$ subject to (8.2) and (8.6) for a fixed θ, are given by

$$\hat{p}_i(\theta) = \frac{\tilde{d}_i(\mathbf{S})}{1 + \lambda(y_i - \theta)}\,, \quad i \in \mathbf{S}\,,$$

where the Lagrange multiplier λ is the solution to

$$g_0(\lambda) = \sum_{i \in \mathbf{S}} \tilde{d}_i(\mathbf{S}) \frac{y_i - \theta}{1 + \lambda(y_i - \theta)} = 0\,. \tag{8.8}$$

The maximizers $\hat{p}_i(\theta) > 0$ exist for any $\theta \in \big(y_{(1)}\,,\ y_{(n)}\big)$, where $y_{(1)} = \min_{i \in \mathbf{S}} y_i$ and $y_{(n)} = \max_{i \in \mathbf{S}} y_i$. The key argument, which is common for empirical likelihood methods in other areas of statistics, is to show that $\lambda = O_p(n^{-1/2})$ when $\theta = \mu_y$. Rewriting $1/\{1 + \lambda(y_i - \theta)\}$ as $1 - \lambda(y_i - \theta)/\{1 + \lambda(y_i - \theta)\}$ in (8.8), we have

$$\lambda \sum_{i \in \mathbf{S}} \tilde{d}_i(\mathbf{S}) \frac{(y_i - \theta)^2}{1 + \lambda(y_i - \theta)} = \sum_{i \in \mathbf{S}} \tilde{d}_i(\mathbf{S}) y_i - \theta\,. \tag{8.9}$$

Noting that $0 < 1 + \lambda(y_i - \theta) \le 1 + |\lambda| k$, where $k = \max_{i \in \mathbf{S}} |y_i - \theta| = o_p(n^{1/2})$, we further have

$$\frac{|\lambda|}{1 + |\lambda| k} \sum_{i \in \mathbf{S}} \tilde{d}_i(\mathbf{S})(y_i - \theta)^2 \le \Big| \sum_{i \in \mathbf{S}} \tilde{d}_i(\mathbf{S}) y_i - \theta \Big|\,.$$

Since $\sum_{i \in \mathbf{S}} \tilde{d}_i(\mathbf{S}) y_i - \mu_y = O_p(n^{-1/2})$ and $\sum_{i \in \mathbf{S}} \tilde{d}_i(\mathbf{S})(y_i - \mu_y)^2 = \sigma_y^2 + o_p(1) = O_p(1)$ under conditions C2 and C3, we must have $\lambda = O_p(n^{-1/2})$ for any $\theta = \mu_y + O(n^{-1/2})$.

The stochastic order of λ implies that $\lambda(y_i - \theta) = o_p(1)$ and $\{1 + \lambda(y_i - \theta)\}^{-1} = 1 + o_p(1)$, uniformly over i. It follows from (8.9) that

$$\begin{aligned} \lambda &= \Big\{ \sum_{i \in \mathbf{S}} \tilde{d}_i(\mathbf{S}) y_i - \theta \Big\} \Big/ \Big\{ \sum_{i \in \mathbf{S}} \tilde{d}_i(\mathbf{S})(y_i - \theta)^2 \Big\} + o_p\big(n^{-1/2}\big) \\ &= \Big\{ \sum_{i \in \mathbf{S}} \tilde{d}_i(\mathbf{S}) y_i - \theta \Big\} \Big/ \sigma_y^2 + o_p\big(n^{-1/2}\big)\,, \end{aligned}$$

and the result holds for any $\theta = \mu_y + O(n^{-1/2})$. By the second order Taylor series expansion $\log(1 + u) = u - u^2/2 + o(u^2)$, we have

$$\begin{aligned} -2 r_{\mathrm{WR}}(\theta) &= -2n \sum_{i \in \mathbf{S}} \tilde{d}_i(\mathbf{S}) \log\big\{ \hat{p}_i(\theta)/\hat{p}_i \big\} \\ &= 2n \sum_{i \in \mathbf{S}} \tilde{d}_i(\mathbf{S}) \log\big\{ 1 + \lambda(y_i - \theta) \big\} \end{aligned}$$

$$
\begin{aligned}
&= 2n \sum_{i \in \mathbf{S}} \tilde{d}_i(\mathbf{S}) \Big\{ \lambda(y_i - \theta) - \lambda^2 (y_i - \theta)^2/2 \Big\} + o_p(1) \\
&= \Big\{ \sum_{i \in \mathbf{S}} \tilde{d}_i(\mathbf{S}) y_i - \theta \Big\}^2 \Big/ \Big\{ \sigma_y^2/n \Big\} + o_p(1) \,.
\end{aligned}
$$

Part (a) of the theorem follows from the result that $\big\{V_p(\hat{\mu}_{y\mathrm{H}})\big\}^{-1/2}\big\{\sum_{i\in\mathbf{S}} \tilde{d}_i(\mathbf{S})y_i - \mu_y\big\}$ converges in distribution to $N(0,1)$ under Conditions C2 and C3. The denominator σ_y^2/n in deff_H is the variance of the Hájek estimator $\hat{\mu}_{y\mathrm{H}}$ under simple random sampling with replacement. The design effect reduces to one for the two designs stated in Part (b). □

The design effect deff_H is an unknown population parameter involving the variance of the Hájek estimator and the population variance. The asymptotic distribution of $-2r_{\mathrm{WR}}(\theta)/\mathrm{deff}_\mathrm{H}$ at $\theta = \mu_y$ remains the same if the design effect is replaced by a consistent estimator. The $(1-\alpha)$-level pseudo EL ratio confidence interval for μ_y can be constructed as

$$
\Big\{ \theta \mid -2r_{\mathrm{WR}}(\theta)/\mathrm{deff}_\mathrm{H} \le \chi_1^2(\alpha) \Big\} \,, \tag{8.10}
$$

where $\chi_1^2(\alpha)$ is the upper α-quantile of the χ^2 distribution with one degree of freedom. Finding the lower and upper bounds of the interval (8.10) requires an iterative search algorithm; see Sect. 8.4 for details. For certain sampling designs, the design effect can be bypassed through a bootstrap procedure; see Example 10.2 of Chap. 10 for further detail.

8.2.2 Pseudo EL with Auxiliary Information

Suppose that the population means $\mu_\mathbf{x}$ of the vector of auxiliary variables $\mathbf{x}$ are known. The information can be incorporated into pseudo EL inferences through the additional constraints

$$
\sum_{i \in \mathbf{S}} p_i \mathbf{x}_i = \mu_\mathbf{x} \,. \tag{8.11}
$$

Let $\mathbf{u}_i = \mathbf{x}_i - \mu_\mathbf{x}$, $i \in \mathbf{S}$. The constraints (8.11) are equivalent to $\sum_{i\in\mathbf{S}} p_i \mathbf{u}_i = \mathbf{0}$ under the normalization constraint (8.2). The maximum pseudo EL estimator of μ_y is computed as $\hat{\mu}_{y\mathrm{PEL}} = \sum_{i\in\mathbf{S}} \hat{p}_i y_i$, where the $\hat{p}_i$'s maximize $\ell_{\mathrm{WR}}(\mathbf{p})$ subject to both (8.2) and (8.11). It can be shown by using the standard Lagrange multiplier method that

$$\hat{p}_i = \frac{\tilde{d}_i(\mathbf{S})}{1 + \lambda'\mathbf{u}_i}, \quad i \in \mathbf{S}, \tag{8.12}$$

where the vector λ of Lagrange multipliers is the solution to

$$g_1(\lambda) = \sum_{i \in \mathbf{S}} \tilde{d}_i(\mathbf{S}) \frac{\mathbf{u}_i}{1 + \lambda'\mathbf{u}_i} = \mathbf{0}. \tag{8.13}$$

Following similar arguments in the proof of Theorem 8.1 and assuming that condition C1 extends to the $\mathbf{x}$ variables, we can show that $\lambda = O_p(n^{-1/2})$ and

$$\lambda = \Big\{ \sum_{i \in \mathbf{S}} \tilde{d}_i(\mathbf{S})\mathbf{u}_i\mathbf{u}_i' \Big\}^{-1} \sum_{i \in \mathbf{S}} \tilde{d}_i(\mathbf{S})\mathbf{u}_i + o_p\big(n^{-1/2}\big).$$

This leads to $(1+\lambda'\mathbf{u}_i)^{-1} = 1-\lambda'\mathbf{u}_i + o_p(n^{-1/2})$, with the term $o_p(n^{-1/2})$ applying uniformly over i, and the following asymptotic expansion of the maximum pseudo EL estimator of μ_y:

$$\hat{\mu}_{y\mathrm{PEL}} = \sum_{i \in \mathbf{S}} \hat{p}_i y_i = \hat{\mu}_{y\mathrm{H}} + \hat{\mathbf{B}}_{\mathrm{PEL}}'\big(\mu_{\mathbf{x}} - \hat{\mu}_{\mathbf{x}\mathrm{H}}\big) + o_p\big(n^{-1/2}\big), \tag{8.14}$$

where $\hat{\mu}_{y\mathrm{H}}$ and $\hat{\mu}_{\mathbf{x}\mathrm{H}}$ are the Hájek estimators of μ_y and $\mu_{\mathbf{x}}$, respectively, and $\hat{\mathbf{B}}_{\mathrm{PEL}} = \big\{\sum_{i \in \mathbf{S}} \tilde{d}_i(\mathbf{S})\mathbf{u}_i\mathbf{u}_i'\big\}^{-1} \sum_{i \in \mathbf{S}} \tilde{d}_i(\mathbf{S})\mathbf{u}_i y_i = \big\{\sum_{i \in \mathbf{S}} d_i\mathbf{u}_i\mathbf{u}_i'\big\}^{-1} \sum_{i \in \mathbf{S}} d_i\mathbf{u}_i y_i$. The maximum pseudo EL estimator $\hat{\mu}_{y\mathrm{PEL}}$ is asymptotically equivalent to the GREG estimator defined by (5.19). A design-based consistent variance estimator for $\hat{\mu}_{y\mathrm{PEL}}$ can be developed based on the expansion given in (8.14) (Problem 8.1).

The pseudo EL ratio statistic $r_{\mathrm{WR}}(\theta) = \ell_{\mathrm{WR}}(\hat{\mathbf{p}}(\theta)) - \ell_{\mathrm{WR}}(\hat{\mathbf{p}})$ has the same definition as in (8.7) but is computed with different maximizers $\hat{\mathbf{p}}$ and $\hat{\mathbf{p}}(\theta)$. Under the current setting, the "global" maximizer $\hat{p}_i$ are subject to both the normalization constraint (8.2) and the constraints (8.11) on auxiliary variables and are given by (8.12); the "restricted" maximizer $\hat{p}_i(\theta)$ for the given θ are obtained under the constraints (8.2), (8.11) and the additional constraint (8.6) induced by the parameter $\theta_0 = \mu_y$.

Under conditions C1–C3 and assuming that C1 and C3 also apply to each auxiliary variable, it can be shown (Problem 8.2) that the adjusted pseudo EL ratio statistic

$$-2r_{\mathrm{WR}}(\theta)/\mathrm{deff}_{\mathrm{GREG}}$$

converges in distribution to a χ^2 random variable with one degree of freedom when $\theta = \mu_y$. The design effect is given by

$$\mathrm{deff}_{\mathrm{GREG}} = V_p\big(\tilde{\mu}_{y\mathrm{GREG}}\big)/\big(\sigma_r^2/n\big).$$

The details of $\tilde{\mu}_{y\text{GREG}}$ and σ_r^2 are outlined in Problem 8.2. The denominator σ_r^2/n in $\text{deff}_{\text{GREG}}$ is the variance of $\tilde{\mu}_{y\text{GREG}}$ under simple random sampling with replacement. The pseudo EL ratio confidence interval can be similarly constructed as (8.10), with $\hat{\mathbf{p}}$ and $\hat{\mathbf{p}}(\theta)$ computed under the current setting and the estimator of deff_{H} replaced by a consistent design-based estimator of $\text{deff}_{\text{GREG}}$.

8.2.3 Examples of Pseudo EL Ratio Confidence Intervals

The traditional $(1-\alpha)$-level confidence interval for a finite population parameter θ_{N} based on normal approximations has the form of $\big(\hat{\theta} - Z_{\alpha/2}\{v(\hat{\theta})\}^{1/2}\,,\ \hat{\theta} + Z_{\alpha/2}\{v(\hat{\theta})\}^{1/2}\big)$, where $\hat{\theta}$ is a point estimator and $v(\hat{\theta})$ is a variance estimator, and $Z_{\alpha/2}$ is the upper $\alpha/2$ quantile from the $N(0,1)$ distribution. The upper and lower bounds of the interval are symmetrically located towards the point estimator and could land outside the range of the parameter space. For instance, the lower bound could take a negative value when $\theta_{\text{N}} \in [0,1]$ is a population proportion.

Pseudo empirical likelihood ratio confidence intervals are range-respecting, their shape is determined by the sample data, and auxiliary population information can be incorporated through additional constraints. Denote the profile pseudo EL ratio confidence interval (8.10) by $(\hat{\theta}_{\text{L}}, \hat{\theta}_{\text{U}})$. The interval is confined by the range of the data: $y_{(1)} \le \hat{\theta}_{\text{L}} < \hat{\theta}_{\text{U}} \le y_{(n)}$. If y is a binary variable (i.e., $y_i = 1$ or 0) and $\theta_{\text{N}} = \mu_y$ is the population proportion, we have $0 \le \hat{\theta}_{\text{L}} < \hat{\theta}_{\text{U}} \le 1$ unless $y_i = 0$ for all $i \in \mathbf{S}$ or $y_i = 1$ for all $i \in \mathbf{S}$. In the latter cases the pseudo EL ratio confidence interval cannot be constructed. The pseudo EL ratio confidence interval usually performs better than the traditional confidence interval if the underlying distribution of the data is seriously skewed. We present two examples of the finite sample performance of the pseudo EL ratio confidence intervals.

Example 8.1 (Confidence Intervals for the Population Distribution Function) The first example concerns the estimation of the population distribution function $F_y(t)$ at $t = t_\gamma$. This is part of the simulation study reported in Wu and Rao (2006). The finite population follows the model $y_i = 1 + x_i + \varepsilon_i$, $i = 1, \ldots, N = 800$, where $x_i \sim 4 + \exp(1)$ and $\varepsilon_i \sim N(0, \sigma^2)$. The value of σ^2 is chosen to control the correlation between y and x ($\rho(y,x) = 0.5$ for Table 8.1). Samples of size $n = 80$ are taken by the Rao-Sampford PPS sampling method with $\pi_i \propto x_i$.

Table 8.1 contains results for the coverage probability (CP), the lower tail error rate (LE), the upper tail error rate (UE) and the average length (AL) of 95% confidence intervals. The method "Pseudo EL (1)" represents the interval (8.10) without using information on x and the method "Pseudo EL (2)" includes the constraint $\sum_{i\in\mathbf{S}} p_i(x_i - \mu_x) = 0$. All three intervals perform well in terms of CP, LE and UE when $t = t_{0.50}$ and $F_y(t) = 0.50$, with the Pseudo EL (2) interval significantly shorter than the normal interval. For $t = t_{0.10}$ and $F_y(t) = 0.10$, the normal interval becomes unacceptable: the coverage probability is below the nominal value and the tail error rates become extremely unbalanced. The pseudo EL

Table 8.1 95% confidence intervals for the distribution function $F_y(t)$ at $t = t_\gamma$

γ	Method	CP	LE	UE	AL
0.10	Normal approximation	90.7	0.2	9.1	0.134
	Pseudo EL (1)	94.1	1.7	4.2	0.134
	Pseudo EL (2)	94.5	1.9	3.6	0.127
0.50	Normal approximation	95.3	2.4	2.3	0.212
	Pseudo EL (1)	95.5	2.4	2.1	0.208
	Pseudo EL (2)	95.4	2.8	1.8	0.187
0.90	Normal approximation	93.9	5.0	1.1	0.116
	Pseudo EL (1)	95.2	2.7	2.1	0.115
	Pseudo EL (2)	93.5	4.0	2.5	0.110

Table 8.2 95% confidence intervals for μ_y when the population contains many zero values

ρ	Zeros (%)	Method	CP	LE	UE	ALB
0.30	95%	Normal approximation	86.0	0.3	13.7	−0.01
		Pseudo EL	91.0	2.4	6.6	0.09
	70%	Normal approximation	93.0	1.6	5.4	1.11
		Pseudo EL	94.5	2.7	2.8	1.23
0.80	95%	Normal approximation	84.1	0.4	15.5	0.00
		Pseudo EL	92.8	3.1	4.1	0.10
	70%	Normal approximation	93.8	2.1	4.1	1.18
		Pseudo EL	94.9	2.8	2.3	1.24

ratio intervals have much better performance on both aspects and the improvement does not come with any inflation on average length of the interval. ◇

Example 8.2 (Populations with Many Zero Values of the Response Variable) The second example is on confidence intervals for the population mean μ_y when the population contains many zero values of y. The scenarios are commonly encountered in areas such as auditing sampling, where the response variable y denotes the amount of money owned to the government and μ_y is the average amount of excessive claims. Most claims are legitimate with corresponding y's being zero, but a small portion of claims may be excessive. The lower bound of the 95% confidence interval on μ_y is often used to compute the total amount of money owned to the government. Chen et al. (2003) contains discussions on the topic and simulation results on empirical likelihood method under simple random sampling with comparisons to commonly used methods based on parametric mixture models.

When unequal probability sampling procedures are used, which is often the case for auditing sampling, methods based on parametric mixture models become difficult to justify. The pseudo empirical likelihood methods are readily available for constructing confidence intervals for μ_y. Table 8.2 presents results from a simulation study reported in Rao and Wu (2009). The initial settings for the population are the same as the first example on $F_y(t)$, but a random portion of the y's are set to be zero for the final population. All non-zero y's are positive. The value of ρ in the first column of the table denotes the correlation between x and y for the initial population. Samples of size $n = 60$ are taken by the Rao-Sampford πps sampling method with $\pi_i \propto x_i$. The primary measure of interest is the average lower bound (ALB) of the 95% confidence interval. The measure is deemed to be reliable if (1) the coverage probability (CP) of the interval is close to 95% and (2) the lower tail error rate (LE) is close to 2.5%. The pseudo EL ratio confidence interval (without using any auxiliary information) has excellent performance for the population with 70% of the y's being zero. The sample size $n = 60$, however, seems a bit too small for the population with 95% zeros and the pseudo EL interval has issues with under-coverage. The problem disappears if we increase the sample size to $n = 120$ (results not shown in the table). As a comparison, the interval based on normal approximations is virtually useless for populations with 95% zeros. The average lower bound (ALB) of the interval is negative for the population with $\rho = 0.3$ and zero for $\rho = 0.8$. In general, the ALB of the pseudo EL interval is always bigger than the ALB of the normal based interval due to the skewed population distributions. ◇

8.3 Pseudo Empirical Likelihood for Stratified Sampling

Under stratified sampling designs, samples from different strata are selected independently. Auxiliary information might be available at the stratum population level or for the overall population. Both stratum population parameters and the overall population parameters can be of interest for inferences. The pseudo empirical likelihood approach provides a flexible tool for stratified sampling designs.

Let $\{(y_{hi}, \mathbf{x}_{hi}), i \in \mathbf{S}_h\,,\ h = 1, \ldots, H\}$ be the stratified survey sample where $\mathbf{S}_h$ is the set of n_h sampled units for stratum h with inclusion probabilities $\pi_{hi} = P(i \in \mathbf{S}_h)$. Let $d_{hi} = 1/\pi_{hi}$ be the stratum design weights and $n = \sum_{h=1}^{H} n_h$ be the overall sample size. It is assumed that the stratum weights $W_h = N_h/N$ are known. The pseudo empirical (log) likelihood function defined by Wu and Rao (2006) for the stratified sample is given by

$$\ell_{\mathrm{WR}}(\mathbf{p}_1, \ldots, \mathbf{p}_{\mathrm{H}}) = n \sum_{h=1}^{H} W_h \sum_{i\in\mathbf{S}_h} \tilde{d}_{hi}(\mathbf{S}_h) \log(p_{hi})\,,$$

where $\mathbf{p}_h = (p_{h1}, \ldots, p_{hn_h})$ is a discrete probability measure over the sample $\mathbf{S}_h$, and $\tilde{d}_{hi}(\mathbf{S}_h) = d_{hi}/\sum_{j\in\mathbf{S}_h} d_{hj}$, $i \in \mathbf{S}_h$, $h = 1, \ldots, H$. Under stratified simple random sampling with proportional sample size allocations (i.e., $n_h = nW_h$), the pseudo EL function reduces to $\ell(\mathbf{p}_1, \ldots, \mathbf{p}_{\mathrm{H}}) = \sum_{h=1}^{H} \sum_{i\in\mathbf{S}_h} \log(p_{hi})$, which is used by Zhong and Rao (2000) for the same sampling design.

The normalization constraints are imposed for each stratum sample and are given by

$$\sum_{i\in\mathbf{S}_h} p_{hi} = 1\,, \quad h = 1,\ldots,H\,. \tag{8.15}$$

Auxiliary information in the form of known $\mu_{\mathbf{x}}$ for the overall population can be incorporated into inferences through the constrains

$$\sum_{h=1}^{H} W_h \sum_{i\in\mathbf{S}_h} p_{hi}\mathbf{x}_{hi} = \mu_{\mathbf{x}}\,. \tag{8.16}$$

If auxiliary information $\mu_{\mathbf{x}h}$ is available for the hth stratum population, it is possible to include the constraint $\sum_{i\in\mathbf{S}_h} p_{hi}\mathbf{x}_{hi} = \mu_{\mathbf{x}h}$ for the individual stratum.

The maximum pseudo empirical likelihood estimator of μ_y for the overall population is computed as $\hat{\mu}_{y\text{PEL}} = \sum_{h=1}^{H} W_h \sum_{i\in\mathbf{S}_h} \hat{p}_{hi} y_{hi}$, where the $\hat{p}_{hi}$'s maximize $\ell_{\text{WR}}(\mathbf{p}_1,\ldots,\mathbf{p}_{\text{H}})$ subject to (8.15) and (8.16). The maximum pseudo EL estimator of the stratum population mean μ_{yh} is computed as $\hat{\mu}_{yh\text{PEL}} = \sum_{i\in\mathbf{S}_h} \hat{p}_{hi} y_{hi}$, $h = 1,\ldots,H$. The pseudo empirical (log) likelihood ratio statistic for $\theta_0 = \mu_y$ is computed as

$$r_{\text{WR}}(\theta) = \ell_{\text{WR}}\big(\hat{\mathbf{p}}_1(\theta),\ldots,\hat{\mathbf{p}}_{\text{H}}(\theta)\big) - \ell_{\text{WR}}\big(\hat{\mathbf{p}}_1,\ldots,\hat{\mathbf{p}}_{\text{H}}\big)\,,$$

where the $\hat{\mathbf{p}}_h = (\hat{p}_{h1},\ldots,\hat{p}_{hn_h})$, $h = 1,\ldots,H$ are the "global" maximizer of $\ell_{\text{WR}}(\mathbf{p}_1,\ \ldots,\mathbf{p}_{\text{H}})$ under the constraints (8.15) and (8.16), the $\hat{\mathbf{p}}_h(\theta) = (\hat{p}_{h1}(\theta),\ldots,\hat{p}_{hn_h}(\theta))$, $h = 1,\ldots,H$ are the "restricted" maximizer of $\ell_{\text{WR}}(\mathbf{p}_1,\ldots,\mathbf{p}_{\text{H}})$ under the constraints (8.15), (8.16) and the additional constraint induced by the parameter of interest:

$$\sum_{h=1}^{H} W_h \sum_{i\in\mathbf{S}_h} p_{hi} y_{hi} = \theta\,. \tag{8.17}$$

In the absence of auxiliary information on $\mathbf{x}$, the constraints (8.16) should be removed when $\hat{\mathbf{p}}_h$ and $\hat{\mathbf{p}}_h(\theta)$ are computed. For this simple scenario the "global" maximizer is given by $\hat{p}_{hi} = \tilde{d}_{hi}(\mathbf{S}_h)$. When the constrained maximization of $\ell_{\text{WR}}(\mathbf{p}_1,\ldots,\mathbf{p}_{\text{H}})$ involves (8.16) or (8.17), there are computational issues in finding the solutions. See Sect. 8.4.2 for further detail.

Under the asymptotic framework for stratified sampling where the number of strata H is bounded and the stratum sample sizes n_h all go to infinity, and under the extended regularity conditions C1–C3 to cover stratum sampling designs, the adjusted pseudo empirical likelihood ratio statistic

$$-2r_{\text{WR}}(\theta)/\text{deff}_{\text{ST}}$$

converges in distribution to a χ^2 random variable with one degree of freedom when $\theta = \mu_y$. The asymptotic result also holds for scenarios where H is large and

n_h is bounded, as seen in equation (8.21) for the Lagrange multiplier using the combined sample. The design effect deff_{ST} has an explicit expression based on the computational algorithm for stratified sampling and is given in Sect. 8.4.2.

8.4 Computational Procedures

There are three major computational tasks for the pseudo empirical likelihood methods presented in Sects. 8.2 and 8.3: Constrained maximization of the pseudo empirical likelihood function for non-stratified samples, the same maximization problem for stratified samples, and the search for the lower and upper bounds of the pseudo empirical likelihood ratio confidence intervals.

8.4.1 Constrained Maximization for Non-stratified Samples

In general, the maximizer of the pseudo EL function $\ell_{\text{WR}}(\mathbf{p}) = n\sum_{i\in\mathbf{S}}\tilde{d}_i(\mathbf{S})\log(p_i)$ is given by $\hat{p}_i = \tilde{d}_i(\mathbf{S})/(1+\lambda'\mathbf{u}_i)$, $i\in\mathbf{S}$, where the Lagrange multiplier λ is the solution to

$$g(\lambda) = \sum_{i\in\mathbf{S}}\tilde{d}_i(\mathbf{S})\frac{\mathbf{u}_i}{1+\lambda'\mathbf{u}_i} = \mathbf{0}\,. \tag{8.18}$$

The variable $\mathbf{u}_i$ depends on the constraints involved: (i) $\mathbf{u}_i = y_i - \theta$ under constraints (8.2) and (8.6); (ii) $\mathbf{u}_i = \mathbf{x}_i - \mu_{\mathbf{x}}$ under constraints (8.2) and (8.11); and (iii) $\mathbf{u}_i = \big(y_i - \theta, (\mathbf{x}_i - \mu_{\mathbf{x}})'\big)'$ under constraints (8.2), (8.6) and (8.11). The computational task reduces to solving (8.18) to find the value of λ (for a given θ if constraint (8.6) is used).

The modified Newton-Raphson procedures proposed by Chen et al. (2002), briefly described in Sect. 6.2.2 as **Step 0** to **Step 3** for the generalized pseudo empirical likelihood method for calibration weighting, can be used to solve (8.18). The only major change is to use

$$\Delta_1(\lambda) = \sum_{i\in\mathbf{S}}\tilde{d}_i(\mathbf{S})\frac{\mathbf{u}_i}{1+\lambda'\mathbf{u}_i} \quad \text{and} \quad \Delta_2(\lambda) = -\sum_{i\in\mathbf{S}}\tilde{d}_i(\mathbf{S})\frac{\mathbf{u}_i\mathbf{u}_i'}{(1+\lambda'\mathbf{u}_i)^2}\,,$$

and replace "If $1 - \big(\lambda^{(m)} - \delta^{(m)}\big)'\mathbf{x}_i q_i \le 0$ for any $i\in\mathbf{S}$" in **Step 2** by "If $1 + \big(\lambda^{(m)} - \delta^{(m)}\big)'\mathbf{u}_i \le 0$ for any $i\in\mathbf{S}$". In the original algorithm presented in Chen et al. (2002), their Step 2 also checks a related dual objective function. It is more of a theoretical device for the proof of the convergence of the algorithm rather than a required step for practical applications. The algorithm is implemented in R (Wu 2005); see the Appendix A.3 for the R code.

8.4.2 Constrained Maximization for Stratified Samples

The pseudo EL function $\ell_{\mathrm{WR}}(\mathbf{p}_1, \ldots, \mathbf{p}_{\mathrm{H}}) = n\sum_{h=1}^{H} W_h \sum_{i\in\mathbf{S}_h} \tilde{d}_{hi}(\mathbf{S}_h)\log(p_{hi})$, the calibration constraints (8.16) on $\mathbf{x}$ and the parameter constraint (8.17) on θ all involve double summations for the combined stratified samples while the normalization constraints (8.15) are imposed for each stratum sample with a single summation. The computational complexity induced by the two-level constraints can be mitigated through a simple reformulation technique proposed by Wu (2004a).

Let $\mathbf{z}_{hi} = (z_{1hi}, \ldots, z_{(\mathrm{H}-1)hi})'$ be the vector of indicator variables for the first $H-1$ strata, i.e., $z_{jhi} = 1$ if $h = j$ and $i \in \mathbf{S}_h$, and $z_{jhi} = 0$ if $h \neq j$, $j = 1, \ldots, H-1$. In other words, the variable z_{1hi} takes the value 1 for all units in the first stratum and the value 0 elsewhere, and the variable $z_{(\mathrm{H}-1)hi}$ takes the value 1 for all units in the $(H-1)$th stratum and the value 0 otherwise. Let $\mu_{\mathbf{z}} = (W_1, \ldots, W_{\mathrm{H}-1})'$. The normalization constraints (8.15) involving the H single summations at the stratum level can be equivalently reformulated as

$$\sum_{h=1}^{H} W_h \sum_{i\in\mathbf{S}_h} p_{hi} = 1\,, \tag{8.19}$$

$$\sum_{h=1}^{H} W_h \sum_{i\in\mathbf{S}_h} p_{hi}\mathbf{z}_{hi} = \mu_{\mathbf{z}}\,. \tag{8.20}$$

Constraints (8.20) specify that $\sum_{i\in\mathbf{S}_h} p_{hi} = 1$ for $h = 1, \ldots, H-1$. This together with (8.19) implies that $\sum_{i\in\mathbf{S}_h} p_{hi} = 1$ for $h = H$.

Using the standard Lagrange multiplier method, it can be shown that the maximizers of $\ell_{\mathrm{WR}}(\mathbf{p}_1, \ldots, \mathbf{p}_{\mathrm{H}})$ are given by

$$\hat{p}_{hi} = \frac{\tilde{d}_{hi}(\mathbf{S}_h)}{1 + \lambda'\mathbf{u}_{hi}}\,, \quad i \in \mathbf{S}_h, \quad h = 1, \ldots, H\,,$$

where the Lagrange multiplier λ is the solution to the following system of equations

$$g_{st}(\lambda) = \sum_{h=1}^{H} W_h \sum_{i\in\mathbf{S}_h} \tilde{d}_{hi}(\mathbf{S}_h)\frac{\mathbf{u}_{hi}}{1 + \lambda'\mathbf{u}_{hi}} = \mathbf{0}\,. \tag{8.21}$$

The variables $\mathbf{u}_{hi}$ used for computation are given by (i) $\mathbf{u}_{hi} = \big((\mathbf{z}_{hi} - \mu_{\mathbf{z}})', y_{hi} - \theta\big)'$ under constraints (8.15) and (8.17); (ii) $\mathbf{u}_{hi} = \big((\mathbf{z}_{hi} - \mu_{\mathbf{z}})', (\mathbf{x}_{hi} - \mu_{\mathbf{x}})'\big)'$ under constraints (8.15) and (8.16); and (iii) $\mathbf{u}_{hi} = \big((\mathbf{z}_{hi} - \mu_{\mathbf{z}})', y_{hi} - \theta, (\mathbf{x}_{hi} - \mu_{\mathbf{x}})'\big)'$ under constraints (8.15), (8.16) and (8.17).

There are three major results as immediate consequences of the aforementioned reformulation of constraints:

1. The modified Newton-Raphson procedure described in Sect. 8.4.1 for non-stratified samples can be used for solving (8.21). Let

$$\Delta_1(\lambda) = \sum_{h=1}^{H} \sum_{i \in \mathbf{S}_h} W_h \tilde{d}_{hi}(\mathbf{S}_h) \frac{\mathbf{u}_{hi}}{1 + \lambda' \mathbf{u}_{hi}} \ ,$$

$$\Delta_2(\lambda) = -\sum_{h=1}^{H} \sum_{i \in \mathbf{S}_h} W_h \tilde{d}_{hi}(\mathbf{S}_h) \frac{\mathbf{u}_{hi}\mathbf{u}'_{hi}}{(1 + \lambda' \mathbf{u}_{hi})^2} \ .$$

   In addition to the combined data matrix for the stratified samples, we only need to create the stratum indicator variables $\mathbf{z}_{hi}$ and to specify the overall survey weights $W_h \tilde{d}_{hi}(\mathbf{S}_h)$.
2. The unification of the two constrained maximization problems for non-stratified and stratified samples through the use of stratum indicator variables provides an easy route to derive the design effect (deff$_{\text{ST}}$, briefly mentioned at the end of Sect. 8.3), which leads to the conclusion that the scaled pseudo EL ratio statistic $-2r_{\text{WR}}(\theta)/\text{deff}_{\text{ST}}$ has an asymptotic χ^2 distribution with one degree of freedom. It follows from similar arguments in Sect. 8.2.2 (with similar details from Problem 8.2) that

$$\text{deff}_{\text{ST}} = \left\{ \sum_{h=1}^{H} W_h^2 \, V_p \left(\sum_{i \in \mathbf{S}_h} \tilde{d}_{hi}(\mathbf{S}_h) r_{hi} \right) \right\} \left(\frac{\sigma_r^2}{n} \right)^{-1} ,$$

   where $r_{hi} = y_{hi} - \mu_y - \mathbf{B}'\mathbf{u}_{hi}$, $\mathbf{u}_{hi} = \mathbf{z}_{hi} - \mu_{\mathbf{z}}$ if the calibration constraints (8.16) are not used, $\mathbf{u}_{hi} = \big((\mathbf{z}_{hi} - \mu_{\mathbf{z}})', (\mathbf{x}_{hi} - \mu_{\mathbf{x}})' \big)'$ if the constraints (8.16) are used, and $\mathbf{B}$ is the population regression coefficients defined as

$$\mathbf{B} = \left(\sum_{h=1}^{H} \sum_{i=1}^{N_h} \mathbf{u}_{hi}\mathbf{u}'_{hi} \right)^{-1} \sum_{h=1}^{H} \sum_{i=1}^{N_h} \mathbf{u}_{hi}(y_{hi} - \mu_y) \, .$$

   The quantity σ_r^2 is the finite population variance defined over the variable r_{hi} and n is the overall sample size.
3. The reformulation technique for stratified sampling can be used to handle constrained maximization for two-sample empirical likelihood problems (Wu and Yan 2012) and empirical likelihood with multiple samples (Fu et al. 2009). See Problem 8.3 for an example of the two-sample empirical likelihood.

A sample R code for finding the solution to (8.21) using the non-stratified algorithm described in Sect. 8.4.1 is included in the Appendix A.3.3. In practice, a simple check for coding and computational correctness is to see whether $\hat{p}_{hi} > 0$ for all units (hi) and $\sum_{i \in \mathbf{S}_h} \hat{p}_{hi} = 1$ for $h = 1, \ldots, H$. The design effect (deff$_{\text{ST}}$) is an unknown population quantity and needs to be replaced by a design-consistent estimator for constructing the pseudo EL ratio confidence interval $\{\theta \mid -2r_{\mathrm{WR}}(\theta)/\mathrm{deff}_{\mathrm{ST}} \le \chi_1^2(\alpha)\}$. For certain sampling designs, the design effect can be bypassed through a bootstrap method. See Wu and Rao (2010) for further detail.

8.4.3 Profile Pseudo EL Ratio Confidence Intervals

The pseudo EL ratio confidence intervals for the four scenarios discussed in Sects. 8.2 and 8.3 (non-stratified or stratified samples, with or without calibration over the **x** variables) have the general form of $\{\theta \mid -2r(\theta)/\hat{c} \le \chi_1^2(\alpha)\}$, where $r(\theta)$ is the suitably defined pseudo EL ratio function and $\hat{c}$ is the estimated design effect. Without loss of generality, we treat $\hat{c} = 1$ in the following discussion.

The pseudo empirical likelihood ratio function $K(\theta) = -2r(\theta) = -2\{\ell(\hat{\mathbf{p}}(\theta)) - \ell(\hat{\mathbf{p}})\}$ has the following common features for all scenarios:

1. The function $K(\theta)$ has a minimum value 0 at $\theta = \hat{\theta}_{\mathrm{PEL}}$, with $\hat{\theta}_{\mathrm{PEL}}$ being the maximum pseudo EL estimator of θ.
2. The function $K(\theta)$ is convex, monotone decreasing when $\theta < \hat{\theta}_{\mathrm{PEL}}$ and monotone increasing when $\theta > \hat{\theta}_{\mathrm{PEL}}$.
3. The range of feasible values of θ is bounded by $y_{(1)}$ and $y_{(n)}$, the minimum and the maximum of the observed response variable y.

Figure 8.1 provides a graphic representation of the function $K(\theta)$. The horizontal line segment represents the value of $\chi_1^2(\alpha)$ (3.8413 in the graph, corresponding to $\alpha = 0.05$). The locations of the two vertical line segments are the lower and

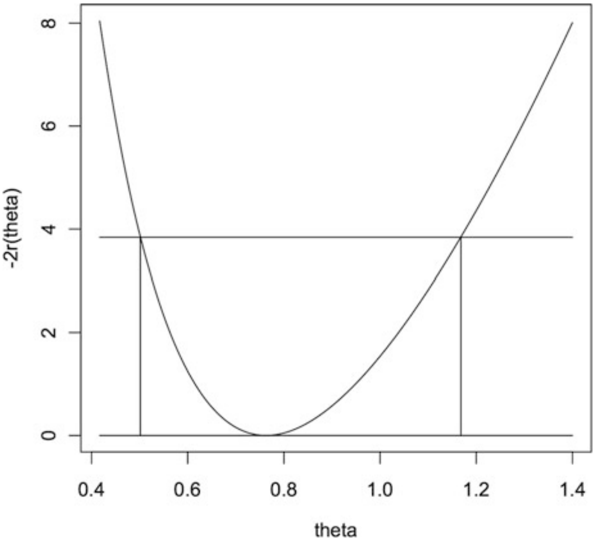

Fig. 8.1 A graphic representation of the pseudo EL ratio function $K(\theta) = -2r(\theta)$ and the bi-section search method for finding the confidence interval $(\hat{\theta}_{\mathrm{L}}, \hat{\theta}_{\mathrm{U}})$

upper bounds, denoted respectively by $\hat{\theta}_{\text{L}}$ and $\hat{\theta}_{\text{U}}$, of the pseudo EL ratio confidence interval for θ.

We provide a brief justification of the features of $K(\theta)$ for non-stratified samples without calibration over the $\mathbf{x}$ variables. Proofs for non-stratified samples with calibration constraints can be adapted from the proof of Theorem 3 in Rao and Wu (2010a). The other two scenarios for stratified samples follow similar arguments. From the proof of Theorem 8.1 we have $\hat{\theta}_{\text{PEL}} = \sum_{i\in\mathbf{S}} \tilde{d}_i(\mathbf{S})y_i$ and

$$K(\theta) = -2r_{\text{WR}}(\theta) = 2n\sum_{i\in\mathbf{S}} \tilde{d}_i(\mathbf{S})\log\{1+\lambda(y_i-\theta)\}\,,$$

where λ satisfies (8.8). It follows that

$$\frac{d}{d\theta}K(\theta) = 2n\sum_{i\in\mathbf{S}} \tilde{d}_i(\mathbf{S})\frac{(\partial\lambda/\partial\theta)(y_i-\theta)-\lambda}{1+\lambda(y_i-\theta)} = -2n\lambda\,.$$

Noting that λ also satisfies (8.9) leads to the conclusion that $dK(\theta)/d\theta < 0$ if $\theta < \hat{\theta}_{\text{PEL}}$ and $dK(\theta)/d\theta > 0$ if $\theta > \hat{\theta}_{\text{PEL}}$.

The lower bound $\hat{\theta}_{\text{L}}$ (and similarly the upper bound $\hat{\theta}_{\text{U}}$) can be found through the following bi-section search algorithm:

```
Step 0  Let $a = y_{(1)}$ and $b = \hat{\theta}_{\text{PEL}}$; Set $\delta = b - a$ and $\varepsilon = 10^{-8}$.
Step 1  Let $M = (a+b)/2$; Compute $K(\theta) = -2r(\theta)$ at $\theta = M$; Set $a = M$ if
        $K(M) \geq \chi_1^2(\alpha)$; Set $b = M$ if $K(M) < \chi_1^2(\alpha)$.
Step 2  Let $\delta = b - a$. If $\delta < \varepsilon$, stop and report $\hat{\theta}_{\text{L}} = M$; Otherwise go to Step 1.
```

The bi-section search algorithm is simple and highly efficient and is guaranteed to converge. A sample R code is included in Appendix A.3.4.

8.5 Generalized Pseudo Empirical Likelihood

Empirical likelihood methods are developed based on a nonparametric discrete probability measure $\mathbf{p} = (p_1, \ldots, p_n)$ over the n sampled units (or several probability measures over multiple samples). Due to the normalization constraint $\sum_{i\in\mathbf{S}} p_i = 1$, it is convenient to handle parameters or auxiliary information in the form of population means such as μ_y and $\mu_{\mathbf{x}}$ or defined through unbiased estimating equations. Pseudo EL ratio confidence intervals for the mean μ_y have been shown to have desirable properties for variables with skewed population distributions.

A substantial part of design-based inference, however, has been developed for population totals, with the Horvitz-Thompson estimator as one of the pillars. The calibration weighting methods discussed in Chap. 6 focus on modifying the design weights $\mathbf{d} = (d_1, \ldots, d_n)$ to incorporate auxiliary information consisting of known population totals $T_{\mathbf{x}}$. The calibrated weights $\mathbf{w} = (w_1, \ldots, w_n)$ provide design-consistent estimators of population totals T_y of study variables. Confidence intervals are constructed through variance estimation and normal approximation to the Z-statistic.

There are two gaps between the pseudo EL methods and the calibration methods. First, the two approaches are disconnected when the population size N is unknown, which is the case for most multi-stage sampling designs. Second, the q-factor used in the distance measure for calibration estimation (Sect. 6.1) does not fit into the formulation of the empirical likelihood function.

It turns out that the gaps between the two approaches can be bridged through the generalized pseudo empirical likelihood method. It is defined in Sect. 6.2.2 that

$$EL(\mathbf{w}, \mathbf{d}) = \sum_{i \in \mathbf{S}} q_i^{-1} \Big\{ d_i \log\Big(\frac{d_i}{w_i}\Big) - d_i + w_i \Big\} ,$$

where $d_i = 1/\pi_i$ are the basic design weights and w_i are the calibrated weights. The q-factor q_i provides an additional tool for efficiency considerations or for achieving other desirable results. Suppose that the population size N is known and $\sum_{i \in \mathbf{S}} w_i = N$ is included as part of the calibration constraints, and suppose that we also choose $q_i = 1$. It can be seen that minimizing the distance measure $EL(\mathbf{w}, \mathbf{d})$ with respect to the w_i under a given set of constraints over the w_i is equivalent to maximizing the pseudo empirical likelihood $\ell_{\text{CS}}(\mathbf{p}) = \sum_{i \in \mathbf{S}} d_i \log(p_i)$ with respect to the p_i under the same set of constraints over the p_i, where $p_i = w_i/N$ and $\sum_{i \in \mathbf{S}} p_i = 1$. The connection provides a justification for the term "*generalized pseudo empirical likelihood*" since the distance measure $EL(\mathbf{w}, \mathbf{d})$ allows N to be unknown, the q_i to be arbitrary and the parameter of interest to be population totals.

Let $T_{\mathbf{x}}$ be the known population totals for the auxiliary variables $\mathbf{x}$. Let $\hat{\mathbf{w}} = (\hat{w}_1, \ldots, \hat{w}_n)$ be the minimizer of $EL(\mathbf{w}, \mathbf{d})$ with respect to the w_i under the calibration constraints

$$\sum_{i \in \mathbf{S}} w_i \mathbf{x}_i = T_{\mathbf{x}} . \tag{8.22}$$

Let $\hat{\mathbf{w}}(\theta) = \big(\hat{w}_1(\theta), \ldots, \hat{w}_n(\theta)\big)$ be the minimizer of $EL(\mathbf{w}, \mathbf{d})$ under both the calibration constraints (8.22) and the parameter constraint

$$\sum_{i \in \mathbf{S}} w_i y_i = \theta \tag{8.23}$$

for a given θ. Tan and Wu (2015) define the generalized pseudo EL ratio function for the population total $\theta = T_y$ of the response variable y as

$$r(\theta) = EL(\hat{\mathbf{w}}, \mathbf{d}) - EL(\hat{\mathbf{w}}(\theta), \mathbf{d}) . \tag{8.24}$$

We assume that the q-factor satisfies $N^{-1} \sum_{i=1}^N q_i^2 = O(1)$ and the conditions C1 and C3 described in Sect. 8.2.1 also apply to the $\mathbf{x}$ variables.

Theorem 8.2 *Under suitable asymptotic framework and the regularity conditions C1–C3 on the sampling design and the finite population involving y and $\mathbf{x}$, the adjusted generalized pseudo EL statistic $-2r(\theta)/C$ converges in distribution to a χ^2 random variable with one degree of freedom when $\theta = T_y$. The scaling constant C is given by*

$$C = V_p(\hat{\eta})/\Big(\sum_{i=1}^{N} q_i e_i^2\Big)\,, \tag{8.25}$$

where $\hat{\eta} = \sum_{i\in\mathbf{S}} d_i e_i$, $e_i = y_i - \mathbf{B}'\mathbf{x}_i$, and $\mathbf{B} = \Big(\sum_{i=1}^{N} q_i \mathbf{x}_i \mathbf{x}_i'\Big)^{-1}\Big(\sum_{i=1}^{N} q_i \mathbf{x}_i y_i\Big)$.

The proof of the theorem is left as an exercise (Problem 8.4). The scaling constant C is an unknown population quantity and needs to be estimated by a consistent estimator $\hat{C}$. The $(1-\alpha)$-level generalized pseudo EL ratio confidence interval for $\theta = T_y$ can be constructed as

$$\Big\{\theta \mid -2r(\theta)/\hat{C} \le \chi_1^2(\alpha)\Big\}\,. \tag{8.26}$$

For single-stage unequal probability sampling designs with small sampling fractions, the interval (8.26) can be constructed through a bootstrap calibration method which does not require the estimated scaling constant $\hat{C}$. See Tan and Wu (2015) for further detail.

An interesting and practically useful result is obtained under rejective sampling and the choice $q_i = \pi_i^{-1} - 1$. Here rejective sampling refers to Poisson sampling conditional on the fixed sample size n (Sect. 4.6). If the inclusion probabilities π_i are used as an auxiliary variable in the calibration constraints (8.22), it can be shown that $C = 1 + o(1)$ if we further assume $n/N \to f \in (0, 1)$ and $\sum_{i=1}^{N} \pi_i(1-\pi_i) \to \infty$. It follows that $-2r(\theta)$ converges in distribution to a χ^2 random variable with one degree of freedom when $\theta = T_y$ (Tan and Wu 2015, Corollary 1). The scaling constant C is not required for the construction of the interval (8.26).

The results on rejective sampling with $q_i = \pi_i^{-1} - 1$ can be extended to high entropy sampling designs. High entropy means that all feasible samples have probabilities as close to uniform as allowed by the design constraints. Rejective sampling achieves maximum entropy subject to pre-specified inclusion probabilities for a fixed sample size. The Rao-Sampford sampling method is a high entropy sampling design and the result $C = 1 + o(1)$ also holds with the choice $q_i = \pi_i^{-1} - 1$ (Tan and Wu 2015, Corollary 2). Hájek (1964, 1981), Berger (1998) and Tan (2013) and additional references therein contain further discussions on rejective sampling and high entropy sampling designs.

8.6 Pseudo Empirical Likelihood and Estimating Equations

The survey weighted estimating equations approach has been discussed in Sect. 7.2 under the setting that the number of parameters in θ is equal to the number of functions in $\mathbf{g}(y_i, \mathbf{x}_i; \theta)$ and the census parameters θ_{N} are uniquely defined as the solution to the system of estimating equations (7.7). This is the so-called *just-identified* scenario, and the discussions in Sect. 7.2 focus primarily on point and variance estimation on θ_{N}.

There are two important directions in which to extend the results of Sect. 7.2. The first is the *over-identified* scenario where the number of estimating functions, r, is bigger than the number of parameters, k. This can happen, for instance, when a single parameter satisfies two moment conditions. If Y follows a Poisson distribution with mean μ, we have $E(Y-\mu) = 0$ and $E\{(Y-\mu)^2-\mu\} = 0$. Another example is the estimation of $\mu = E(Y)$ with a known variance $\sigma_0^2 = Var(Y)$, which leads to $E(Y - \mu) = 0$ and $E\{(Y - \mu)^2 - \sigma_0^2\} = 0$. A practically more important over-identified scenario is the inclusion of additional calibration constraints as discussed in Sects. 8.2 and 8.3. For over-identified scenarios, a direct solution to the survey weighted estimating equations (7.8) is usually not available. Note that scenarios with $r < k$ are not of interest here since the parameter values are not well defined.

The second direction in which to extend the results of Sect. 7.2 is to conduct general linear or nonlinear hypothesis tests on the parameters involving some or all of the components of θ_{N}. Analytic use of survey data may require building a statistical model from an initial set of variables, which may involve hypothesis testing and variable selection. Design-based variable selection using survey data is discussed in Sect. 8.8.

8.6.1 *General Results on Point and Variance Estimation*

Let θ_{N} be $k \times 1$ and $\mathbf{g}(y_i, \mathbf{x}_i; \theta)$ be $r \times 1$ with $r \geq k$. We assume that $\mathbf{g}(y_i, \mathbf{x}_i; \theta)$ is differentiable with respect to θ. Some additional comments are given at the end of Sect. 8.6 on non-smooth estimating functions. The finite population parameters θ_{N} satisfy the census estimating equations (7.7). We consider non-stratified sampling and derive the maximum pseudo EL estimator of θ_{N}. Let

$$\ell_{\mathrm{WR}}(\theta) = n \sum_{i \in \mathbf{S}} \tilde{d}_i(\mathbf{S}) \log\{\hat{p}_i(\theta)\} \tag{8.27}$$

be the profile pseudo empirical likelihood function, where $\hat{\mathbf{p}}(\theta) = (\hat{p}_1(\theta), \cdots, \hat{p}_n(\theta))$ maximizes the pseudo EL function $\ell_{\mathrm{WR}}(\mathbf{p})$ given by (8.5) subject to the normalization constraint (8.2) and the following parameter constraints

$$\sum_{i \in \mathbf{S}} p_i \, \mathbf{g}(y_i, \mathbf{x}_i; \theta) = \mathbf{0} \tag{8.28}$$

for a fixed θ. The maximum pseudo EL estimator of θ_{N}, denoted as $\hat{\theta}_{\mathrm{PEL}}$, is the maximizer of $\ell_{\mathrm{WR}}(\theta)$, i.e., $\hat{\theta}_{\mathrm{PEL}} = \arg\max_{\theta \in \Theta} \ell_{\mathrm{WR}}(\theta)$, where Θ is the parameter space.

Development of asymptotic properties of $\hat{\theta}_{\mathrm{PEL}}$ requires certain regularity conditions similar to conditions C1–C3 presented in Sect. 8.2.1, with y_i being replaced by $\mathbf{g}_i = \mathbf{g}(y_i, \mathbf{x}_i; \theta_{\mathrm{N}})$. For instance, the condition C3 under the current setting states that the Horvitz-Thompson estimator $N^{-1} \sum_{i \in \mathbf{S}} d_i \mathbf{g}_i$ is asymptotically normally distributed.

Let $\Sigma = V_p(N^{-1} \sum_{i \in \mathbf{S}} d_i \mathbf{g}_i)$ be the design-based variance-covariance matrix, which is the same as $V_p\big\{\mathbf{G}_n(\theta_{\mathrm{N}})\big\}$ in (7.9). Let $\mathbf{W}_1 = N^{-1} \sum_{i=1}^{N} \mathbf{g}_i \mathbf{g}_i'$. The notation $\mathbf{W}_1$ as well as $\mathbf{V}_1$ below is to distinguish these quantities from $\mathbf{W}_2$ and $\mathbf{V}_2$ of the next section on sample empirical likelihood. Let $\mathbf{H} = \mathbf{H}_{\mathrm{N}}(\theta_{\mathrm{N}})$ where $\mathbf{H}_{\mathrm{N}}(\theta) = N^{-1} \sum_{i=1}^{N} \partial \mathbf{g}(y_i, \mathbf{x}_i; \theta)/\partial \theta$, which also appears in (7.9). Note that both Σ and $\mathbf{W}_1$ are $r \times r$ matrices but $\mathbf{H}$ is $r \times k$. We have the following results for $\hat{\theta}_{\mathrm{PEL}}$.

Theorem 8.3 *Under suitable regularity conditions on the estimation functions $\mathbf{g}_i = \mathbf{g}(y_i, \mathbf{x}_i; \theta_{\mathrm{N}})$ and the sampling design,*

(a) The maximum pseudo EL estimator $\hat{\theta}_{\mathrm{PEL}}$ is design-consistent.

(b) The maximum pseudo EL estimator $\hat{\theta}_{\mathrm{PEL}}$ is asymptotically normally distributed with mean θ_{N} and the $k \times k$ design-based variance-covariance matrix given by

$$\mathbf{V}_1 = \Big(\mathbf{H}'\mathbf{W}_1^{-1}\mathbf{H}\Big)^{-1} \mathbf{H}'\mathbf{W}_1^{-1} \Sigma \mathbf{W}_1^{-1} \mathbf{H} \Big(\mathbf{H}'\mathbf{W}_1^{-1}\mathbf{H}\Big)^{-1} .$$

Proof We sketch major steps in proving part (b). The proof of part (a) involves tedious arguments similar to those in Zhao et al. (2020c) on sample empirical likelihood. Further details are also available in Zhao and Wu (2019). It can be shown that for a given θ,

$$\ell_{\mathrm{WR}}(\theta) = n \sum_{i \in \mathbf{S}} \tilde{d}_i(\mathbf{S}) \log \big\{\tilde{d}_i(\mathbf{S})\big\} - n \sum_{i \in \mathbf{S}} \tilde{d}_i(\mathbf{S}) \log \big\{1 + \lambda' \mathbf{g}(y_i, \mathbf{x}_i; \theta)\big\} ,$$

where the Lagrange multiplier λ satisfies

$$\sum_{i \in \mathbf{S}} \frac{\tilde{d}_i(\mathbf{S}) \mathbf{g}(y_i, \mathbf{x}_i; \theta)}{1 + \lambda' \mathbf{g}(y_i, \mathbf{x}_i; \theta)} = \mathbf{0} . \tag{8.29}$$

Let $\ell(\theta,\lambda)=\sum_{i\in\mathbf{S}}\tilde{d}_i(\mathbf{S})\log\left\{1+\lambda'\mathbf{g}(y_i,\mathbf{x}_i;\theta)\right\}$. The first major step is to change the problem of finding the maximum pseudo EL estimator $\hat{\theta}_{\rm PEL}$ to an equivalent dual problem of maximizing $\ell(\theta,\lambda)$ with respect to λ, i.e., solving (8.29) for λ, and then minimizing $\ell(\theta,\lambda)$ with respect to θ. Let $(\hat{\theta},\hat{\lambda})=\min_\theta\max_\lambda\ell(\theta,\lambda)$, where $\hat{\theta}=\hat{\theta}_{\rm PEL}=\theta_{\rm N}+o_p(1)$. It can also be shown that $\hat{\lambda}=O_p(n^{-1/2})$.

The second major step is to show that the objective function $\ell(\theta,\lambda)$ can be replaced by a quadratic function in (θ,λ) obtained from a second order Taylor series expansion of $\ell(\theta,\lambda)$ when (θ,λ) is in the neighborhood of $(\theta_{\rm N},\mathbf{0})$ without changing the solutions asymptotically. The new objective function is given by (Problem 8.5)

$$L(\theta,\lambda)=(\theta-\theta_{\rm N})'\mathbf{H}'\lambda+\left\{\mathbf{G}_n(\theta_{\rm N})\right\}'\lambda-\frac{1}{2}\lambda'\mathbf{W}_1\lambda\,, \tag{8.30}$$

where $\mathbf{G}_n(\theta_{\rm N})=N^{-1}\sum_{i\in\mathbf{S}}d_i\mathbf{g}_i$ following the notation in (7.8). The solution $(\hat{\theta},\hat{\lambda})$ is asymptotically equivalent to $\min_\theta\max_\lambda L(\theta,\lambda)$, which can be found by solving $\partial L(\theta,\lambda)/\partial\theta=\mathbf{0}$ and $\partial L(\theta,\lambda)/\partial\lambda=\mathbf{0}$, i.e., the solution to

$$\begin{pmatrix}\mathbf{0}\\ -\mathbf{G}_n(\theta_{\rm N})\end{pmatrix}-\begin{pmatrix}\mathbf{0} & \mathbf{H}'\\ \mathbf{H} & -\mathbf{W}_1\end{pmatrix}\begin{pmatrix}\theta-\theta_{\rm N}\\ \lambda\end{pmatrix}=\mathbf{0}\,. \tag{8.31}$$

It follows that the solution $\tilde{\theta}$ from (8.31) satisfies

$$\tilde{\theta}-\theta_{\rm N}=-\left(\mathbf{H}'\mathbf{W}_1^{-1}\mathbf{H}\right)^{-1}\mathbf{H}'\mathbf{W}_1^{-1}\mathbf{G}_n(\theta_{\rm N})\,. \tag{8.32}$$

The results of part (b) of the theorem follow immediately due to the asymptotic equivalence between $\hat{\theta}_{\rm PEL}$ and $\tilde{\theta}$. □

For the special just-identified cases where $r=k$, the maximum pseudo EL estimator $\hat{\theta}_{\rm PEL}$ is identical to the survey weighted estimating equations estimator discussed in Sect. 7.2.1 and the asymptotic design-based variance $\mathbf{V}_1$ given in Theorem 8.3 reduces to the same design-based variance given in (7.9).

Calibration constraints are commonly used in survey sampling to create an over-identified estimating equations system. Let

$$\mathbf{g}(y_i,\mathbf{x}_i;\theta)=\begin{pmatrix}\mathbf{g}_1(y_i,\mathbf{x}_i;\theta)\\ \mathbf{g}_2(\mathbf{x}_i)\end{pmatrix}\,,$$

where $\mathbf{g}_1(y_i,\mathbf{x}_i;\theta)$ are the $k\times 1$ estimating functions for defining the $k\times 1$ parameters θ and $\mathbf{g}_2(\mathbf{x}_i)=\mathbf{h}(\mathbf{x}_i)-N^{-1}\sum_{j=1}^N\mathbf{h}(\mathbf{x}_i)$ are the $(r-k)\times 1$ calibration variables satisfying $N^{-1}\sum_{i=1}^N\mathbf{g}_2(\mathbf{x}_i)=\mathbf{0}$. The parameter constraints (8.28) can be rewritten as

$$\sum_{i\in\mathbf{S}} p_i\,\mathbf{g}_1(y_i,\mathbf{x}_i;\theta)=\mathbf{0}\,, \tag{8.33}$$

$$\sum_{i\in\mathbf{S}} p_i\,\mathbf{g}_2(\mathbf{x}_i)=\mathbf{0}\,. \tag{8.34}$$

The maximum pseudo EL estimator $\hat{\theta}_{\text{PEL}}$ under the this setting can be computed directly without using profile empirical likelihood. Let $\hat{\mathbf{p}}=(\hat{p}_1,\cdots,\hat{p}_n)$ be the maximizer of $\ell_{\text{WR}}(\mathbf{p})$ under the normalization constraint (8.2) and the calibration constraints (8.34). The $\hat{p}_i$'s can be viewed as the calibration weights. The maximum pseudo EL estimator $\hat{\theta}_{\text{PEL}}$ solves the "survey weighted estimating equations" specified by (8.33) with the p_i's replaced by the $\hat{p}_i$'s.

Variance estimation for $\hat{\theta}_{\text{PEL}}$ requires estimation of the components involved in $\mathbf{V}_1$. Estimation of $\mathbf{H}$ and $\Sigma=V_p(N^{-1}\sum_{i\in\mathbf{S}}d_i\mathbf{g}_i)$ has been discussed in Sect. 7.2.2. The component $\mathbf{W}_1$ can be estimated by $\hat{\mathbf{W}}_1=N^{-1}\sum_{i\in\mathbf{S}}d_i\hat{\mathbf{g}}_i\hat{\mathbf{g}}_i'$, where $\hat{\mathbf{g}}_i=\mathbf{g}(y_i,\mathbf{x}_i;\hat{\theta}_{\text{PEL}})$. The factor N^{-1} can be dropped from all involved components for the computation, since they all cancel out in the expression for $\mathbf{V}_1$.

8.6.2 General Results on Hypothesis Tests

We consider general cases where $r\geq k$ and the inferential problem is to test H_0: $\theta_{\text{N}}=\theta_{\text{N0}}$ versus H_1: $\theta_{\text{N}}\neq\theta_{\text{N0}}$ for a given θ_{N0}. The pseudo EL ratio statistic is computed as

$$r_{\text{PEL}}(\theta_{\text{N0}})=-2\Big\{\ell_{\text{WR}}(\theta_{\text{N0}})-\ell_{\text{WR}}(\hat{\theta}_{\text{PEL}})\Big\}\,.$$

Under the same regularity conditions as in Theorem 8.3, it can be shown that

$$r_{\text{PEL}}(\theta_{\text{N0}})=\mathbf{Q}'\Delta_1\mathbf{Q}+o_p(1)\,,$$

where $\mathbf{Q}$ follows the standard multivariate normal distribution $\mathbf{N}(\mathbf{0},\mathbf{I}_r)$ and

$$\Delta_1=n\Sigma^{1/2}\mathbf{W}_1^{-1}\mathbf{H}\Big(\mathbf{H}'\mathbf{W}_1^{-1}\mathbf{H}\Big)^{-1}\mathbf{H}'\mathbf{W}_1^{-1}\Sigma^{1/2}\,. \tag{8.35}$$

The matrix Δ_1 reduces to $\Delta_1=n\Sigma^{1/2}\mathbf{W}_1^{-1}\Sigma^{1/2}$ for just-identified cases with $r=k$. Note that there is a factor n in Δ_1 which comes from the definition of $\ell_{\text{WR}}(\theta)$. Since $\Sigma=O(n^{-1})$, we have $\Delta_1=O(1)$. The quadratic form $\mathbf{Q}'\Delta_1\mathbf{Q}$ can be re-expressed as

$$\mathbf{Q}'\Delta_1\mathbf{Q}=\sum_{j=1}^{k}\delta_i\chi_j^2(1)\,,$$

where $\delta_j, j = 1, \cdots, k$ are the non-zero eigenvalues of Δ_1 and $\chi_j^2(1), j = 1, \cdots, k$ are independent random variables all following a chisquare distribution with one degree of freedom. In other words, the pseudo EL ratio statistic $r_{\text{PEL}}(\theta_{\text{N0}})$ follows asymptotically a weighted χ^2 distribution. Proof of the result involves using $L(\theta, \lambda)$ defined in (8.30) to approximate $\ell_{\text{WR}}(\theta)$ at $\theta = \hat{\theta}_{\text{PEL}}$ and $\theta = \theta_{\text{N0}}$. Details can be found in Zhao et al. (2020c) where the focus is on sample empirical likelihood, to be discussed in the next section.

We further consider a general hypothesis H_0: $\mathbf{K}(\theta_{\text{N}}) = \mathbf{0}$ versus H_1: $\mathbf{K}(\theta_{\text{N}}) \neq \mathbf{0}$, where $\mathbf{K}(\theta_{\text{N}}) = \mathbf{0}$ imposes k_1 $(\leq k)$ linear or nonlinear constraints on the $k \times 1$ parameters θ_{N}. Let $\hat{\theta}^*_{\text{PEL}} = \arg\max_{\theta \in \Theta^*} \ell_{\text{WR}}(\theta)$ be the restricted maximum pseudo EL estimator of θ_{N} under the restricted parameter space $\Theta^* = \{\theta \mid \theta \in \Theta \text{ and } \mathbf{K}(\theta) = \mathbf{0}\}$. The pseudo EL statistic under H_0: $\mathbf{K}(\theta_{\text{N}}) = \mathbf{0}$ is computed as

$$r_{\text{PEL}}(\theta_{\text{N}} \mid H_0) = -2\Big\{\ell_{\text{WR}}(\hat{\theta}^*_{\text{PEL}}) - \ell_{\text{WR}}(\hat{\theta}_{\text{PEL}})\Big\},$$

where $\hat{\theta}_{\text{PEL}} = \arg\max_{\theta \in \Theta} \ell_{\text{WR}}(\theta)$ is the global maximum pseudo EL estimator. It can be shown that

$$r_{\text{PEL}}(\theta_{\text{N}} \mid H_0) = \mathbf{Q}'\Delta_1^*\mathbf{Q} + o_p(1),$$

where $\mathbf{Q} \sim \mathbf{N}(\mathbf{0}, \mathbf{I}_r)$ and

$$\Delta_1^* = n\Sigma^{1/2}\mathbf{W}_1^{-1}\mathbf{H}\Gamma_1\Phi'\big(\Phi\Gamma_1\Phi'\big)^{-1}\Phi\Gamma_1\mathbf{H}'\mathbf{W}_1^{-1}\Sigma^{1/2}, \tag{8.36}$$

where $\Gamma_1 = \mathbf{H}'\mathbf{W}_1^{-1}\mathbf{H}$, $\Phi = \Phi(\theta_{\text{N}})$ and $\Phi(\theta) = \partial\mathbf{K}(\theta)/\partial\theta$. Note that Φ is a $k_1 \times k$ matrix and Δ_1^* is a $r \times r$ matrix with rank k_1.

Proof of the result relies once again on using $L(\theta, \lambda)$ to approximate $\ell_{\text{WR}}(\theta)$ and handles the restriction $\mathbf{K}(\theta) = \mathbf{0}$ through the Lagrange multiplier method. The restricted maximum pseudo EL estimator $\hat{\theta}^*_{\text{PEL}}$ is asymptotically equivalent to $\tilde{\theta}$ where

$$(\tilde{\theta}, \tilde{\lambda}, \tilde{\eta}) = \arg\min_{\theta,\lambda,\eta} \Big\{L(\theta, \lambda) + \eta\mathbf{K}(\theta)\Big\}.$$

The solutions $(\tilde{\theta}, \tilde{\lambda}, \tilde{\eta})$ can be found by taking partial derivatives of $L(\theta, \lambda) + \eta\mathbf{K}(\theta)$ with respect to θ, λ and η and setting them to be zeros. Details can be found in Zhao et al. (2020c).

The hypothesis H_0: $\theta_{\text{N}} = \theta_{\text{N0}}$ discussed at the beginning of this section can be rewritten as H_0: $\mathbf{I}_k(\theta_{\text{N}} - \theta_{\text{N0}}) = \mathbf{0}$, which leads to $\Phi(\theta) = \mathbf{I}_k$. The matrix Δ_1^* given by (8.36) reduces to Δ_1 given by (8.35). Hypotheses commonly encountered in practice have a linear form, i.e., H_0: $\mathbf{A}\theta_{\text{N}} = \mathbf{b}$, where $\mathbf{A}$ is a $k_1 \times k$ constant matrix and $\mathbf{b}$ is a $k_1 \times 1$ vector with $1 \leq k_1 \leq k$. We have $\Phi = \mathbf{A}$ for this type of problems.

Implementation of the pseudo empirical likelihood ratio tests discussed in this section requires estimation of components $\mathbf{W}_1$, $\mathbf{H}$ and the design-based variance-covariance matrix Σ. Both $\mathbf{W}_1$ and $\mathbf{H}$ are population totals and can be estimated by using the Horvitz-Thompson estimator. The distribution of the quadratic form $\mathbf{Q}'\Delta_1\mathbf{Q}$ or $\mathbf{Q}'\Delta_1^*\mathbf{Q}$ can be estimated through a simulation-based approach. Our discussion in this section does not cover parameters defined through non-smooth estimating functions such as population quantiles. Further details can be found in Zhao and Wu (2019).

8.7 Sample Empirical Likelihood and Estimating Equations

Sample empirical likelihood was first briefly mentioned in Chen and Kim (2014) as an alternative formulation to the population empirical likelihood discussed in their paper. The formulation uses the standard empirical likelihood function $\ell(\mathbf{p}) = \sum_{i\in\mathbf{S}} \log(p_i)$ subject to the normalization constraint $\sum_{i\in\mathbf{S}} p_i = 1$, and incorporates the survey design features through the survey weighted estimating equations,

$$\sum_{i\in\mathbf{S}} p_i \,\{d_i\mathbf{g}(y_i, \mathbf{x}_i; \theta)\} = \mathbf{0}\,. \tag{8.37}$$

There are parallel results to those presented in Sect. 8.6 with slightly different structures in defining the matrix $\mathbf{W}_1$. The major advantage of the sample empirical likelihood approach is the unification of computational procedures with those of the standard empirical likelihood (Owen 1988; Qin and Lawless 1994). The empirical likelihood approach discussed by Berger and De La Riva Torres (2016) and Oguz-Alper and Berger (2016) is closely related to the sample empirical likelihood.

8.7.1 General Results on Point and Variance Estimation

The profile sample empirical likelihood function for θ is given by (with the constant term $-n\log(n)$ omitted)

$$\ell(\theta) = -\sum_{i\in\mathbf{S}} \log\Big[1 + \lambda'\{d_i\mathbf{g}(y_i, \mathbf{x}_i; \theta)\}\Big]\,,$$

where the Lagrange multiplier $\lambda = \lambda(\theta)$ with the given θ is the solution to

$$\frac{1}{n}\sum_{i\in\mathbf{S}} \frac{d_i\mathbf{g}(y_i, \mathbf{x}_i; \theta)}{1 + \lambda'\{d_i\mathbf{g}(y_i, \mathbf{x}_i; \theta)\}} = \mathbf{0}\,. \tag{8.38}$$

The computational procedures are identical to standard empirical likelihood with independent data if we treat $d_i\mathbf{g}(y_i, \mathbf{x}_i; \theta)$ as the estimating functions. The maximum sample empirical likelihood estimator of θ_{N} is the maximizer of $\ell(\theta)$, denoted as $\hat{\theta}_{\mathrm{SEL}} = \arg\max_\theta \ell(\theta)$.

Under suitable regularity conditions, the estimator $\hat{\theta}_{\mathrm{SEL}}$ is design consistent with design-based variance-covariance matrix given by

$$\mathbf{V}_2 = \left(\mathbf{H}'\mathbf{W}_2^{-1}\mathbf{H}\right)^{-1}\mathbf{H}'\mathbf{W}_2^{-1}\Sigma\mathbf{W}_2^{-1}\mathbf{H}\left(\mathbf{H}'\mathbf{W}_2^{-1}\mathbf{H}\right)^{-1}, \tag{8.39}$$

where $\mathbf{H}$ and Σ are defined in the previous section, and $\mathbf{W}_2 = nN^{-2}\sum_{i=1}^N d_i\mathbf{g}_i\mathbf{g}_i'$ with $\mathbf{g}_i = \mathbf{g}(y_i, \mathbf{x}_i; \theta_{\mathrm{N}})$. For special cases where $r = k$, the maximum sample EL estimator is identical to the survey weighted estimating equations estimator presented in Sect. 7.2.1. The asymptotic variance $\mathbf{V}_2$ reduces to the sandwich variance given in (7.9). The component $\mathbf{W}_2$ can be estimated by $\hat{\mathbf{W}}_2 = nN^{-2}\sum_{i\in\mathbf{S}} d_i^2\hat{\mathbf{g}}_i\hat{\mathbf{g}}_i'$, where $\hat{\mathbf{g}}_i = \mathbf{g}(y_i, \mathbf{x}_i; \hat{\theta}_{\mathrm{SEL}})$. Once again, the factors N^{-1} and N^{-2} can be dropped from all involved components for the purpose of computation.

8.7.2 General Results on Hypothesis Tests

The sample EL ratio statistic for testing H_0: $\theta_{\mathrm{N}} = \theta_{\mathrm{N0}}$ versus H_1: $\theta_{\mathrm{N}} \neq \theta_{\mathrm{N0}}$ is computed as

$$r_{\mathrm{SEL}}(\theta_{\mathrm{N0}}) = -2\left\{\ell(\theta_{\mathrm{N0}}) - \ell(\hat{\theta}_{\mathrm{SEL}})\right\}.$$

Under the same regularity conditions of Theorem 8.3, it can be shown that $r_{\mathrm{SEL}}(\theta_{\mathrm{N0}}) = \mathbf{Q}'\Delta_2\mathbf{Q} + o_p(1)$, where $\mathbf{Q} \sim \mathbf{N}(\mathbf{0}, \mathbf{I}_r)$ and

$$\Delta_2 = n\Sigma^{1/2}\mathbf{W}_2^{-1}\mathbf{H}\left(\mathbf{H}'\mathbf{W}_2^{-1}\mathbf{H}\right)^{-1}\mathbf{H}'\mathbf{W}_2^{-1}\Sigma^{1/2}. \tag{8.40}$$

The matrix Δ_2 reduces to $\Delta_2 = n\Sigma^{1/2}\mathbf{W}_2^{-1}\Sigma^{1/2}$ for just-identified cases with $r = k$.

Theorem 8.4 *Under suitable regularity conditions on the estimating functions $\mathbf{g}_i = \mathbf{g}(y_i, \mathbf{x}_i; \theta_{\mathrm{N}})$ and the sampling design, the sample EL ratio statistic $r_{\mathrm{SEL}}(\theta_{\mathrm{N0}})$ converges in distribution to a standard chisquare random variable with k degrees of freedom if $r = k$ and the survey design is single-stage PPS sampling with a negligible sampling fraction.*

The result of the theorem follows from Part (c) of Problem 7.1 which shows that $\Sigma = n^{-1}\mathbf{W}_2$ and hence $\Delta_2 = \mathbf{I}_k$.

For a general hypothesis H_0: $\mathbf{K}(\theta_{\mathrm{N}}) = \mathbf{0}$ versus H_1: $\mathbf{K}(\theta_{\mathrm{N}}) \neq \mathbf{0}$, let $\hat{\theta}^*_{\mathrm{SEL}} = \arg\max_{\theta\in\Theta^*} \ell(\theta)$ be the restricted maximum sample EL estimator of θ_{N} under the restricted parameter space $\Theta^* = \{\theta \mid \theta \in \Theta \quad \text{and} \quad \mathbf{K}(\theta) = \mathbf{0}\}$. The sample EL statistic under H_0: $\mathbf{K}(\theta_{\mathrm{N}}) = \mathbf{0}$ is computed as

$$r_{\mathrm{SEL}}(\theta_{\mathrm{N}} \mid H_0) = -2\Big\{\ell(\hat{\theta}^*_{\mathrm{SEL}}) - \ell(\hat{\theta}_{\mathrm{SEL}})\Big\}\,.$$

It can be shown that $r_{\mathrm{SEL}}(\theta_{\mathrm{N}} \mid H_0) = \mathbf{Q}'\Delta_2^*\mathbf{Q} + o_p(1)$, where $\mathbf{Q} \sim \mathbf{N}(\mathbf{0}, \mathbf{I}_r)$ and

$$\Delta_2^* = n\Sigma^{1/2}\mathbf{W}_2^{-1}\mathbf{H}\Gamma_2\Phi'\big(\Phi\Gamma_2\Phi'\big)^{-1}\Phi\Gamma_2\mathbf{H}'\mathbf{W}_2^{-1}\Sigma^{1/2}\,,$$

where $\Gamma_2 = \mathbf{H}'\mathbf{W}_2^{-1}\mathbf{H}$ and Φ is defined as before. If $r = k$ and the survey design is single-stage PPS sampling with a small sampling fraction, the sample EL ratio statistic $r_{\mathrm{SEL}}(\theta_{\mathrm{N}} \mid H_0)$ follows asymptotically a standard χ^2 distribution with k_1 degrees of freedom. Zhao et al. (2020c) contains a proof of the results as well as other technical details on the sample empirical likelihood method. Zhao and Wu (2019) presents a comparative study and discussions on practical issues with the pseudo and the sample empirical likelihood methods.

8.8 Design-Based Variable Selection Methods

Complex surveys often collect information on a large number of variables. Some of those variables measure basic characteristics of the units and some are specifically designed for broad scientific objectives. For instance, the International Tobacco Control (ITC) survey to be discussed in Chap. 12 collects data on many variables related to demographic, psychosocial, behavioral and health aspects of the units as well as measures of knowledge and attitude towards smoking. Variable selection is an important problem at the initial stage of model building to identify relevant factors for a particular response variable such as addiction or quitting behaviors.

In the non-survey context, the basic aim of variable selection is to identify variables in a regression model for which the coefficients are zero. Design-based variable selection using survey data focuses on the finite population regression coefficients θ_{N} defined as the solution to the census estimating equations. The components of θ_{N} are usually not exactly equal to zero even if the corresponding superpopulation parameters are zero. The components are typically of the order $O(N^{-1/2})$ if the model parameters are zero and the model holds for the finite population. We consider practical scenarios where N is very large and certain components of θ_{N} can be treated as zero. The inferential objectives are to identify the zero components and to obtain estimates for the non-zero components using the survey dataset.

Under the model-based framework, variable selection via the penalized empirical likelihood has been studied by several authors, including Tang and Leng (2010) and Leng and Tang (2012). It turns out that both the pseudo empirical likelihood and the sample empirical likelihood can be modified through suitable penalization to provide design-based variable selection with oracle properties. That is, with probability tending to 1 in the asymptotic framework, the methods correctly identify the known zero components of the finite population parameters.

8.8.1 Penalized Pseudo Empirical Likelihood

Variable selection was first discussed under the framework of least squares estimation for linear regression models. The *least absolute shrinkage and selection operator* (LASSO) proposed by Tibshirani (1996) is a well-known penalized least squares method for variable selection. For likelihood-based procedures, including variable selection for generalized linear models, the *smoothly clipped absolute deviation* (SCAD) penalty function proposed by Fan and Li (2001) has been shown to achieve variable selection and unbiased parameter estimation simultaneously.

Let $p_\tau(\cdot)$ be a pre-specified penalty function with regularization parameter τ. Let the parameters θ be a $k \times 1$ vector defined through estimating functions $\mathbf{g}(y, \mathbf{x}; \theta)$. Following the proof of Theorem 8.3 and ignoring the constant term $n\sum_{i\in\mathbf{S}} \tilde{d}_i(\mathbf{S}) \log\left\{\tilde{d}_i(\mathbf{S})\right\}$, the penalized pseudo empirical likelihood function is defined as

$$\ell_{\text{PPEL}}(\theta) = -n\sum_{i\in\mathbf{S}} \tilde{d}_i(\mathbf{S}) \log\left\{1 + \lambda'\mathbf{g}(y_i, \mathbf{x}_i; \theta)\right\} - n\sum_{j=1}^{k} p_\tau(|\theta_j|)\,,$$

where θ_j is the jth component of θ and the Lagrange multiplier λ is the solution to (8.29). We consider the continuous differentiable SCAD penalty function $p_\tau(t)$ which satisfies $p_\tau(0) = 0$ with its first order derivative given by

$$p'_\tau(t) = \tau\left\{I(t \le \tau) + \frac{(a\tau - t)_+}{(a-1)\tau} I(t > \tau)\right\},$$

where $(b)_+ = b$ if $b \ge 0$ and $(b)_+ = 0$ if $b < 0$. The penalty function contains two regularization parameters: a and τ. The choice $a = 3.7$ works well under the universal thresholding $\tau = \{2\log(k)\}^{1/2}$ when $k \le 100$. More refined data-driven choice of (a, τ) can be determined using criteria such as BIC or generalized cross-validation. See Fan and Li (2001) and Tang and Leng (2010) for further discussion.

The maximum penalized pseudo empirical likelihood estimator of θ_{N} is given by $\hat{\theta}_{\text{PPEL}} = \arg\max_\theta \ell_{\text{PPEL}}(\theta)$. A major advantage of using the SCAD penalty function is an efficient algorithm for finding $\hat{\theta}_{\text{PPEL}}$ based on a local quadratic approximation. For an initial value $\theta_0 = (\theta_{10}, \cdots, \theta_{k0})'$, we set $\hat{\theta}_j = 0$ if θ_{j0} is very close to 0. Otherwise we have

$$\left\{\frac{d}{d\theta_j} p_\tau(|\theta_j|)\right\}\bigg|_{\theta_j=\theta_{j0}} = p'_\tau(|\theta_{j0}|)\,\text{sign}(\theta_{j0}) \doteq p'_\tau(|\theta_{j0}|)\frac{\theta_j}{|\theta_{j0}|}\,,$$

where θ_j is close to θ_{j0}, and the first order Taylor series expansion further leads to

$$\begin{aligned} p_\tau(|\theta_j|) &\doteq p_\tau(|\theta_{j0}|) + \left\{p'_\tau(|\theta_{j0}|)/|\theta_{j0}|\right\}\theta_j(\theta_j - \theta_{j0}) \\ &\doteq p_\tau(|\theta_{j0}|) + \frac{1}{2}\left\{p'_\tau(|\theta_{j0}|)/|\theta_{j0}|\right\}(\theta_j^2 - \theta_{j0}^2)\,. \end{aligned}$$

The Newton-Raphson procedure can now be used to find $\hat{\theta}_{\text{PPEL}}$ based on the local quadratic approximation to $\ell_{\text{PPEL}}(\theta)$. Zhao and Wu (2019) and Zhao et al. (2020c) contain further details on the computational algorithm.

8.8.2 Penalized Sample Empirical Likelihood

The penalized sample empirical likelihood function (omitting the constant term $-n\log(n)$) is defined as

$$\ell_{\text{PSEL}}(\theta) = -\sum_{i\in\mathbf{S}} \log\left[1 + \lambda'\{d_i\mathbf{g}(y_i, \mathbf{x}_i; \theta)\}\right] - n\sum_{j=1}^{k} p_\tau(|\theta_j|)\,,$$

where the Lagrange multiplier λ with the given θ is the solution to (8.38) and $p_\tau(\cdot)$ is a pre-specified penalty function. The sample empirical likelihood treats the survey weights d_i as part of the estimating functions and uses $d_i\mathbf{g}(y_i, \mathbf{x}_i; \theta)$ in the constraints (8.37) under the formulation of standard empirical likelihood for independent sample data. Computational algorithms for standard empirical likelihood (Owen 2001) can be used to find $\hat{\theta}_{\text{PSEL}} = \arg\max_\theta \ell_{\text{PSEL}}(\theta)$ if we use the SCAD penalty function with the local quadratic approximation discussed in Sect. 8.8.1.

The design-based penalized pseudo empirical likelihood and penalized sample empirical likelihood methods retain the oracle property for variable selection. Suppose that $\theta_{\text{N}} = (\theta'_{\text{N}1}, \theta'_{\text{N}2})'$ and $\theta_{\text{N}2} = \mathbf{0}$. Let $\hat{\theta} = (\hat{\theta}'_1, \hat{\theta}'_2)'$ be the estimator $\hat{\theta}_{\text{PPEL}}$ or $\hat{\theta}_{\text{PSEL}}$ following the same partition. Then with probability approaching to 1 in the asymptotic framework, we have $\hat{\theta}_2 = \mathbf{0}$. Further technical details can be found in Zhao et al. (2020c).

8.9 Problems

8.1 (Variance Estimation for the Maximum Pseudo EL Estimator) With known population means $\mu_{\mathbf{x}}$, the maximum pseudo EL estimator $\hat{\mu}_{y\text{PEL}}$ has an asymptotic expansion given in (8.14). Note that the expansion involves two Hájek estimators $\hat{\mu}_{y\text{H}}$ and $\hat{\mu}_{\mathbf{x}\text{H}}$.

(a) Derive the asymptotic design-based variance formula $AV_p\big(\hat{\mu}_{y\text{PEL}}\big)$.
(b) Develop a consistent design-based variance estimator for $\hat{\mu}_{y\text{PEL}}$.

8.2 (Pseudo EL Ratio Statistic with Auxiliary Information) Let $r_{\text{WR}}(\theta) = \ell_{\text{WR}}(\hat{\mathbf{p}}(\theta)) - \ell_{\text{WR}}(\hat{\mathbf{p}})$, where the "global" maximizers $\hat{p}_i, i \in \mathbf{S}$ are subject to both the normalization constraint (8.2) and the constraints (8.11) on auxiliary variables, and are given by (8.12); the "restricted" maximizers $\hat{p}_i(\theta), i \in \mathbf{S}$ are obtained under the constraints (8.2), (8.11) and the additional constraint (8.6) induced by the parameter $\theta_0 = \mu_y$. Let

$$K = \sum_{i \in \mathbf{S}} \tilde{d}_i(\mathbf{S}) \log\{\tilde{d}_i(\mathbf{S})\}\,.$$

Note that $\mathbf{u}_i = \mathbf{x}_i - \mu_{\mathbf{x}}$ and we are interested in the asymptotic distribution of $r_{\text{WR}}(\theta)$ at $\theta = \mu_y$.

(a) Show in detail that

$$\ell_{\text{WR}}\big(\hat{\mathbf{p}}\big) = nK - \frac{n}{2}\Big\{\sum_{i \in \mathbf{S}} \tilde{d}_i(\mathbf{S})\mathbf{u}_i\Big\} \left(\frac{1}{N}\sum_{i=1}^{N} \mathbf{u}_i\mathbf{u}_i'\right)^{-1} \Big\{\sum_{i \in \mathbf{S}} \tilde{d}_i(\mathbf{S})\mathbf{u}_i\Big\} + o_p(1)\,.$$

(b) Argue that the constraint (8.6) used for computing the "restricted" maximizer $\hat{p}_i(\theta)$'s can be replaced by

$$\sum_{i \in \mathbf{S}} p_i(y_i - \theta - \mathbf{B}'\mathbf{u}_i) = 0$$

for any constant vector $\mathbf{B}$.

(c) Argue that, for any $\theta = \mu_y + O(n^{-1/2})$, $\ell_{\text{WR}}\big(\hat{\mathbf{p}}(\theta)\big)$ has the same asymptotic expansion as $\ell_{\text{WR}}\big(\hat{\mathbf{p}}\big)$ given in Part (a) except that $\mathbf{u}_i$ needs to be changed to $\mathbf{u}_i^* = (\mathbf{u}_i', r_i)'$, where $r_i = y_i - \theta - \mathbf{B}'\mathbf{u}_i$.

(d) Show that the matrix $\sum_{i=1}^{N}(\mathbf{u}_i^*)(\mathbf{u}_i^*)'$ is block diagonal at $\theta = \mu_y$ if we choose

$$\mathbf{B} = \mathbf{B}_{\text{PEL}} = \Big(\sum_{i=1}^{N} \mathbf{u}_i\mathbf{u}_i'\Big)^{-1} \sum_{i=1}^{N} \mathbf{u}_i(y_i - \mu_y)\,.$$

(e) Show that $-2r_{\text{WR}}(\theta)/\text{deff}_{\text{GREG}}$ converges in distribution to a χ^2 random variable with one degree of freedom, where the design effect is given by

$$\text{deff}_{\text{GREG}} = V_p\big(\tilde{\mu}_{y\text{GREG}}\big)/\big(\sigma_r^2/n\big)\,,$$

with $\tilde{\mu}_{y\text{GREG}} = N^{-1}\sum_{i\in\mathbf{S}} d_i r_i$, $\sigma_r^2 = (N-1)^{-1}\sum_{i=1}^{N} r_i^2$ and $r_i = y_i - \mu_y - \mathbf{B}'_{\text{PEL}}\mathbf{u}_i$, $i = 1, \ldots, N$.

8.3 (Algorithms for Two-Sample Empirical Likelihood) Let $\{(y_{1i}, \mathbf{x}_i), i = 1, \ldots, n_1\}$ be the first sample and $\{(y_{0j}, \mathbf{x}_j), j = 1, \ldots, n_0\}$ be the second sample. The two samples are independent. The response variables y_1 and y_0 can be different but the auxiliary variables $\mathbf{x}$ have common measures such as age and gender. The parameter of interest is $\theta = \mu_1 - \mu_0$, where $\mu_1 = E(y_1)$ and $\mu_0 = E(y_0)$ (Wu and Yan 2012). The standard two-sample empirical log-likelihood function is given by

$$\ell(\mathbf{p}, \mathbf{q}) = \sum_{i=1}^{n_1} \log(p_i) + \sum_{j=1}^{n_0} \log(q_j)\,,$$

(a) Using techniques similar to stratified sampling presented in Sect. 8.4.2, describe an algorithm to maximize $\ell(\mathbf{p}, \mathbf{q})$ subject to the following constraints:

$$\sum_{i=1}^{n_1} p_i = 1\,,\quad \sum_{j=1}^{n_0} q_j = 1\,,$$

$$\sum_{i=1}^{n_1} p_i \mathbf{x}_i = \sum_{j=1}^{n_0} q_j \mathbf{x}_j\,,$$

$$\sum_{i=1}^{n_1} p_i y_{1i} - \sum_{j=1}^{n_0} q_j y_{0j} = \theta$$

for a fixed value of θ.

Hint: Reformulate the problem as a stratified sample with $H = 2$ and $W_1 = W_2 = 1/2$. Put the constraints in the form of (8.19), (8.20), (8.16) and (8.17).

(b) Write an R code to carry out the constrained maximization of $\ell(\mathbf{p}, \mathbf{q})$ with respect to $\mathbf{p}$ and $\mathbf{q}$, assuming that there are two auxiliary variables. Use simulated data to test the code.

8.4 (Generalized Pseudo Empirical Likelihood) The generalized pseudo EL ratio statistic $r(\theta)$ for $\theta = T_y$ is defined in Sect. 8.5 and the related computational details on constrained minimization of $EL(\mathbf{w}, \mathbf{d})$ are given in Sect. 6.2.2.

(a) Give a detailed proof of Theorem 8.2 using techniques similar to those of Problem 8.2 on the pseudo EL ratio statistic for $\theta = \mu_y$.
(b) Provide a design consistent estimator $\hat{C}$ for the scaling constant C.

8.5 (Proof of Part (b) of Theorem 8.3)

(a) Show that $\ell(\theta, \lambda) = \sum_{i \in \mathbf{S}} \tilde{d}_i(\mathbf{S}) \log \left\{1 + \lambda' \mathbf{g}(y_i, \mathbf{x}_i; \theta)\right\}$ can be approximated by $L(\theta, \lambda) = (\theta - \theta_{\mathrm{N}})' \mathbf{H}' \lambda + \left\{\mathbf{G}_n(\theta_{\mathrm{N}})\right\}' \lambda - (1/2) \lambda' \mathbf{W}_1 \lambda$ when (θ, λ) is in the neighborhood of $(\theta_{\mathrm{N}}, \mathbf{0})$.
(b) Show that $\partial L(\theta, \lambda)/\partial \theta = \mathbf{0}$ and $\partial L(\theta, \lambda)/\partial \lambda = \mathbf{0}$ lead to the equation system (8.31).
(c) Show that (8.31) leads to (8.32).

Chapter 9
Methods for Handling Missing Data

In an ideal world for surveys, the finite population is well defined, sampling frames are complete and up-to-date, units for the sample are selected based on the survey design, and variables of interest are measured without any error for all units selected in the sample. In practice, however, there might be issues associated with some or all of the steps in survey operations, resulting in problems such as under-coverage, nonresponse and measurement error. If any of the problems is not handled with care, naive statistical analyses using the observed survey data often lead to invalid results. While missing data itself is a general topic arising from many fields, this chapter discusses issues related to and methods for handling missing survey data.

9.1 Issues with Missing Survey Data

We make a distinction between two types of variables. Let $\mathbf{x}$ be a vector of baseline variables describing the basic characteristics of the units, such as gender, age, region, education, etc. Let $\mathbf{y}$ be a vector of study variables which are the main focus of the data collection and subsequent scientific investigations. Let $\mathscr{F}_N = \{(\mathbf{x}_i, \mathbf{y}_i), i = 1, 2, \cdots, N\}$ represent the data structure for the entire finite population, with $i = 1, 2, \cdots, N$ representing the N individual units of the finite population.

There are two types of nonresponse for survey data: *unit nonresponse* and *item nonresponse*. Unit nonresponse refers to cases where the unit is selected by the sampling procedure but nothing is measured, either because of inaccessibility of the unit during the data collection period or because of a refusal of the unit to be part of the survey. For multi-stage cluster sampling, unit nonresponse can occur at cluster levels. See Chap. 12 for an example from the International Tobacco Control (ITC) China Survey. Item nonresponse refers to missing values for a particular variable, due to either the failure of the data collection process or the refusal of the

C. Wu, M. E. Thompson, *Sampling Theory and Practice*, ICSA Book Series in Statistics, https://doi.org/10.1007/978-3-030-44246-0_9

Table 9.1 An example of survey datasets with item nonresponse on two study variables

i	y_{i1}	y_{i2}	x_{i1}	x_{i2}	x_{i3}	w_i
1	y_{11}	y_{12}	x_{11}	x_{12}	x_{13}	w_1
2	$*$	y_{22}	x_{21}	x_{22}	x_{23}	w_2
3	y_{31}	$*$	x_{31}	x_{32}	x_{33}	w_3
4	y_{41}	y_{42}	x_{41}	x_{42}	x_{43}	w_4
5	$*$	$*$	x_{51}	x_{52}	x_{53}	w_5
6	y_{61}	y_{62}	x_{61}	x_{62}	x_{63}	w_6
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	y_{n1}	y_{n2}	x_{n1}	x_{n2}	x_{n3}	w_n

respondent. For item nonresponse, we focus mainly on a univariate study variable y in this chapter. A bivariate $\mathbf{y}$ is used in Table 9.1 to demonstrate possible data file structures.

9.1.1 Missing Data Mechanisms

A common setting for item nonresponse is that the baseline variables $\mathbf{x}$ are observed for all sampled units while some or all of the study variables $\mathbf{y}$ are subject to missingness. The variables $\mathbf{x}$ may also include auxiliary variables or survey design variables, as long as they are available for all the units in the final sample. Table 9.1 depicts possible patterns of missing values for $\mathbf{y} = (y_1, y_2)'$ while $\mathbf{x} = (x_1, x_2, x_3)'$ are observed for all the units in the final sample with size n. The symbol "$*$" indicates a missing value. The last column represents the survey weights.

Let y be a univariate study variable with missing observations. Let $\delta_i = 1$ if y_i is observed and $\delta_i = 0$ otherwise. One of the most crucial concepts in dealing with missing data is the so-called *missing mechanism*. This can be defined through the response probability $\tau_i = P(\delta_i = 1 \mid y_i, \mathbf{x}_i)$. There are three possible scenarios for the missing mechanism of y as defined by Rubin (1976) and Little and Rubin (2002):

(i) missing completely at random (MCAR) if $\tau_i = \tau$ is a constant for all i;
(ii) missing at random (MAR) if $\tau_i = \tau(\mathbf{x}_i)$ depends only on $\mathbf{x}_i$;
(iii) missing not at random (MNAR) if $\tau_i = \tau(y_i, \mathbf{x}_i)$ depends on both y_i and $\mathbf{x}_i$.

MCAR means the study variable is missing purely by chance. MAR implies that the reason behind the missingness of y is characterized by the baseline variables such as gender and age but is unrelated to the value of the study variable. MNAR indicates that both the baseline variables $\mathbf{x}$ and the study variable y itself affect the outcome whether y is observed or not. MCAR is a special case of MAR, and MAR can be viewed as a special case of MNAR.

Under the missing at random assumption where $\tau_i = P(\delta_i = 1 \mid y_i, \mathbf{x}_i) = P(\delta_i = 1 \mid \mathbf{x}_i)$, the missing mechanism is called *ignorable*. If the missing mechanism also satisfies $\tau_i = P(\delta_i = 1 \mid \mathbf{x}_i) > 0$ for all i, the scenario is referred to as the *strong ignorability* by Rosenbaum and Rubin (1983) in the context of causal inference. Strong ignorability also implies that each unit in the finite population has a non-zero probability to be observed in the survey sample, which is an important condition for unbiased estimation of finite population parameters. As pointed out by Rivers (2007), the term "ignorable" is an unfortunate choice of terminology for the missing data and causal inference literature, since it certainly cannot be ignored by the data analyst. Similarly, the term "missing at random" should not be confused with "randomly missing". The missing mechanism plays a crucial role in how missing data are handled in statistical analyses.

9.1.2 Frameworks for Statistical Inference

The indicator variable δ_i for defining the missing mechanisms is a random variable, which is conceptually different from the indicator variable $A_i = I(i \in \mathbf{S})$. The latter is defined by the probability sampling design. The response probability $\tau_i = P(\delta_i = 1 \mid y_i, \mathbf{x}_i)$ is also called *the propensity score* in the literature on causal inference (Rosenbaum and Rubin 1983). The probability model associated with δ_i is called *the propensity score model*. It is often assumed that the missing mechanism and the probability sampling design are not confounded, i.e., the two indicator variables δ_i and A_i are independent given $\mathbf{x}_i$.

When the missing values of y are handled by imputation (Sect. 9.4), an imputation model for y given $\mathbf{x}$ is required. This is the same concept as the superpopulation model discussed in previous chapters. In the general literature on missing data and causal inference, it is also called *the outcome regression model*, which specifies the conditional distribution or the conditional moments of y given $\mathbf{x}$.

There are three possible sources of randomization for dealing with missing survey data: (i) the probability sampling design, p, for selecting the initial sample from the finite population; (ii) the propensity score model, denoted as q, for the missing data mechanism; and (iii) the outcome regression model, denoted as ξ, for the imputation. Statistical inferences with missing survey data require a joint randomization framework involving the probability sampling design p and at least one of the two models q and ξ. Theoretical developments on consistency of point estimators and their variance or asymptotic variance need to be carried out under the required framework. In particular, the variance or asymptotic variance can often be decomposed into two or three variance components depending on how the point estimator is constructed.

9.1.3 Validity and Efficiency

Let $\mathbf{S}$ be the set of sampled units. Let $\mathbf{S}_R = \{i \mid i \in \mathbf{S} \text{ and } \delta_i = 1\}$ be the set of respondents and $\mathbf{S}_M = \{i \mid i \in \mathbf{S} \text{ and } \delta_i = 0\}$ be the set of (missing) nonrespondents. We have $\mathbf{S} = \mathbf{S}_R \cup \mathbf{S}_M$ and $n = n_R + n_M$, where n_R and n_M are respectively the number of respondents and nonrespondents in the initial sample $\mathbf{S}$ and n is the initial sample size. The observed sample data consist of $\{(y_i, \mathbf{x}_i), i \in \mathbf{S}_R\}$ and $\{\mathbf{x}_i, i \in \mathbf{S}_M\}$. Sometimes the two parts of the data, together with the survey weights $\{w_i, i \in \mathbf{S}\}$, can be combined as $\{(\delta_i y_i, \mathbf{x}_i, \delta_i, w_i), i \in \mathbf{S}\}$.

There are two critical aspects of statistical analysis with missing data: *validity* and *efficiency*. Validity refers to the consistency of the point estimators for the parameters of interest, and efficiency is measured by the relative variances or asymptotic variances among alternative consistent estimators. Validity and efficiency depend on the analysis problems (i.e., the unknown parameters) and also on how missing data are handled. The so-called complete-case analysis (CCA), which deletes units with missing y and only uses the data from $\{(y_i, \mathbf{x}_i), i \in \mathbf{S}_R\}$, typically leads to invalid results unless y is missing completely at random. See Problem 9.1 for a simple example. Achieving valid statistical inference is the primary goal in analyzing data with missing values, and finding a more efficient method among possible alternatives is an added bonus.

9.2 Methods for Handling Unit Nonresponse

Let $\mathbf{S}_0$ be the initial intended sample of size n_0 with units selected by a probability sampling design. Let $\pi_i^{(0)} = P(i \in \mathbf{S}_0)$ be the initial inclusion probabilities and $w_i^{(0)} = 1/\pi_i^{(0)}$ be the initial survey weights. However, there are only n $(< n_0)$ units with full or partial measurements and there are $n_0 - n$ units that fail to respond completely. In survey practice, unit nonrespondents are deleted from the sample, and the survey weights for the remaining n units in the final sample $\mathbf{S} \subset \mathbf{S}_0$ are adjusted to compensate for the reduced sample size.

The inclusion probabilities for the final sample $\mathbf{S}$ involve the missing mechanism for the unit nonresponse and cannot be determined by the initial survey design. Computing them involves both the propensity score model and the initial survey design, and the resulting survey weights are sometimes called the quasi-randomization weights (Kott 1994). The quasi-randomization inclusion probability can be written as

$$\pi_i = P(i \in \mathbf{S}) = P(i \in \mathbf{S}, i \in \mathbf{S}_0) = P(i \in \mathbf{S}_0)P(i \in \mathbf{S} \mid i \in \mathbf{S}_0) = \pi_i^{(0)}\tau_i\,,$$

where τ_i is the propensity score for unit i and is assumed to be independent of the sample selection process. If the missing mechanism for the unit nonresponse is MCAR, i.e., $\tau_i = \tau$ is a constant, then τ can be estimated by N_1/N, where N_1 is the number of potential unit respondents in the population. A design-based estimator of τ is given by

$$\hat{\tau} = \frac{\hat{N}_1}{\hat{N}} = \left(\sum_{i \in \mathbf{S}} w_i^{(0)} \right) \Big/ \left(\sum_{i \in \mathbf{S}_0} w_i^{(0)} \right) .$$

The final adjusted survey weights are given by $w_i = 1/\hat{\pi}_i = w_i^{(0)} \hat{\tau}^{-1}$, $i \in \mathbf{S}$. This is called *the ratio adjustment* to survey weights for unit nonresponse. It can be seen that $w_i \propto w_i^{(0)}$ and $\sum_{i \in \mathbf{S}} w_i = \sum_{i \in \mathbf{S}_0} w_i^{(0)}$.

Although the main survey variables $\mathbf{x}$ and $\mathbf{y}$ are not available for unit nonrespondents, sometimes certain information is available at the survey design stage for all the units in the initial sample $\mathbf{S}_0$, such as the size variable of the unit or variables for defining stratification of the finite population. Under such scenarios, it is possible to consider the more general missing mechanism MAR for unit nonresponses. Let $\mathbf{z}_i$ be the measurements available for all $i \in \mathbf{S}_0$. Let $\tau_i = \tau(\mathbf{z}_i)$ be the propensity score which depends on $\mathbf{z}_i$. Let $\hat{\tau}_i$ be a suitable estimate for τ_i. Then the final adjusted survey weights can be computed similarly to the MCAR case and are given by $w_i = w_i^{(0)} \hat{\tau}_i^{-1}$, $i \in \mathbf{S}$.

The most popular method for estimating $\tau = \tau(\mathbf{z})$ is to assume a logistic regression model for the response indicator variable δ. We have $\delta_i = 1$ if $i \in \mathbf{S}$ and $\delta_i = 0$ otherwise. The parametric form of τ_i is given by

$$\tau_i = 1 - \{1 + \exp(\mathbf{z}_i' \theta)\}^{-1} ,$$

where θ is the vector of unknown model parameters. The set of variables $\mathbf{z}_i$ is assumed to have 1 as its first component so that the model contains an intercept. The likelihood function of θ is given by

$$L(\theta) = \prod_{i \in \mathbf{S}_0} \tau_i^{\delta_i} (1 - \tau_i)^{1 - \delta_i} ,$$

and the maximum likelihood estimator $\hat{\theta}$ of θ can be obtained by using a suitable iterative procedure (Problem 9.2).

A special case for the MAR scenario is that the variables $\mathbf{z}$ consist of one or several categorical variables. For instance, we may have $\mathbf{z} = (z_1, z_2)'$ where z_1 is the size variable having three categories (small, medium and large) and z_2 represents geographical areas (ten provinces). In this case, the $\mathbf{z}$ variables define a total of $3 \times 10 = 30$ subpopulations (post-strata), and within each post-stratum the propensity score values $\tau_i = \tau(\mathbf{z}_i)$ are constant. Weight adjustment for unit nonresponse under such scenarios can be done by first dividing the initial sample $\mathbf{S}_0$ into 30 post-stratified subsamples based on $\mathbf{z}$, called *the weighting classes*, and within each weighting class the ratio adjustment described earlier can be applied.

There are other weight adjustment methods such as calibration weighting. Kim and Kim (2007), Haziza and Lesage (2016) and Haziza and Beaumont (2017) contain further discussions and reviews of weight adjustment for unit nonresponse.

9.3 Methods for Handling Item Nonresponse

Let y be the univariate response variable which is subject to missingness. Let $\{(\delta_i y_i, \mathbf{x}_i, \delta_i, w_i), i \in \mathbf{S}\}$ be the observed survey dataset. Without complicating the discussions, we assume that the survey weights w_i provide valid design-based inference and any potential impact of weight adjustment on unit nonresponse can be ignored. Let $\mathbf{S} = \mathbf{S}_R \cup \mathbf{S}_M$ and $n = n_R + n_M$ be defined based on respondents and nonrespondents for the study variable y.

9.3.1 Complete-Case Analysis

The complete-case analysis (CCA) approach to missing survey data deletes all units with missing y and uses only the data $\{(y_i, \mathbf{x}_i, w_i), i \in \mathbf{S}_R\}$ from the set of respondents. The auxiliary information $\{(\mathbf{x}_i, w_i), i \in \mathbf{S}_M\}$ from the missing set of nonrespondents is not used.

The CCA approach often leads to inconsistent results unless the missing mechanism is MCAR. In the absence of any missing values, the estimator of the finite population mean μ_y is given by $\hat{\mu}_y = \big(\sum_{i \in \mathbf{S}} w_i y_i\big)/\big(\sum_{i \in \mathbf{S}} w_i\big)$, which satisfies $E(\hat{\mu}_y) \doteq \sum_{i=1}^N y_i/N$ under the survey design. With missing values on y, the CCA estimator of μ_y is given by

$$\hat{\mu}_{y\text{CCA}} = \left(\sum_{i \in \mathbf{S}_R} w_i y_i\right) \Big/ \left(\sum_{i \in \mathbf{S}_R} w_i\right) .$$

The theoretical properties of the estimator $\hat{\mu}_{y\text{CCA}}$ need to be assessed under the joint randomization of the survey design, p, and the propensity score model, q. Noting that $E_q(\delta_i) = \tau_i$, the numerator of $\hat{\mu}_{y\text{CCA}}$ satisfies

$$E\left(\sum_{i \in \mathbf{S}_R} w_i y_i\right) = E_p\left\{E_q\left(\sum_{i \in \mathbf{S}} \delta_i w_i y_i\right)\right\} = E_p\left(\sum_{i \in \mathbf{S}} \tau_i w_i y_i\right) \doteq \sum_{i=1}^N \tau_i y_i .$$

Similarly, $E\big(\sum_{i \in \mathbf{S}_R} w_i\big) \doteq \sum_{i=1}^N \tau_i$. It is apparent that $\hat{\mu}_{y\text{CCA}}$ is not an approximately unbiased estimator of μ_y under the joint (p, q) framework unless $\tau_i = \tau$ is a constant for all i. This statement holds for many other finite population parameters.

The CCA methods are sometimes used in practice due to its ease of implementation with existing computational software. The results may be acceptable when the rate of nonresponse is low and the amount of perturbation from the missing values is negligible. When the nonresponse rate is high, the method often leads to invalid results. There is one important problem, however, for which the CCA approach does provide valid results under the MAR assumption, as described below.

Suppose that the linear regression model with $E_\xi(y \mid \mathbf{x}) = \mathbf{x}'\beta$ and $V_\xi(y \mid \mathbf{x}) = \sigma^2$ holds for the finite population, and the inference goal is to obtain a consistent estimator of the regression coefficients β. Suppose also that y is missing at random (MAR) such that y_i and δ_i are independent given $\mathbf{x}_i$. In addition, the survey design is assumed to be noninformative in the sense that y_i and the sample inclusion indicator A_i are independent given $\mathbf{x}_i$. Under such conditions we have

$$E_\xi(y_i \mid \delta_i, A_i, \mathbf{x}_i) = E_\xi(y_i \mid \mathbf{x}_i) = \mathbf{x}_i'\beta \,.$$

The ordinary least squares estimator of β using the observed sample data $\{(y_i, \mathbf{x}_i), i \in \mathbf{S}_R\}$ (corresponding to units with $\delta_i = 1$) without using the survey weights is consistent. When the noninformativeness of the survey design is questionable, one should use a survey weighted least squares estimator based on the observed data. Section 9.5 contains further discussions on related topics.

9.3.2 Propensity Score Adjusted Analysis

The propensity score adjusted (PSA) method was first proposed in the context of causal inference by Rosenbaum and Rubin (1983). The basic idea comes from the well-known Horvitz-Thompson estimator (Horvitz and Thompson 1952) used in survey sampling, for which each observed unit is inversely weighted by the inclusion probability. The PSA approach is also popularly termed as the Inverse Probability Weighting (IPW) method in the missing data and causal inference literature.

Suppose that y is missing at random and the propensity scores are given by $\tau_i = P(\delta_i = 1 \mid \mathbf{x}_i)$. The propensity score adjusted estimator of the finite population mean μ_y is defined as

$$\hat{\mu}_{y\text{PSA}} = \frac{1}{N} \sum_{i \in \mathbf{S}_R} \frac{w_i y_i}{\tau_i} \,,$$

where N is assumed to be known. The expectation of $\hat{\mu}_{y\text{PSA}}$ under the (q, p)-randomization is given by

$$E\left(\hat{\mu}_{y\text{PSA}}\right) = E_p E_q \left(\frac{1}{N} \sum_{i \in \mathbf{S}} \delta_i \frac{w_i y_i}{\tau_i} \right) = E_p \left(\frac{1}{N} \sum_{i \in \mathbf{S}} w_i y_i \right) \doteq \mu_y \,.$$

The approximate equal sign at the last step is due to potential unit nonresponse adjustment and calibration weighting for the final survey weights w_i. In practice, the population size N is typically unknown and a Hájek-type estimator is used, which is given by

$$\hat{\mu}_{y\text{PSA}} = \left(\sum_{i \in \mathbf{S}_R} \frac{w_i y_i}{\tau_i} \right) \Big/ \left(\sum_{i \in \mathbf{S}_R} \frac{w_i}{\tau_i} \right) . \tag{9.1}$$

Similar to the discussions presented in Chap. 4, the Hájek-type estimator is preferred in practice even if the population size N is known. The exact asymptotic properties of $\hat{\mu}_{y\text{PSA}}$ given by (9.1) with the fixed set of τ_i can be derived as a ratio-type estimator (Problem 9.3).

The propensity scores τ_i are unknown in practice but can be estimated under MAR with an assumed logistic regression model or other suitable models for the binary responses δ_i based on the dataset $\{(\delta_i, \mathbf{x}_i), i \in \mathbf{S}\}$. The consistency of the PSA estimator with the estimated τ_i remains unchanged but the variance will be affected by the estimation process. See Sect. 9.6 for further details.

9.3.3 Analysis with Imputation for Missing Values

Missing values of y may be replaced by imputed ones to create an artificial "complete dataset" without missing cells. The imputed value for a missing y_i, denoted as y_i^*, represents a guess of what is missing based on what is known. In other words, *imputation is prediction*, and prediction requires a model. Under the context of imputation for missing values, it is called *the imputation model*. The most sensible imputation model for the study variable y given $\mathbf{x}$ is the outcome regression model, ξ, which is a superpopulation model for the finite population. The imputation model can be fully parametric or nonparametric, or something in between.

Consider the following semiparametric imputation model (ξ),

$$y_i = m(\mathbf{x}_i; \beta) + \varepsilon_i \,, \quad i = 1, 2, \cdots, N \,, \tag{9.2}$$

where β is the vector of unknown model parameters but the function $m(\cdot;\cdot)$ has a known form, and the error terms $\varepsilon_1, \cdots, \varepsilon_N$ are assumed to be independent with mean zero and variance $v(\mathbf{x}_i)\sigma^2$. It follows that $E_\xi(y_i \mid \mathbf{x}_i) = m(\mathbf{x}_i; \beta)$ and $V_\xi(y_i \mid \mathbf{x}_i) = v(\mathbf{x}_i)\sigma^2$. The variance function $v(\cdot)$ is assumed to have a known form as well. To simplify discussions, we assume the true values of β, denoted as β_0, are known for the rest of Sect. 9.3. In practice, the regression coefficients β need to be replaced by a suitable estimator, $\hat{\beta}$, as discussed in Sects. 9.4–9.6.

Let $y_i^* = m(\mathbf{x}_i; \beta_0)$ be the imputed value for $y_i, i \in \mathbf{S}_M$. Let $\{(\tilde{y}_i, \mathbf{x}_i, w_i), i \in \mathbf{S}\}$ be the imputed dataset, where $\tilde{y}_i = \delta_i y_i + (1 - \delta_i) y_i^*$. It can be seen that $\tilde{y}_i = y_i$ if y_i is observed and $\tilde{y}_i = y_i^*$ if y_i is missing. The imputation-based estimator of μ_y is given by

$$\hat{\mu}_{y\text{IMP}} = \frac{1}{N} \sum_{i \in \mathbf{S}} w_i \tilde{y}_i = \frac{1}{N} \left(\sum_{i \in \mathbf{S}_R} w_i y_i + \sum_{i \in \mathbf{S}_M} w_i y_i^* \right), \tag{9.3}$$

where N is assumed to be known. Theoretical properties of $\hat{\mu}_{y\text{IMP}}$ need to be evaluated under the joint randomization of the probability sampling design, p, and the imputation model ξ specified by (9.2). Under the model ξ, the target parameter, namely the finite population mean μ_y, needs to be viewed as a random quantity. Approximate unbiasedness of $\hat{\mu}_{y\text{IMP}}$ requires that $E\big(\hat{\mu}_{y\text{IMP}} - \mu_y\big) \doteq 0$ under the joint (p, ξ) framework, and variance estimation is for $V\big(\hat{\mu}_{y\text{IMP}} - \mu_y\big)$. See Problem 9.4 for further details. Practical applications of the estimator $\hat{\mu}_{y\text{IMP}}$ given by (9.3) under model (9.2) require two modifications: replace β_0 by a sample-based estimate and substitute N by $\hat{N} = \sum_{i \in \mathbf{S}} w_i$. The consistency of the estimator under the (p, ξ) framework remains unchanged but both sources of randomness affect the asymptotic variance. Some of the commonly used imputation methods are described in Sect. 9.4.

One of the motivations behind the imputation approach is a more efficient use of auxiliary information that is available from the observed dataset. The CCA approach deletes the observed data $\{(\mathbf{x}_i, w_i), i \in \mathbf{S}_M\}$ for which the y_i are missing. The PSA approach uses $\{(\mathbf{x}_i, w_i), i \in \mathbf{S}_M\}$ indirectly through the estimation of the propensity scores τ_i for $i \in \mathbf{S}_R$. The imputation-based approach provides a direct way of using $\mathbf{x}_i$ to predict y_i when it is missing, which often leads to more efficient estimation of the parameters of interest with the imputed datasets.

9.3.4 Doubly Robust Estimation

The propensity score adjusted estimator $\hat{\mu}_{y\text{PSA}}$ requires that the propensity score model q be correctly specified. The imputation-based estimator $\hat{\mu}_{y\text{IMP}}$ depends on the correct specification of the imputation model ξ. The estimator $\hat{\mu}_{y\text{PSA}}$ or $\hat{\mu}_{y\text{IMP}}$ becomes inconsistent when the respective model is misspecified. An estimator is called *doubly robust* if the estimator is consistent whenever one of the two models is correctly specified. Double robustness has been a popular concept in dealing with missing data problems or causal inference (Robins et al. 1994; Scharfstein et al. 1999).

Let τ_i be the propensity scores which are specified by the model q, $i \in \mathbf{S}_R$; let $m(\mathbf{x}_i, \beta_0)$ be given by the outcome regression model ξ, $i \in \mathbf{S}$. Consider the following estimator of μ_y,

$$\hat{\mu}_{y\mathrm{DR}} = \frac{1}{N}\left\{\sum_{i\in\mathbf{S}_R}\frac{w_i y_i}{\tau_i} - \sum_{i\in\mathbf{S}_R}\frac{w_i\, m(\mathbf{x}_i,\beta_0)}{\tau_i} + \sum_{i\in\mathbf{S}} w_i\; m(\mathbf{x}_i,\beta_0)\right\}. \tag{9.4}$$

This estimator is doubly robust under the MAR assumption in the sense that it is consistent under the correctly specified propensity score model q regardless of the model ξ and it is also consistent under the correctly specified outcome regression model ξ irrespective of the τ_i (Problem 9.5). Note that the sampling design p is always part of the theoretical framework.

When the τ_i and the β_0 used in $\hat{\mu}_{y\mathrm{DR}}$ given by (9.4) are replaced by sample-based estimates, the double robustness of the estimator remains unchanged, i.e., the estimator is still consistent for μ_y when one of the q and ξ models is correctly specified. However, variance estimation for doubly robust estimators turns out to be a very challenging problem. The asymptotic variance formula, which is the base for constructing a variance estimator, needs to be derived under the given models. The doubly robust estimator only requires that one of the two models q and ξ be correct, but does not specify which one. Kim and Haziza (2014) proposed a variance estimation strategy to first construct a variance estimator under the propensity score model and then adjust it for potential bias under the outcome regression model, so that the final variance estimator is approximately unbiased under either the (p, q) or the (p, ξ) framework. The method of Kim and Haziza (2014) also requires the estimation of the model parameters for the propensity scores and the outcome regression to be done through a special set of unbiased estimating equations. One useful result in deriving the asymptotic variance formula is that the estimation of the parameters β has no impact on the asymptotic variance of $\hat{\mu}_{y\mathrm{DR}}$ under the propensity score model. Chen et al. (2020) contains useful technical details and related discussions on variance estimation for doubly robust estimators.

9.4 Imputation for Missing Values

Imputation for missing values is a prediction process, and the choice of the imputation model depends on the availability of auxiliary information. Suppose that y_i is missing and the y_i^* is an imputed value by a chosen imputation method. The *first principle* is that the imputation method needs to be unbiased. *Unbiased imputation* is defined as

$$E_\xi\big(y_i^* - y_i\big) = 0\,, \quad i \in \mathbf{S}_M\,, \tag{9.5}$$

where ξ denotes the imputation model. Most commonly used imputation methods use a conditional model for y given $\mathbf{x}$. A stronger version of unbiased imputation is defined as

$$E_\xi\big(y_i^* - y_i \mid \mathbf{x}_i\big) = 0\,, \quad i \in \mathbf{S}_M\,. \tag{9.6}$$

It is apparent that (9.6) implies (9.5). The imputation model used in practice typically involves unknown model parameters, and unbiased imputation defined through (9.5) or (9.6) becomes an approximate process under the fitted imputation model.

The *second principle* for imputation is that the imputed dataset preserves *the distributional structure* of the full dataset without missing values. This may refer to the marginal distribution of a particular study variable y or the joint distribution of multivariate study variables. The discussions in this section focus primarily on the first principle with a single study variable y.

9.4.1 Single Imputation

Single imputation replaces each missing y_i by one imputed value y_i^* to fill the missing cell, and creates a single "complete dataset". The imputation procedure could be random or deterministic. Under *random imputation*, the imputed values for the same missing y_i may be different if the imputation procedure is repeated under the same process. *Deterministic imputation* leads to a unique imputed value for each missing y_i even if the imputation procedure is carried out by different data producers.

Another practically useful concept is the so-called *hot deck* imputation procedure, which imputes the missing y_i for $i \in \mathbf{S}_M$ by $y_i^* = y_j$ for some $j \in \mathbf{S}_R$. The units in $\mathbf{S}_R$ are called the *donors* and the units in $\mathbf{S}_M$ are called the *recipients*. Hot deck imputation uses real observed values which are always within the feasible range of the missing value. For instance, if the study variable y is binary taking two possible values 0 and 1, a hot deck imputation procedure would impute a missing y_i by 0 or 1. This is not always the case with certain other imputation methods.

We now discuss some commonly used single imputation methods. As indicated earlier, the choice of the imputation method is often dictated by the available auxiliary information. We consider three practical scenarios depending on the format and availability of the auxiliary information.

Scenario I: No Auxiliary Information Available from the Dataset The observed dataset is $\{(y_i, w_i),\ i \in \mathbf{S}_R\}$ plus $\{w_i, i \in \mathbf{S}_M\}$. The only sensible model for imputation is the common mean model ξ,

$$y_i = \mu + \varepsilon_i\,, \quad i = 1, 2, \cdots, N\,,$$

where μ is the mean of the superpopulation. There are two popular imputation methods for this simple scenario. The first is *mean imputation*. For each missing

y_i, we impute y_i by $y_i^* = \bar{y}_R = \big(\sum_{i\in\mathbf{S}_R} w_i y_i\big)/\big(\sum_{i\in\mathbf{S}_R} w_i\big)$, the survey weighted mean of the observed sample. The second method is *random hot deck imputation*. For each missing y_i, we impute y_i by $y_i^* = y_j$, which is the observed value of y from donor j, where j is randomly selected from the donor set $\mathbf{S}_R$. This process is carried out independently for each $i \in \mathbf{S}_M$. There is also a modified procedure to select the donor j with unequal probability proportional to the survey weights w_i, which is termed the *weighted random hot deck imputation*.

The mean imputation and random hot deck imputation methods are both unbiased under the common mean model ξ. The mean imputation method produces an imputed dataset with a "spike" at the imputed value $\bar{y}_R$, and hence does not preserve the distributional structure of the y variable. It is seldom used in practice. The random hot deck imputation method is often used in practice in conjunction with the imputation classes discussed below.

Scenario II: Auxiliary Information with Categorical Auxiliary Variables The observed sample dataset is given by $\{(y_i, \mathbf{x}_i, w_i), i \in \mathbf{S}_R\}$ plus $\{(\mathbf{x}_i, w_i), i \in \mathbf{S}_M\}$, where the vector of auxiliary variables $\mathbf{x}$ consists of several categorical variables such as gender and age groups. Suppose that we have an imputation model ξ such that $E_\xi(y_i \mid \mathbf{x}_i) = m(\mathbf{x}_i), i = 1, 2, \cdots, N$, where $m(\cdot)$ is an unspecified function. If we partition the finite population into K subpopulations based on the classifications formed by $\mathbf{x}$, then $y_i = m(\mathbf{x}_i) + \varepsilon_i$ becomes a common mean model within each subpopulation.

Let $\mathbf{S} = \mathbf{S}_1 \cup \cdots \cup \mathbf{S}_K$ be the partition formed by the categorical $\mathbf{x}$ variables, where each subsample $\mathbf{S}_k$ is called an *imputation class*. Each imputation class $\mathbf{S}_k$ can further be divided as $\mathbf{S}_k = \mathbf{S}_{Rk} \cup \mathbf{S}_{Mk}$, where $\mathbf{S}_{Rk}$ is the set of respondents and $\mathbf{S}_{Mk}$ is the set of units with missing y for the kth imputation class. The random hot deck imputation method can be applied within each imputation class. The imputation procedure is unbiased under the model with $E_\xi(y_i \mid \mathbf{x}_i) = m(\mathbf{x}_i)$.

Random hot deck imputation with imputation classes is one of the most popular procedures in practice. It can be extended to the general Scenario III (to be described below) with a large number of mixed-types auxiliary variables. The most crucial step is to form imputation classes that resemble the scenario with categorical auxiliary variables and the common mean model. This can be done by first fitting a model on y given $\mathbf{x}$ and obtaining fitted values $\hat{y}_i = \hat{m}(\mathbf{x}_i)$ for all $i \in \mathbf{S}$, and then ordering all the units in the sample based on the magnitude of $\hat{y}_i$, i.e., $\hat{y}_{(1)} \le \hat{y}_{(2)} \le \cdots \hat{y}_{(n)}$. The imputation classes are formed by grouping units based on the ordered sequence. The mean values $m(\mathbf{x}_i)$ for units in the same imputation class have relatively small variations, and the random hot deck imputation method provides an approximately unbiased imputation procedure. Another popular method for constructing imputation classes under the general scenario is to use the propensity scores τ_i. The imputation classes are formed based on the ordered sequence of the estimated propensity scores. Within each imputation class, the propensity scores are close to being uniform, and the missing mechanism can be viewed as MCAR. It therefore justifies the use of random hot deck imputation within each imputation class. The paper by Haziza and Beaumont (2007) contains discussions and relevant references on construction of imputation classes and other related issues such as the choice of the number of imputation classes to be formed.

Scenario III: Auxiliary Information with Mixed-Type Auxiliary Variables The observed sample dataset is still given by $\{(y_i, \mathbf{x}_i, w_i), i \in \mathbf{S}_R\}$ plus $\{(\mathbf{x}_i, w_i), i \in \mathbf{S}_M\}$, where the vector $\mathbf{x}$ consists of auxiliary variables with different types, such as binary, categorical, ordinal or continuous variables. Imputation procedures under this general scenario are developed based on different modeling strategies.

Parametric imputation requires a fully specified parametric distribution $f(y \mid \mathbf{x}, \gamma)$ on y given $\mathbf{x}$, where the γ is the vector of model parameters. The missing y_i with the observed $\mathbf{x}_i$ can be imputed by y_i^*, which is randomly generated from the distribution $f(y \mid \mathbf{x}, \gamma)$ at $\mathbf{x} = \mathbf{x}_i$, with the model parameters γ replaced by suitable estimates. The imputation procedure is approximately unbiased when y is missing at random, which allows consistent estimation of the model parameters γ using the observed data $\{(y_i, \mathbf{x}_i, w_i), i \in \mathbf{S}_R\}$.

Regression imputation is a popular procedure based on a semiparametric model which specifies that $E_\xi(y_i \mid \mathbf{x}_i) = \mathbf{x}_i'\beta$ and $V_\xi(y_i \mid \mathbf{x}_i) = v(\mathbf{x}_i)\sigma^2$. Let $\hat{\beta}$ be the ordinary or the survey weighted least squares estimator discussed in Sect. 7.3.1 using the observed dataset $\{(y_i, \mathbf{x}_i, w_i), i \in \mathbf{S}_R\}$. Under the *deterministic regression imputation* procedure, the missing y_i is imputed as $y_i^* = \mathbf{x}_i'\hat{\beta}$. It can be seen that $E_\xi(y_i - y_i^* \mid \mathbf{x}_i) = 0$ holds exactly under the regression model. The distributional structure of the study variable y from the imputed dataset can be better preserved through *random regression imputation*, which imputes the missing y_i by $y_i^* = \mathbf{x}_i'\hat{\beta} + \hat{\varepsilon}_j$, where $\hat{\varepsilon}_j = y_j - \mathbf{x}_j'\hat{\beta}$ is the fitted residual for unit j, which is randomly selected from the donor set $\mathbf{S}_R$ for which the fitted residuals can be computed. Random regression imputation is also exactly unbiased under the assumed regression model.

Nearest neighbor imputation (NNI) is a procedure that has been widely used in practice. The method is nonparametric and requires no explicit specification of the imputation model. Noting that the values of $\mathbf{x}$ are measured for all units, a missing y_i, $i \in \mathbf{S}_M$ may be replaced by y_j for some $j \in \mathbf{S}_R$ such that the donor j has the closest similarity to the recipient i. More formally, the missing y_i is imputed as $y_i^* = y_j$, with the donor j satisfying

$$\|\mathbf{x}_j - \mathbf{x}_i\| = \min_{k \in \mathbf{S}_R} \|\mathbf{x}_k - \mathbf{x}_i\| ,$$

where $\| \cdot \|$ denotes the L_2 norm. When there are multiple candidate donors all satisfying the same minimum distance, the NNI procedure randomly chooses one from the pool of candidates.

The NNI procedure is equivalent to random hot deck imputation with imputation classes when $\mathbf{x}$ consists of several categorical variables, since all the donors have zero distance to all the recipients within the same imputation class. In general, the NNI procedure is approximately unbiased under the nonparametric model $y_i = m(\mathbf{x}_i) + \varepsilon_i$ and some mild regularity conditions. Further theoretical properties can be found in Chen and Shao (2000, 2001) and Shao (2009).

9.4.2 Multiple Imputation

Multiple imputation was proposed by Rubin (1978, 1987) as a general approach to dealing with item nonresponse in surveys. The procedure was initially formulated under the Bayesian framework with imputed values generated from the posterior distribution of the missing values given the observed data. It requires specifications of a joint parametric distribution over the full sample as well as a prior distribution for the model parameters.

Let y be the study variable subject to missingness and denote $\mathbf{y} = (y_1, y_2, \cdots, y_n)'$. The vector $\mathbf{y}$ can then be partitioned into $\mathbf{y} = (\mathbf{y}'_{\text{mis}}, \mathbf{y}'_{\text{obs}})'$, where $\mathbf{y}_{\text{mis}}$ consists of the nonresponses and $\mathbf{y}_{\text{obs}}$ is formed by all the observed values. The missing mechanism is assumed to be MAR. The imputation model specifies a parametric distribution $f(\mathbf{y} \mid \gamma)$ with model parameters γ. Under the ideal situation where $\gamma = \gamma_0$ is known, we could impute the missing values $\mathbf{y}_{\text{mis}}$ by $\mathbf{y}^*_{\text{mis}}$ which are generated from the conditional distribution of $\mathbf{y}_{\text{mis}}$ given the observed values $\mathbf{y}_{\text{obs}}$, i.e.,

$$\mathbf{y}^*_{\text{mis}} \quad \sim \quad f(\mathbf{y}_{\text{mis}} \mid \mathbf{y}_{\text{obs}}; \gamma_0) \,.$$

It follows that the imputation procedure is unbiased since $E(\mathbf{y}^*_{\text{mis}} - \mathbf{y}_{\text{mis}} \mid \mathbf{y}_{\text{obs}}) = \mathbf{0}$ under the assumed model, which further implies $E(\mathbf{y}^*_{\text{mis}} - \mathbf{y}_{\text{mis}}) = \mathbf{0}$. With the imputed values $\mathbf{y}^*_{\text{mis}}$, an imputed dataset is created. The process is repeated multiple times, independent of each other, to create multiple imputed datasets.

There are two practical implementations of the *multiple imputation* procedure. The first is to use the Bayesian framework with an assigned prior distribution $p(\gamma)$ for the parameters γ. Using the marginal distribution of $\mathbf{y}_{\text{obs}}$ with the given γ, one can obtain the posterior distribution of γ given $\mathbf{y}_{\text{obs}}$, denoted as $p(\gamma \mid \mathbf{y}_{\text{obs}})$. The so-called *posterior predictive distribution* for $\mathbf{y}_{\text{mis}}$ is given by

$$f(\mathbf{y}_{\text{mis}} \mid \mathbf{y}_{\text{obs}}) = \int f(\mathbf{y}_{\text{mis}} \mid \mathbf{y}_{\text{obs}}; \gamma) p(\gamma \mid \mathbf{y}_{\text{obs}}) d\gamma \,.$$

The imputed values $\mathbf{y}^*_{\text{mis}}$ can be generated by Monte Carlo sampling techniques from $f(\mathbf{y}_{\text{mis}} \mid \mathbf{y}_{\text{obs}})$ through two steps (Little and Rubin 2002; Kim and Shao 2013):

1. Generate γ^* from $p(\gamma \mid \mathbf{y}_{\text{obs}})$;
2. Generate $\mathbf{y}^*_{\text{mis}}$ from $f(\mathbf{y}_{\text{mis}} \mid \mathbf{y}_{\text{obs}}; \gamma^*)$ with the given γ^*.

The second implementation method uses a frequentist approach to multiple imputation, which was discussed by Wei and Tanner (1990). This is particularly appealing for scenarios with auxiliary information where the study variable y is subject to missingness but $\mathbf{x}$ is observed for all sampled units, which is the basic setting used in this chapter. A parametric model for y given $\mathbf{x}$ is specified as $f(y \mid \mathbf{x}, \gamma)$. Under the MAR assumption, the maximum likelihood estimator $\hat{\gamma}$ of the model parameters γ can be obtained by using data from the set of complete cases, i.e., $\{(y_i, \mathbf{x}_i), i \in \mathbf{S}_R\}$. If y_i is missing, the imputed value y_i^* is generated from the estimated conditional distribution of y given $\mathbf{x} = \mathbf{x}_i$, i.e.,

$$y_i^* \sim f(y \mid \mathbf{x}_i, \hat{\gamma}), \tag{9.7}$$

and the imputation procedure is carried out for each $i \in \mathbf{S}_M$ to create an imputed dataset. Multiple imputed datasets can then be created, independent of each other with the given $\hat{\gamma}$. Wang and Robins (1998) termed the resulting multiple imputation estimator as the *type B estimator*.

Let θ be the main parameter of interest for inference. Let $\hat{\theta}^{(d)}$ be the survey weighted estimator of θ based on the dth imputed dataset and the final survey weights. For instance, if $\theta = \mu_y$ is the finite population mean, then $\hat{\theta}^{(d)}$ is computed using the formula given in (9.3). The multiple imputation estimator of θ is then computed as

$$\hat{\theta}_{\text{MI}} = \frac{1}{D}\sum_{d=1}^{D}\hat{\theta}^{(d)}, \tag{9.8}$$

where D is the total number of imputed datasets. In practice, the value of D is often taken as 3, 5 or 10. Wang and Robins (1998) and Robins and Wang (2000) showed that the type B multiple imputation estimator under the frequentist approach is more efficient than the estimator under the Bayesian framework. We will briefly discuss variance estimation under multiple imputation in Sect. 9.6.

9.4.3 Fractional Imputation

Imputation for missing values brings another source of variation to the imputed estimator, and the inflation in variance can be substantial when the nonresponse rate is high. Single imputation is appealing for large survey organizations where it is highly desirable to create a single data file to be released for public use with different users (Brick and Kalton 1996). One of the main drawbacks for single imputation is the lack of efficiency. Multiple imputation can reduce the imputation variance component due to the averaging over multiple imputed estimators as shown in (9.8). The price for the reduction in variance under multiple imputation is the creation, storage and management of multiple imputed data files, which can be laborious and challenging for dealing with large samples in practice.

It turns out that the imputation variance can be reduced by using fractional imputation (Kalton and Kish 1984; Fay 1996). Under fractional imputation, each missing value is replaced by K (≥ 1) imputed ones, with the observed variables in the corresponding row of the data file duplicated K times, resulting in a single but enlarged data file. Each duplicated row also receives a fractional weight, representing the distributional importance of the imputed value.

Table 9.2 An illustration of fractionally imputed data file with $K = 3$ and two study variables

i	y_{i1}	y_{i2}	x_{i1}	x_{i2}	x_{i3}	w_i	w_{ik}
1	y_{11}	y_{12}	x_{11}	x_{12}	x_{13}	w_1	1
2	y^*_{211}	y_{22}	x_{21}	x_{22}	x_{23}	w_2	1/3
2	y^*_{212}	y_{22}	x_{21}	x_{22}	x_{23}	w_2	1/3
2	y^*_{213}	y_{22}	x_{21}	x_{22}	x_{23}	w_2	1/3
3	y_{31}	y^*_{321}	x_{31}	x_{32}	x_{33}	w_3	1/4
3	y_{31}	y^*_{322}	x_{31}	x_{32}	x_{33}	w_3	1/2
3	y_{31}	y^*_{323}	x_{31}	x_{32}	x_{33}	w_3	1/4
4	y_{41}	y_{42}	x_{41}	x_{42}	x_{43}	w_4	1
5	y^*_{511}	y^*_{521}	x_{51}	x_{52}	x_{53}	w_5	1/3
5	y^*_{512}	y^*_{522}	x_{51}	x_{52}	x_{53}	w_5	1/3
5	y^*_{513}	y^*_{523}	x_{51}	x_{52}	x_{53}	w_5	1/3
6	y_{61}	y_{62}	x_{61}	x_{62}	x_{63}	w_6	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	y_{n1}	y_{n2}	x_{n1}	x_{n2}	x_{n3}	w_n	1

Table 9.2 shows a fractionally imputed dataset based on the missing patterns presented in Table 9.1. The newly added last column w_{ik} represents the fractional weights, which satisfy $\sum_{k=1}^K w_{ik} = 1$ for each unit i. The fractional weights are automatically assigned as 1 for rows without missing values. For instance, the missing y_{21} is replaced by $K = 3$ imputed values y^*_{211}, y^*_{212} and y^*_{213}, with the fractional weights $w_{21} = w_{22} = w_{23} = 1/3$. The missing y_{32} is imputed by y^*_{321}, y^*_{322} and y^*_{323}, with the fractional weights $w_{31} = 1/4$, $w_{32} = 1/2$ and $w_{33} = 1/4$. The missing y_{51} and y_{52} are handled here with the same set of fractional weights but the two missing values may be imputed using the joint distribution of (y_1, y_2). See She and Wu (2019) for an example of a fully efficient joint fractional imputation for bivariate ordinal responses.

Consider a single study variable y where a missing y_i is imputed by y^*_{ik} with the associated fractional weight w_{ik}, $k = 1, \cdots, K$. The survey weighted fractional imputation estimator of μ_y is computed as

$$\hat{\mu}_{y\text{FI}} = \frac{1}{N}\left(\sum_{i\in\mathbf{S}_R} w_i y_i + \sum_{i\in\mathbf{S}_M} w_i \sum_{k=1}^K w_{ik} y^*_{ik}\right) .$$

The total number of rows in the fractionally imputed dataset is $m = n_R + Kn_M$. Suppose that we re-label the rows as $j = 1, 2, \cdots, m$. Let $\tilde{y}_j$ be either the observed value or the imputed one for y; let $\tilde{w}_j = w_i w_{ik}$, where w_i and w_{ik} are the corresponding entries in the last two columns of the imputed dataset shown in Table 9.2 for the jth row under the new labelling system. It follows that the fractional imputation estimator of μ_y can be written as

$$\hat{\mu}_{y\text{FI}} = \frac{1}{N}\sum_{j=1}^m \tilde{w}_j \tilde{y}_j . \tag{9.9}$$

In practice, the last two columns of Table 9.2 can be replaced by a single weight column for $\tilde{w}_j$. The estimator given by (9.9) is the standard survey weighted estimator treating the imputed dataset as if it is the original sample without missing values. It can be seen that $\sum_{j=1}^{m} \tilde{w}_j = \sum_{i \in \mathbf{S}} w_i$, which is the estimator $\hat{N}$ for the population size N and it remains the same with the fractionally imputed dataset.

Fractional imputation can be implemented with commonly used random imputation procedures. Kim and Fuller (2004) discussed fractional random hot deck imputation, which is appealing in practice since all imputed values are selected from the observed ones. The most crucial aspect of fractional imputation is the choice of fractional weights. Under an ideal situation, the K imputed values for y_i along with the fractional weights w_{ik} should approximate the distribution of y_i given $\mathbf{x}_i$. This is difficult to accomplish for general cases but possible for specific scenarios.

Suppose that y is an ordinal variable taking values $1, \cdots, J$. Let $w_{ik} = P(y_i = k \mid \mathbf{x}_i)$ be the conditional probability that y_i takes the value at level k, $k = 1, \cdots, J$, with the given $\mathbf{x}_i$. One strategy for fractional imputation is to impute a missing y_i by all the possible values $k = 1, \cdots, J$ with the probability w_{ik} as the fractional weights. She and Wu (2020) showed that this fractional imputation procedure for missing ordinal responses is optimal in terms of estimation efficiency. The conditional probabilities w_{ik} are unknown but can be estimated under the MAR assumption (She and Wu 2020).

9.5 Estimation with Missing Survey Data

We examine the *validity* of various estimation methods in the presence of item nonresponse. As discussed in Sect. 9.1.3, validity refers to the consistency of the point estimator. The conclusion depends on the parameters of interest and the method for handling missing data. It also depends on the models required for the estimation, which could be the model for the propensity scores or the model for imputation. For analytical use of survey data, there is also an analysis model with the associated parameters of interest. The analysis model and the imputation model may not be the same, due to the disconnection between the data file producers and the dataset users as well as the specific objectives of the scientific investigation. Once again, the missing mechanism is assumed to be MAR.

The discussions in this section are formulated through estimating equations without using survey weights, with the main parameters θ associated with a superpopulation model. The estimator $\hat{\theta}$ is defined as the solution to the standard estimating equations $\sum_{i=1}^{n} \mathbf{g}(y_i, \mathbf{x}_i; \theta) = \mathbf{0}$. Under certain regularity conditions, the estimator $\hat{\theta}$ is consistent for the true value of the parameters, θ_0, if the estimating functions $\mathbf{g}(y, \mathbf{x}; \theta)$ are unbiased, i.e., $E\{\mathbf{g}(y, \mathbf{x}; \theta_0)\} = \mathbf{0}$, under the assumed model. Theoretical details can be found in Godambe (1991b) and Section 3.2 of Tsiatis (2006) on the m-estimators. It can be shown that the survey weighted estimator, which is computed as the solution to $\sum_{i \in \mathbf{S}} w_i \mathbf{g}(y_i, \mathbf{x}_i; \theta) = \mathbf{0}$ with unbiased estimating functions, is consistent for the finite population parameters θ_N defined through the census estimating equations if the survey design is not confounded with the missing data mechanism. We focus on the unbiasedness of the estimating functions for the estimation problems to be discussed in this section. Note that $\delta_i = 1$ if y_i is observed and $\delta_i = 0$ otherwise. Also note that $\tau_i = \tau(\mathbf{x}_i) = P(\delta_i = 1 \mid \mathbf{x}_i) > 0$ for all i. Let $\{(\delta_i y_i, \mathbf{x}_i, \delta_i), i \in \mathbf{S}\}$ be the sample data, with the survey weights w_i omitted from the dataset.

9.5.1 Estimation of the Population Mean

The population mean θ can be defined through $E(y-\theta) = 0$ under the model on y. The estimator $\hat{\theta}$ in the absence of missing values is the solution to $\sum_{i=1}^{n}(y_i - \theta) = 0$, which is given by $\hat{\theta} = \bar{y}$, the simple sample mean.

The CCA Method The complete-case analysis estimator of θ is the solution to $\sum_{i\in\mathbf{S}} \delta_i(y_i - \theta) = 0$, which is given by $\hat{\theta}_{\text{CCA}} = \bar{y}_R = n_R^{-1}\sum_{i\in\mathbf{S}_R} y_i$. The estimator is not consistent unless the missing mechanism is MCAR. It can be checked directly that $E\{\delta(y-\theta)\} \neq 0$ under the MAR assumption. See also Problem 9.1.

The PSA Method The propensity score adjusted estimator of θ is the solution to

$$\sum_{i\in\mathbf{S}} g(y_i, \mathbf{x}_i, \delta_i; \theta) = \sum_{i\in\mathbf{S}} \frac{\delta_i}{\tau(\mathbf{x}_i)}(y_i - \theta) = 0\,,$$

where $g(y_i, \mathbf{x}_i, \delta_i; \theta) = \delta_i\{\tau(\mathbf{x}_i)\}^{-1}(y_i - \theta)$, which is given by

$$\hat{\theta}_{\text{PSA}} = \left\{\sum_{i\in\mathbf{S}_R} \frac{y_i}{\tau(\mathbf{x}_i)}\right\} \Big/ \left\{\sum_{i\in\mathbf{S}_R} \frac{1}{\tau(\mathbf{x}_i)}\right\}.$$

The estimator $\hat{\theta}_{\text{PSA}}$ is consistent since $g(y_i, \mathbf{x}_i, \delta_i; \theta)$ is unbiased under both the propensity score model and the model on y (Problem 9.6). The estimator remains consistent when the propensity scores $\tau(\mathbf{x}_i)$ are estimated by an assumed model. See Sects. 9.5.2 and 9.6 for further discussion.

Imputation-Based Methods The estimator $\hat{\theta}_{\text{IMP}}$ based on a single imputed dataset $\{(\tilde{y}_i, \mathbf{x}_i), i \in \mathbf{S}\}$, where $\tilde{y}_i = \delta_i y_i + (1-\delta_i)y_i^*$, is the solution to

$$\sum_{i\in\mathbf{S}} g(y_i, y_i^*, \delta_i; \theta) = \sum_{i\in\mathbf{S}} (\tilde{y}_i - \theta) = 0\,, \tag{9.10}$$

where $g(y_i, y_i^*, \delta_i; \theta) = \delta_i y_i + (1-\delta_i)y_i^* - \theta$. The estimating function $g(y_i, y_i^*, \delta_i; \theta)$ is unbiased under the imputation model and the model for y, as long as the imputation procedure is unbiased, and consequently, the imputation-based estimator $\hat{\theta}_{\text{IMP}}$ is consistent under the assumed imputation model. For multiple imputation with D imputed datasets, we can modify the equation (9.10) by re-defining

$$g(y_i, y_i^*, \delta_i; \theta) = \delta_i y_i + (1-\delta_i)\frac{1}{D}\sum_{d=1}^{D} y_{id}^* - \theta\,,$$

where y_{id}^* is the imputed value for y_i in the dth imputed dataset. Similarly, for fractional imputation, we can re-define

$$g(y_i, y_i^*, \delta_i; \theta) = \delta_i y_i + (1-\delta_i)\sum_{k=1}^{K} w_{ik} y_{ik}^* - \theta\,.$$

The estimators under multiple imputation or fractional imputation are consistent under the assumed imputation model.

9.5.2 Regression Analysis: Scenario I

Regression analysis is one of the most commonly used methods for scientific investigations. Depending on the specific type of the study variable y, it could be linear regression, logistic regression or analysis with a generalized linear model. The main parameters of interest are the regression coefficients θ defined through the mean function: $E(y \mid \mathbf{x}) = m(\mathbf{x}, \theta)$, where the functional form of $m(\cdot,\cdot)$ is known. For linear regression, we have $m(\mathbf{x}, \theta) = \mathbf{x}'\theta$; for logistic regression where y is a binary variable, we have $m(\mathbf{x}, \theta) = P(y = 1 \mid \mathbf{x})$ and

$$\log\left\{\frac{m(\mathbf{x},\theta)}{1-m(\mathbf{x},\theta)}\right\} = \mathbf{x}'\theta \qquad \text{or} \qquad m(\mathbf{x},\theta) = \frac{\exp(\mathbf{x}'\theta)}{1+\exp(\mathbf{x}'\theta)}\,.$$

Note that $\mathbf{x}$ is the vector of all available covariates from the dataset. We first consider scenarios where the data user conducts regression analysis with the study variable y using the same full set of covariates that are available from the dataset. In the absence of missing values, the estimator of the regression coefficients θ can be obtained by solving the estimating equations

$$\sum_{i\in\mathbf{S}} \mathbf{g}(y_i, \mathbf{x}_i; \theta) = \sum_{i\in\mathbf{S}} \mathbf{D}(\mathbf{x}_i, \theta)\{V(\mathbf{x}_i, \theta)\}^{-1}\big\{y_i - m(\mathbf{x}_i, \theta)\big\} = \mathbf{0}\,,$$

where $\mathbf{D}(\mathbf{x}_i, \theta) = \partial m(\mathbf{x}_i, \theta)/\partial\theta$ and $V(\mathbf{x}_i, \theta)$ is the variance function with a known form. See McCullagh and Nelder (1983) for further details on estimation under generalized linear models and She and Wu (2020) on regression analysis with ordinal y variables.

The CCA Method The complete-case analysis estimator of θ is obtained by solving $\sum_{i\in\mathbf{S}}\delta_i\mathbf{g}(y_i,\mathbf{x}_i;\theta)=\mathbf{0}$. The resulting $\hat{\theta}_{\text{CCA}}$ is consistent since

$$E\big\{\delta\mathbf{g}(y,\mathbf{x};\theta)\bigm|\mathbf{x}\big\}=E(\delta\mid\mathbf{x})\mathbf{D}(\mathbf{x},\theta)\{V(\mathbf{x},\theta)\}^{-1}\big\{E(y\mid\mathbf{x})-m(\mathbf{x},\theta)\big\}=\mathbf{0}\,,$$

due to the independence of δ and y given $\mathbf{x}$, which implies that $E\big\{\delta\mathbf{g}(y,\mathbf{x};\theta)\big\}=\mathbf{0}$. The consistency of $\hat{\theta}_{\text{CCA}}$ can also be argued directly by noting that $E(y\mid\mathbf{x},\delta)=E(y\mid\mathbf{x})$ under the MAR assumption. This conclusion is important since the CCA estimator is often used to build an imputation model on y with missing values.

The PSA Method The propensity scores $\tau_i=P(\delta_i=1\mid\mathbf{x}_i)$ can be estimated by assuming a logistic regression model over $(\delta,\mathbf{x})$ such that

$$\tau_i=\tau(\mathbf{x}_i,\phi)=\exp(\mathbf{x}_i'\phi)/\{1+\exp(\mathbf{x}_i'\phi)\}\,,\quad i\in\mathbf{S}\,,$$

where $\mathbf{x}_i$ is assumed to have 1 as its first component so that the model has an intercept. The maximum likelihood estimator $\hat{\phi}$ is the solution to the score equations

$$\sum_{i\in\mathbf{S}}\mathbf{g}_1(\delta_i,\mathbf{x}_i;\phi)=\sum_{i\in\mathbf{S}}\mathbf{x}_i\big\{\delta_i-\tau(\mathbf{x}_i,\phi)\big\}=\mathbf{0}\,. \tag{9.11}$$

The estimated propensity scores $\tau(\mathbf{x}_i,\hat{\phi})$ are then used to construct the PSA estimator $\hat{\theta}_{\text{PSA}}$ for the main parameters of interest, θ, which is the solution to

$$\sum_{i\in\mathbf{S}}\mathbf{g}_2(y_i,\mathbf{x}_i;\theta;\phi)=\sum_{i\in\mathbf{S}}\frac{\delta_i}{\tau(\mathbf{x}_i,\phi)}\mathbf{g}(y_i,\mathbf{x}_i;\theta)=\mathbf{0}\,, \tag{9.12}$$

where $\mathbf{g}(y_i,\mathbf{x}_i;\theta)=\mathbf{D}(\mathbf{x}_i,\theta)\{V(\mathbf{x}_i,\theta)\}^{-1}\big\{y_i-m(\mathbf{x}_i,\theta)\big\}$ as defined by the regression analysis for the main parameters, θ. It follows that the PSA estimator $\hat{\theta}_{\text{PSA}}$ is consistent, since the estimator $(\hat{\theta},\hat{\phi})$ for the combined parameters (θ,ϕ) can be viewed as the solution to the joint estimating equations (9.11) and (9.12) and both estimating functions $\mathbf{g}_1$ and $\mathbf{g}_2$ are unbiased under the assumed logistic regression model for τ_i and the regression model for the main analysis.

Imputation-Based Methods The estimator $\hat{\theta}_{\text{IMP}}$ based on a single imputed dataset is computed as the solution to

$$\sum_{i\in\mathbf{S}}\mathbf{g}(\tilde{y}_i,\mathbf{x}_i;\theta)=\sum_{i\in\mathbf{S}}\mathbf{D}(\mathbf{x}_i,\theta)\{V(\mathbf{x}_i,\theta)\}^{-1}\big\{\tilde{y}_i-m(\mathbf{x}_i,\theta)\big\}=\mathbf{0}\,, \tag{9.13}$$

where $\tilde{y}_i=\delta_iy_i+(1-\delta_i)y_i^*$. Note that $m(\mathbf{x}_i,\theta)=E(y_i\mid\mathbf{x}_i)$ is specified for the main analysis model by the data user. A crucial question for the consistency of the imputation-based estimator $\hat{\theta}_{\text{IMP}}$ is whether the semiparametric imputation model (9.2) with model parameters β or the parametric imputation model (9.7) with model parameters γ is compatible with the analysis model. If the same type of models are used for imputation and for the main analysis, we have $E(y_i^*\mid\mathbf{x}_i)=m(\mathbf{x}_i;\theta)$ given the true values of β or γ. It follows that the estimator $\hat{\theta}_{\text{IMP}}$

is consistent. For most commonly encountered analysis in practice, such as linear regression analysis or logistic regression analysis, the imputation model and the analysis model would have the same form as long as the same full set of covariates are used for both models.

The estimator $\hat{\theta}_{\text{FI}}$ based on a fractionally imputed dataset is computed in the same way as $\hat{\theta}_{\text{IMP}}$, with $\tilde{y}_i$ in (9.13) replaced by $\tilde{y}_i = \delta_i y_i + (1-\delta_i)\sum_{k=1}^K w_{ik} y_{ik}^*$. The operational procedure for the multiple imputation estimator $\hat{\theta}_{\text{MI}}$ given in (9.8) requires solving the equations (9.13) one at a time for each of the D imputed datasets. If the imputation is done through the frequentist approach, it is shown by Wang and Robins (1998) that the estimator $\hat{\theta}_{\text{MI}}$ is asymptotically equivalent to solving (9.13) only once, with $\tilde{y}_i = \delta_i y_i + (1-\delta_i)D^{-1}\sum_{d=1}^D y_{id}^*$, which amounts to using $K = D$ and $w_{ik} = D^{-1}$ for fractional imputation. The estimators based on fractional imputation and multiple imputation are both consistent if the imputation model and the analysis model are the same.

9.5.3 Regression Analysis: Scenario II

A practically important scenario for regression analysis occurs when the data analyst wishes to explore the relationship between y and a selected subset of covariates that are available in the dataset. This could be the case, for instance, when a specific scientific objective requires only certain covariates to be included in the analysis. Note that the MAR assumption involves all the covariates $\mathbf{x}$ in the dataset. For the imputation based approach, it is assumed that the imputed dataset is created independently by the data file producers and the imputation model involves all the covariates that are available.

Let $\mathbf{x} = (\mathbf{u}', \mathbf{v}')'$. The main analysis model only involves covariates $\mathbf{u}$. Without loss of generality, we use the same notation as in Sect. 9.5.2 and let $E(y \mid \mathbf{u}) = m(\mathbf{u}, \theta)$ and θ be the parameters of interest. In the absence of missing values, the estimator of θ is defined as the solution to $\sum_{i\in\mathbf{S}} \mathbf{g}(y_i, \mathbf{u}_i; \theta) = \mathbf{0}$, where $\mathbf{g}(y_i, \mathbf{u}_i; \theta) = \mathbf{D}(\mathbf{u}_i, \theta)\{V(\mathbf{u}_i, \theta)\}^{-1}\big\{y_i - m(\mathbf{u}_i, \theta)\big\}$.

The CCA Method The complete-case analysis estimator $\hat{\theta}_{\text{CCA}}$ is the solution to $\sum_{i\in\mathbf{S}} \delta_i \mathbf{g}(y_i, \mathbf{u}_i; \theta) = \mathbf{0}$. We have

$$E\big\{\delta_i \mathbf{g}(y_i, \mathbf{u}_i; \theta) \mid \mathbf{x}_i\big\} = \tau(\mathbf{x}_i)\mathbf{D}(\mathbf{u}_i, \theta)\{V(\mathbf{u}_i, \theta)\}^{-1}\big\{E(y_i \mid \mathbf{x}_i) - E(y_i \mid \mathbf{u}_i)\big\}.$$

The estimating function $\delta\mathbf{g}(y, \mathbf{u}; \theta)$ is typically not unbiased unless $E(y \mid \mathbf{u}, \mathbf{v}) = E(y \mid \mathbf{u})$, i.e., y and $\mathbf{v}$ are unrelated given $\mathbf{u}$. The CCA estimator for the regression coefficients θ is invalid under this setting unless the missing mechanism does not depend on $\mathbf{v}$.

The PSA Method The propensity scores $\tau_i = \tau(\mathbf{x}_i, \phi)$ are assumed to depend on the full set of covariates $\mathbf{x}$. The propensity score adjusted estimator $\hat{\theta}_{\text{PSA}}$ is the solution to $\sum_{i\in\mathbf{S}}\{\delta_i/\tau(\mathbf{x}_i,\phi)\}\mathbf{g}(y_i,\mathbf{u}_i;\theta)=\mathbf{0}$, with ϕ satisfying equations (9.11). The estimator $\hat{\theta}_{\text{PSA}}$ is consistent under the MAR assumption (Problem 9.7).

Imputation-Based Methods It is assumed that the imputation model uses the full set of $\mathbf{x}$ variables. Let β be the parameters for the imputation model and $E(y\mid\mathbf{x})=m(\mathbf{x},\beta)$. With the given β, the imputed value y_i^* satisfies $E(y_i^*-y_i\mid\mathbf{x}_i)=0$. The imputation based estimator $\hat{\theta}_{\text{IMP}}$ is the solution to

$$\sum_{i\in\mathbf{S}}\{\delta_i\mathbf{g}(y_i,\mathbf{u}_i;\theta)+(1-\delta_i)\mathbf{g}(y_i^*,\mathbf{u}_i;\theta)\}=\mathbf{0}\,.$$

The estimator $\hat{\theta}_{\text{IMP}}$ is consistent for the regression coefficients θ (Problem 9.7). The multiple imputation estimator and the fractional imputation estimator of θ are also consistent.

9.6 Variance Estimation

Variance estimation with missing survey data is a challenging topic. A joint randomization framework involving related sources of variation is required, and the total variance may consist of two or more variance components from the sampling design, the propensity score model, the imputation model and the imputation method, and the main analysis model. We briefly discuss some standard scenarios in this section. The missing mechanism is assumed to be MAR.

9.6.1 PSA Estimators

Let $\tau_i=\tau(\mathbf{x}_i,\phi)$ be the propensity scores and the estimating functions for the parameter ϕ be given by $\mathbf{g}_1(\delta_i,\mathbf{x}_i;\phi)$, which is specified in (9.11) under the logistic regression model. One of the assumptions for the propensity score adjusted method with a survey sample is that the missing mechanism and the survey design are not confounded. In practice this issue can be alleviated by using the survey weighted estimating equations for ϕ, which are given by

$$\hat{\mathbf{U}}_2(\phi)=\sum_{i\in\mathbf{S}}w_i\mathbf{g}_1(\delta_i,\mathbf{x}_i;\phi)=\mathbf{0}\,. \tag{9.14}$$

Let θ be the main parameters of interest, defined through estimating functions $\mathbf{g}_2(y_i,\mathbf{x}_i;\theta)$. The propensity score adjusted survey weighted estimator of θ is the solution to the estimating equations

$$\hat{\mathbf{U}}_1(\theta,\phi)=\sum_{i\in\mathbf{S}}w_i\,\frac{\delta_i}{\tau(\mathbf{x}_i,\phi)}\,\mathbf{g}_2(y_i,\mathbf{x}_i;\theta)=\mathbf{0}\,. \tag{9.15}$$

In practice, the estimator $\hat{\phi}$ is obtained by solving (9.14), and the estimator $\hat{\theta}$ is computed as the solution to (9.15) with ϕ replaced by $\hat{\phi}$.

Variance estimation for $\hat{\theta}$, which is the estimator for the main parameters of interest, can be handled by combining the two estimating equations into a single system. Let $\hat{\mathbf{U}}(\theta,\phi) = (\hat{\mathbf{U}}_1'(\theta,\phi), \hat{\mathbf{U}}_2'(\phi))'$. With some misuse of notation without causing any confusion, let $\psi = (\theta', \phi')'$ be the true values of the parameters defined through the census estimating equations and $\hat{\psi} = (\hat{\theta}', \hat{\phi}')'$ be the solutions to $\hat{\mathbf{U}}(\theta,\phi) = \mathbf{0}$. Since $\hat{\psi}$ is a consistent estimator of ψ, we have $\hat{\psi} = \psi + o_p(1)$ (component-wise), and the standard Taylor series expansion leads to

$$\hat{\psi} = \psi - \mathbf{H}^{-1}\hat{\mathbf{U}}(\theta,\phi) + O_p\big\{\|\hat{\psi} - \psi\|^2\big\},$$

where

$$\mathbf{H} = \begin{pmatrix} \mathbf{H}_{11} & \mathbf{H}_{12} \\ \mathbf{0} & \mathbf{H}_{22} \end{pmatrix} \quad \text{and} \quad \mathbf{H}^{-1} = \begin{pmatrix} \mathbf{H}_{11}^{-1} & -\mathbf{H}_{11}^{-1}\mathbf{H}_{12}\mathbf{H}_{22}^{-1} \\ \mathbf{0} & \mathbf{H}_{22}^{-1} \end{pmatrix},$$

and the matrix blocks are defined as

$$\mathbf{H}_{11} = E\Big\{\frac{\partial}{\partial\theta}\sum_{i\in\mathbf{S}} w_i \frac{\delta_i}{\tau(\mathbf{x}_i,\phi)}\,\mathbf{g}_2(y_i,\mathbf{x}_i;\theta)\Big\},$$

$$\mathbf{H}_{12} = E\Big\{\frac{\partial}{\partial\phi}\sum_{i\in\mathbf{S}} w_i \frac{\delta_i}{\tau(\mathbf{x}_i,\phi)}\,\mathbf{g}_2(y_i,\mathbf{x}_i;\theta)\Big\},$$

$$\mathbf{H}_{22} = E\Big\{\frac{\partial}{\partial\phi}\sum_{i\in\mathbf{S}} w_i\mathbf{g}_1(\delta_i,\mathbf{x}_i;\phi)\Big\}.$$

The estimating functions $\mathbf{g}_1$ and $\mathbf{g}_2$ are assumed to be smooth in ϕ and θ. Note that the expectations in defining the blocks in the matrix $\mathbf{H}$ are for theoretical purposes only and are not required for the final variance estimator. The asymptotic variance formula for $\hat{\psi} = (\hat{\theta}', \hat{\phi}')'$ has the conventional sandwich form and is given by

$$Var\big(\hat{\psi}\big) \doteq \mathbf{H}^{-1} Var\big\{\hat{\mathbf{U}}(\theta,\phi)\big\}(\mathbf{H}')^{-1}.$$

There are two major sources of variation for the PSA estimator: the probability sampling design p for selecting the sample $\mathbf{S}$ and the missing mechanism described by the propensity score model q on δ. We further have

$$Var\big\{\hat{\mathbf{U}}(\theta,\phi)\big\} = V_pE_q\big\{\hat{\mathbf{U}}(\theta,\phi)\big\} + E_pV_q\big\{\hat{\mathbf{U}}(\theta,\phi)\big\}.$$

Under the model q, we have $E_q(\delta \mid \mathbf{x}) = \tau(\mathbf{x},\phi)$ and $V_q(\delta \mid \mathbf{x}) = \tau(\mathbf{x},\phi)\{1 - \tau(\mathbf{x},\phi)\}$. The variance component $V_p(\cdot)$ depends on the sampling design. The upper left block of $Var\big(\hat{\psi}\big)$ is $Var\big(\hat{\theta}\big)$. A variance estimator can be constructed through plug-in estimators for all the required components. Details need to be worked out case-by-case. See Problem 9.8 for variance estimation for the PSA estimator of the finite population mean.

9.6.2 Single Imputation Based Estimators

The variance of an imputation-based estimator depends on the parameters of interest, the survey design, the imputation model and the imputation method. As discussed in Sect. 9.4.1, there exist a variety of imputation methods under different assumptions for the imputation model. Variance estimation for imputation-based estimators is itself a topic of research with a growing literature. Haziza (2009) and Chen and Haziza (2019) contain overviews on imputation for item non-response and variance estimation with imputed survey data.

We consider random imputation procedures where the missing y_i is imputed by y_i^* such that $E_\xi(y_i^* \mid \mathbf{x}_i) = m(\mathbf{x}_i, \beta)$ and $V_\xi(y_i^* \mid \mathbf{x}_i) = v(\mathbf{x}_i)\sigma^2$ under the imputation model, ξ. Let θ be the main parameters of interest defined through the census estimating equations with estimating functions $\mathbf{g}(y, \mathbf{x}; \theta)$. The imputation-based estimator $\hat{\theta}$ is the solution to

$$\hat{\mathbf{U}}^*(\theta) = \sum_{i \in \mathbf{S}} w_i \big\{\delta_i \mathbf{g}(y_i, \mathbf{x}_i; \theta) + (1 - \delta_i)\mathbf{g}(y_i^*, \mathbf{x}_i; \theta)\big\} = \mathbf{0}\,.$$

It is assumed that the imputation procedure ensures the consistency of the estimator $\hat{\theta}$. The asymptotic variance of $\hat{\theta}$ can be derived under the joint randomization of the sampling design p and the imputation model ξ. This is the so-called *imputation model approach* to variance estimation with imputed values (Haziza 2009) where the missing indicators δ_i are viewed as fixed and are conceptually defined for every unit in the finite population regardless whether the unit is selected in the sample or not. It follows that

$$Var\big(\hat{\theta}\big) \doteq \mathbf{H}^{-1} Var\big\{\hat{\mathbf{U}}^*(\theta)\big\}(\mathbf{H}')^{-1}\,,$$

where

$$\mathbf{H} = E\Big[\frac{\partial}{\partial\theta}\sum_{i \in \mathbf{S}} w_i \big\{\delta_i \mathbf{g}(y_i, \mathbf{x}_i; \theta) + (1 - \delta_i)\mathbf{g}(y_i^*, \mathbf{x}_i; \theta)\big\}\Big]\,,$$

and the expectation is with respect to both p and ξ. The estimating functions $\mathbf{g}$ are assumed to be smooth in θ. We also have

$$Var\big\{\hat{\mathbf{U}}^*(\theta)\big\} = V_p E_\xi\big\{\hat{\mathbf{U}}^*(\theta)\big\} + E_p V_\xi\big\{\hat{\mathbf{U}}^*(\theta)\big\}\,, \tag{9.16}$$

with the evaluations of $E_\xi(\cdot)$ and $V_\xi(\cdot)$ carried out under the assumed imputation model for y_i^*, and $E_p(\cdot)$ and $V_p(\cdot)$ under the sampling design. A variance estimator can then be constructed by collecting estimates for all involved components. See Problem 9.9 on variance estimation for the population mean under random regression imputation.

Variance estimation for estimators based on deterministically imputed datasets can be developed under the so-called *response model approach*. The imputed values y_i^* are treated as fixed and a variance decomposition similar to (9.16) can be done under the joint randomization of the sampling design p for the probability sample $\mathbf{S}$ and the propensity score model q on the missing indicator variables δ_i.

9.6.3 Multiple Imputation Estimators

The multiple imputation estimator of θ is computed as $\hat{\theta}_{\mathrm{MI}} = D^{-1}\sum_{d=1}^{D}\hat{\theta}^{(d)}$, where $\hat{\theta}^{(d)}$ is computed based on the dth imputed dataset using a standard formula by treating imputed values as if they are observed. One of the major appealing features of multiple imputation is that the variance of $\hat{\theta}_{\mathrm{MI}}$ can be estimated using a simple combining rule (Rubin 1987; Little and Rubin 2002)

$$v\big(\hat{\theta}_{\mathrm{MI}}\big) = \bar{W}_{\mathrm{D}} + \frac{D+1}{D}B_{\mathrm{D}}\,, \tag{9.17}$$

where $\bar{W}_{\mathrm{D}}$ is the average of the within-imputation variance estimators and B_{D} is the estimated between-imputation variance component, which are computed respectively as

$$\bar{W}_{\mathrm{D}} = \frac{1}{D}\sum_{d=1}^{D}\hat{V}^{(d)} \quad \text{and} \quad B_{\mathrm{D}} = \frac{1}{D-1}\sum_{d=1}^{D}\big(\hat{\theta}^{(d)} - \hat{\theta}_{\mathrm{MI}}\big)\big(\hat{\theta}^{(d)} - \hat{\theta}_{\mathrm{MI}}\big)'\,.$$

The term $\hat{V}^{(d)}$ is the variance estimator of $\hat{\theta}^{(d)}$ computed from the dth imputed dataset using a standard formula, i.e., treating the imputed values as if they are observed. The term $(D+1)/D$ in (9.17) is an adjustment for the finite number D of imputations.

The variance formula (9.17) is derived under the assumption that the analysis procedure for the parameters of interest, θ, is congenial to the imputation model. In other words, the analysis model and the imputation model have the same set of assumptions. See Meng (1994) and Xie and Meng (2017) for detailed discussions on the congeniality condition and the validity of Rubin's combining rule for variance estimation. In practice, the data imputer and the data analyst are typically disconnected, and the model used for analysis can be different from (i.e., uncongenial to) the model used for imputation. Under such scenarios, the variance formula given by (9.17) is not valid. See Fay (1992), Meng (1994) and Kim and Shao (2013) for further discussions.

9.7 Some Additional Notes on Missing Data Problems

Nonresponse and missing data are ubiquitous problems across many scientific fields and have remained as an active research area for the past 30+ years. In dealing with these, the first crucial issue is the missing data mechanism, which determines how missing data should be handled and analyzed. If data are missing completely at random, the complete-case analysis method would provide valid inferences. There exist statistical tests for whether the missing mechanism is MCAR. See, for instance, Little (1988), Chen and Little (1999) and Zhang et al. (2019a). Most commonly used methods for handling missing data are developed under the MAR assumption. The MAR assumption itself, however, cannot be tested without additional information. It is possible to conduct a sensitivity analysis on the MAR assumption, as shown by Zhao et al. (2019). Scenarios with data not-missing-at-random (i.e., nonignorable missing data) are extremely difficult to handle. Further assumptions or additional information are needed to develop practically useful inferential procedures. Chapter 6 of Kim and Shao (2013) contains discussions and some related references on nonignorable missing data.

Joint imputation procedures for multivariate study variables with missing values are another important problem that requires further investigation. The general goal is to preserve the correlation structure or the joint distribution for the imputed data files to facilitate valid inferences. The sequential regression multiple imputation procedure (Raghunathan et al. 2001) is popular among practitioners but it lacks theoretical justification. For longitudinal study variables, the issue becomes even more complicated due to different missing data patterns (i.e., monotone versus intermittent). Carrillo et al. (2011) developed a marginal imputation procedure for longitudinal surveys with missing observations, which works well when the rate of missing data is not too high. For bivariate ordinal study variables with missing values, She and Wu (2019) proposed a fully efficient joint fractional imputation procedure.

Data could also be missing by design. One such problem arises with the popular pretest-posttest study designs, which assign each participating unit to either a control group or a treatment group. The main objective is to assess the average treatment effect for the study variable. When the treatment assignments are randomized, the analysis can be handled by standard two-sample techniques. For many medical studies of disease treatment, or examinations of the effectiveness of an intervention, the treatment assignments are not randomized. Units in the treatment group do not have the study variable measured under the control conditions, and vice versa, which leads to a data missing-by-design scenario. Chen et al. (2015), Chen et al. (2016) and Chen et al. (2018) applied missing data techniques to analyze data from such designs. Zhang et al. (2019b) further discussed scenarios under pretest-posttest designs with missing observations among the two groups and developed empirical likelihood-based inferential procedures for the average treatment effect.

9.8 Problems

9.1 (A Simple Example of Complete Case Analysis) Suppose that **S** is selected by simple random sampling without replacement and y is subject to missingness. Let $\mathbf{S}_R$ be the set of respondents such that y_i is observed if $i \in \mathbf{S}_R$ and y_i is missing otherwise. Let $\bar{y}_R = n_R^{-1} \sum_{i \in \mathbf{S}_R} y_i$ be the simple sample mean for the "complete cases", where n_R is the number of respondents.

(a) Show that $\bar{y}_R$ is a consistent estimator of the population mean μ_y if y is missing completely at random.
(b) Show (analytically or by giving a counter example) that $\bar{y}_R$ is not a consistent estimator of μ_y when the missing mechanism for y is not MCAR.

9.2 (A Newton-Raphson Procedure for Logistic Regression)

(a) Show in detail how to derive the score function for the maximum likelihood estimator of θ for the logistic regression model described in Section 9.2.
(b) Describe the Newton-Raphson iterative procedure for finding the maximum likelihood estimator $\hat{\theta}$.
(c) Write an R function to implement the Newton-Raphson procedure.

Note: The Newton-Raphson iterative procedure for the survey weighted logistic regression analysis has been described in Sect. 7.3.2.

9.3 (Asymptotic Properties of the PSA Estimator) Suppose that y is missing at random with known propensity scores $\tau_i = P(\delta_i = 1 \mid \mathbf{x}_i)$. Let $w_i = 1/\pi_i$ be the basic design weights and suppose the population size N is known. The PSA estimator of μ_y under this setting is given by $\hat{\mu}_{y\text{PSA}} = N^{-1} \sum_{i \in \mathbf{S}_R} y_i/(\pi_i \tau_i)$.

(a) Show that $\hat{\mu}_{y\text{PSA}}$ is an exactly unbiased estimator of μ_y under the joint (q, p)-randomization framework.
(b) Show how to derive the theoretical variance of $\hat{\mu}_{y\text{PSA}}$.
(**Hint**: $Var(\cdot) = V_p\{E_q(\cdot)\} + E_p\{V_q(\cdot)\}$)
(c) Show how to develop a variance estimator based on the observed dataset $\{(y_i, \tau_i, \pi_i), i \in \mathbf{S}_R\}$.
(d) Suppose that the PSA estimator of μ_y is re-defined as

$$\hat{\mu}_{y\text{PSA}} = \left(\sum_{i \in \mathbf{S}_R} \frac{y_i}{\tau_i \pi_i} \right) / \left(\sum_{i \in \mathbf{S}_R} \frac{1}{\tau_i \pi_i} \right).$$

Show how to derive the asymptotic variance of $\hat{\mu}_{y\text{PSA}}$ using the linearization method.

9.4 (Asymptotic Properties of the Imputation-Based Estimator) Consider the semiparametric imputation model specified in (9.2) with the imputed value of y_i given by $y_i^* = m(\mathbf{x}_i, \beta_0)$, $i \in \mathbf{S}_M$, where β_0 are the true values of the model parameters and are assumed to be known. Furthermore, the final survey weights are the basic design weights $w_i = 1/\pi_i$ with no adjustment.

(a) Show that the estimator $\hat{\mu}_{y\text{IMP}}$ is an exactly unbiased estimator of μ_y under the joint (p, ξ) framework.
(b) Find the exact expression for $V\big(\hat{\mu}_{y\text{IMP}} - \mu_y\big)$.
(c) Develop a variance estimator using the imputed dataset.

9.5 (Doubly Robust Estimator for the Finite Population Mean) Consider the estimator $\hat{\mu}_{y\text{DR}}$ described in equation (9.4) with the given τ_i and β_0. Suppose that the final survey weights are the basic design weights $w_i = 1/\pi_i$.

(a) Show that the estimator $\hat{\mu}_{y\text{DR}}$ is an exactly unbiased estimator of μ_y under the joint (p, q) framework, regardless of the outcome regression model.
(b) Show that the estimator $\hat{\mu}_{y\text{DR}}$ is an exactly unbiased prediction estimator of μ_y under the joint (p, ξ) framework, irrespective of the model for the propensity scores.

9.6 (Consistency of the PSA Estimator for the Population Mean) Consider the propensity score adjusted estimator for the population mean where $g(y_i, \mathbf{x}_i, \delta_i; \theta) = \delta_i\{\tau(\mathbf{x}_i)\}^{-1}(y_i - \theta)$.

(a) Show that the estimating function $g(y, \mathbf{x}, \delta; \theta)$ is unbiased under the propensity score model and the model on y when the missing mechanism is MAR.
(b) Argue that $g(y, \mathbf{x}, \delta; \theta)$ is not unbiased when the missing mechanism is MNAR.

9.7 (Consistency of the PSA and Imputation-Based Estimators for the Regression Coefficients) Consider the second scenario of regression analysis with missing data. Let $E(y \mid \mathbf{u}) = m(\mathbf{u}, \theta)$, where $\mathbf{u}$ is a subset of $\mathbf{x}$. Let $\mathbf{g}(y, \mathbf{u}; \theta) = \mathbf{D}(\mathbf{u}, \theta)\{V(\mathbf{u}, \theta)\}^{-1}\big\{y - m(\mathbf{u}, \theta)\big\}$.

(a) The PSA estimator: Let $\tau(\mathbf{x}, \phi)$ be the propensity score. Show that the estimating functions $\{\delta/\tau(\mathbf{x}, \phi)\}\mathbf{g}(y, \mathbf{u}; \theta)$ are unbiased under the MAR assumption.
(b) The imputation-based estimator: With the given β for the imputation model, we have $E(y_i^* - y_i \mid \mathbf{x}_i) = 0$. Show that the estimating functions $\{\delta_i \mathbf{g}(y_i, \mathbf{u}_i; \theta) + (1 - \delta_i)\mathbf{g}(y_i^*, \mathbf{u}_i; \theta)\}$ are unbiased.

9.8 (Variance Estimation for the PSA Estimator of μ_y) Consider the PSA estimator of μ_y with the logistic regression model for the propensity scores and assume that a design-based variance estimator for the survey weighted estimator $\sum_{i\in \mathbf{S}} w_i y_i$ without missing values is available.

(a) Derive the asymptotic variance formula for $\hat{\mu}_{y\text{PSA}}$ as outlined in Sect. 9.6.1.
(b) Construct a variance estimator for $\hat{\mu}_{y\text{PSA}}$.

Note: Consider $w_i = 1/\pi_i$ to be the basic design weights and assume that π_{ij} are available.

9.9 (Variance Estimation for μ_y Under Random Regression Imputation) Consider the random regression imputation described in Sect. 9.4.1 under Scenario III. The imputation model ξ satisfies $E_\xi(y_i^* \mid \mathbf{x}_i) = \mathbf{x}_i'\beta$ and $V_\xi(y_i^* \mid \mathbf{x}_i) = \sigma^2$, and the model parameters β and σ^2 can be estimated by the least squares method using observed data. The estimator $\hat{\mu}_y$ is the solution to $\sum_{i\in\mathbf{S}} w_i \big\{\delta_i(y_i-\theta)+(1-\delta_i)(y_i^*-\theta)\big\} = 0$.

(a) Derive the asymptotic variance formula for $\hat{\mu}_y$ using the approach described in Sect. 9.6.2.

(b) Construct a variance estimator for $\hat{\mu}_y$.

Note: Consider $w_i = 1/\pi_i$ to be the basic design weights and assume that π_{ij} are available.

Chapter 10
Resampling and Replication Methods

Estimation of standard errors and approximation to the sampling distribution of test statistics are crucial components in analysis of complex survey data. Linearization variance estimation and normal-based approximation theory are used for scenarios where analytic variance expressions can be developed and sufficient design information is available to the data analyst. For more complicated inferential problems, the linearization method requires high level analytic tools to tackle problems one-at-a-time and may not always be feasible. This is especially cumbersome for large survey datasets which are often analyzed by multiple data users with different inferential objectives.

Resampling and replication methods are computer intensive techniques for mitigating the analytic burden of data users through structured and often Monte Carlo simulation-based procedures. Commonly used *resampling methods* include *jackknife*, *balanced repeated replications* and *bootstrap*, where artificial samples are selected or created from the original sample (i.e., resampled). Each of these samples is used to create a replicated version of the original estimator or test statistic, and the sequence of the replicates is used to construct a variance estimator or to approximate the sampling distribution of the original test statistic. Any resampling method can be referred to as a *replication method*, but the latter may not always involve a resampling procedure, as shown in Sect. 10.5. We focus on bootstrap methods in this chapter. Other methods along with a few relevant references are briefly discussed in Sect. 10.7.

10.1 The With-Replacement Bootstrap

Let $\{(y_i, \mathbf{x}_i, d_i),\ i \in \mathbf{S}\}$ be the sample data, where the basic design weights d_i are treated as part of the dataset. Let n be the sample size and $\hat{T}_{y\text{HT}} = \sum_{i \in \mathbf{S}} d_i y_i$ be the Horvitz-Thompson estimator of the population total T_y. It is assumed that additional

C. Wu, M. E. Thompson, *Sampling Theory and Practice*, ICSA Book
Series in Statistics, https://doi.org/10.1007/978-3-030-44246-0_10

design information such as second order inclusion probabilities is not available to data users. The primary goal of this section is to construct a variance estimator for $\hat{T}_{y\mathrm{HT}}$ using bootstrap methods. If a bootstrap method provides valid variance estimator for the Horvitz-Thompson estimator, it is also immediately applicable to other more complicated finite population parameters where the estimator is a function of several Horvitz-Thompson estimators.

10.1.1 Single-Stage Sampling

We first consider general single-stage unequal probability sampling designs. The point estimator $\hat{T}_{y\mathrm{HT}}$ involves only the first order inclusion probabilities but its theoretical variance generally requires the second order inclusion probabilities. When the sampling fraction n/N is small, a common practice for the purpose of variance estimation is to treat the sample as if it is selected with replacement. This leads to overestimation of the variance but the difference is negligible if the sampling fraction is very small. Let z_i be the normalized size variable for a PPS without replacement sampling design such that $z_i > 0$, $\sum_{i=1}^N z_i = 1$ and $\pi_i = P(i \in \mathbf{S}) = nz_i$. The Hansen-Hurwitz estimator of T_y under PPS sampling with replacement is given by $\hat{T}_{y\mathrm{HH}} = \sum_{i=1}^n y_i/(nz_i)$ as specified by (4.13), which is algebraically identical to the Horvitz-Thompson estimator $\hat{T}_{y\mathrm{HT}} = \sum_{i\in\mathbf{S}} d_i y_i$ if no unit is selected more than once under the with-replacement procedure. The estimator $\hat{T}_{y\mathrm{HH}}$ is practically identical to $\hat{T}_{y\mathrm{HT}}$ when n/N is small. The unbiased variance estimator of $\hat{T}_{y\mathrm{HH}}$ given in Part (c) of Theorem 4.6 can be re-written as

$$v\big(\hat{T}_{y\mathrm{HH}}\big) = \frac{1}{n(n-1)} \sum_{i=1}^n \Big(\frac{y_i}{z_i} - \hat{T}_{y\mathrm{HH}}\Big)^2 = \frac{n}{n-1} \sum_{i\in\mathbf{S}} \Big(d_i y_i - \frac{\hat{T}_{y\mathrm{HT}}}{n}\Big)^2 . \qquad (10.1)$$

The following with-replacement bootstrap procedures were proposed by McCarthy and Snowden (1985):

1. Select a bootstrap sample $\mathbf{S}^*$ of size n from the original sample $\mathbf{S}$ using simple random sampling with replacement; denote the bootstrap sample data by $\{(y_i^*, \mathbf{x}_i^*, d_i^*), i \in \mathbf{S}^*\}$.
2. Compute the bootstrap version of $\hat{T}_{y\mathrm{HT}}$ using the bootstrap sample data: $\hat{T}^*_{y\mathrm{HT}} = \sum_{i\in\mathbf{S}^*} d_i^* y_i^*$.
3. Repeat Steps 1 and 2 a large number B times, independently, to obtain $\hat{T}^{*(b)}_{y\mathrm{HT}}$, $b = 1, \cdots, B$.

The bootstrap variance estimator for $\hat{T}_{y\mathrm{HT}}$ is given by $V_*\big(\hat{T}^*_{y\mathrm{HT}}\big)$, where $V_*(\cdot)$ denotes the variance under the bootstrap procedures (Steps 1 and 2), which can be approximated by the empirical variance of the simulated sequence from Step 3:

$$v_{\text{boot}}\big(\hat{T}_{y\text{HT}}\big) = \frac{1}{B-1}\sum_{b=1}^{B}\Big(\hat{T}_{y\text{HT}}^{*(b)} - \bar{T}_{y\text{HT}}^{*}\Big)^2, \tag{10.2}$$

where $\bar{T}_{y\text{HT}}^{*} = B^{-1}\sum_{b=1}^{B}\hat{T}_{y\text{HT}}^{*(b)}$. The factor $1/(B-1)$ in (10.2) may be replaced by $1/B$ since B is large, and $\bar{T}_{y\text{HT}}^{*}$ can be replaced by the original estimator $\hat{T}_{y\text{HT}}$. The bootstrap variance estimator $v_{\text{boot}}\big(\hat{T}_{y\text{HT}}\big)$ is valid under PPS sampling with replacement when $(n-1)/n \doteq 1$ and approximately valid under PPS sampling without replacement when $(n-1)/n \doteq 1$ and the sampling fraction n/N is small (Problem 10.1).

When the sampling fraction is not small, the with-replacement bootstrap variance estimator becomes invalid. This can be seen from the simple case where the original sample is selected by SRSWOR and $\hat{T}_{y\text{HT}} = N\bar{y}$. The design-unbiased variance estimator and the bootstrap variance estimator (using the result of Problem 10.1) are given respectively by

$$v\big(\hat{T}_{y\text{HT}}\big) = N^2(1-f)\frac{s_y^2}{n} \quad \text{and} \quad V_*\big(\hat{T}_{y\text{HT}}^*\big) = N^2\Big(1-\frac{1}{n}\Big)\frac{s_y^2}{n},$$

where $f = n/N$ is the sampling fraction and s_y^2 is the sample variance. The two estimators are the same if $f = n^{-1}$ or equivalently $n/N = N^{-1/2}$, which requires the sampling fraction to be small. A possible modification to the with-replacement bootstrap under designs with non-negligible sample fractions is to rescale the variance estimator $V_*\big(\hat{T}_{y\text{HT}}^*\big)$ by multiplying the factor $(1-f)/(1-n^{-1})$. Another modification to the standard with-replacement bootstrap is to select bootstrap samples of size m instead of the original sample size n. The value of m is chosen such that the bootstrap variance $V_*\big(\hat{T}_{y\text{HT}}^*\big)$ matches exactly the design-unbiased variance estimator $v\big(\hat{T}_{y\text{HT}}\big)$ when the original sample $\mathbf{S}$ is selected by SRSWOR. See Problem 10.2 for further details.

10.1.2 Stratified Sampling

Let $\{(y_{hi}, \mathbf{x}_{hi}, d_{hi}),\ i \in \mathbf{S}_h,\ h = 1, \cdots, H\}$ be the sample data under a stratified sampling design, where $d_{hi} = 1/\pi_{hi}$ are the design weights and $\pi_{hi} = P(i \in \mathbf{S}_h)$ are the inclusion probabilities for stratum h. Let $(n_1, \cdots, n_{\text{H}})$ be the stratum sample sizes. The Horvitz-Thompson estimator of T_y is given by $\hat{T}_{y\text{HT}} = \sum_{h=1}^{H}\sum_{i\in\mathbf{S}_h} d_{hi}y_{hi}$. The with-replacement bootstrap procedures for stratified sampling are as follows.

1. Select a bootstrap sample $\mathbf{S}_h^*$ of size n_h from the original stratum sample $\mathbf{S}_h$ using simple random sampling with replacement, independently for each stratum; denote the bootstrap sample data by $\{(y_{hi}^*, \mathbf{x}_{hi}^*, d_{hi}^*), i \in \mathbf{S}_h^*,\ h = 1, \cdots, H\}$.

2. Compute the bootstrap version of $\hat{T}_{y\mathrm{HT}}$ using the bootstrap sample data: $\hat{T}^*_{y\mathrm{HT}} = \sum_{h=1}^{H} \sum_{i \in \mathbf{S}^*_h} d^*_{hi} y^*_{hi}$.
3. Repeat Steps 1 and 2 a large number B times, independently, to obtain $\hat{T}^{*(b)}_{y\mathrm{HT}}$, $b = 1, \cdots, B$.

The bootstrap variance estimator $v_{\mathrm{boot}}\big(\hat{T}_{y\mathrm{HT}}\big)$ computed in the same form of (10.2) is valid if $(n_h - 1)/n_h \doteq 1$ and n_h/N_h is small for all h.

A practically important scenario under stratified sampling design is that the number of strata, H, is large but the stratum sample sizes n_h are bounded. In the most extreme case we have $n_h = 2$ for all h. The with-replacement bootstrap variance estimator is not valid under such scenarios. Under stratified random sampling where $d_{hi} = N_h/n_h$, the with-replacement bootstrap variance estimator reduces to

$$V_*\big(\hat{T}^*_{y\mathrm{HT}}\big) = N^2 \sum_{h=1}^{H} W_h^2 \Big(\frac{n_h - 1}{n_h}\Big) \frac{s^2_{yh}}{n_h},$$

where $W_h = N_h/N$ are the stratum weights and s^2_{yh} are the stratum sample variance. Even under stratified random sampling, there is no obvious way to directly rescale $V_*\big(\hat{T}^*_{y\mathrm{HT}}\big)$ to make it a consistent variance estimator. The scaling issue for the with-replacement bootstrap under stratified random sampling was recognized by Efron (1982) and Bickel and Freedman (1984). Rao and Wu (1988) proposed an ingenious solution to this problem, which provides valid variance estimation for cases with bounded n_h and non-negligible stratum sampling fractions. It is popularly called the Rao-Wu rescaling bootstrap and is widely used by statistical agencies and survey organizations.

10.1.3 Multi-Stage Cluster Sampling

Two-stage or multi-stage cluster sampling designs generally impose difficulties for variance estimation. We consider multi-stage cluster sampling with first stage clusters selected by PPS sampling without replacement with negligible sampling fractions. For variance estimation, we treat the design as if the first stage clusters are selected by PPS sampling with replacement as discussed in Example 4.5.

The population consists of K first-stage clusters with T_i as the population total for the ith cluster. Let $\mathbf{S}_{1c}$ be the set of k first stage clusters selected in the sample and $d_{1i} = 1/(kz_i)$ be the first stage survey weights, where z_i are the values of the normalized size variable for the PPS sampling of first stage clusters. The subscript "1" in $\mathbf{S}_{1c}$ and d_{1i} indicates "first stage". Let $\hat{T}_i$ be a design-unbiased estimator of the cluster total T_i using data from the second and higher stage samples selected within the ith first stage cluster. The overall sample data can be summarized by

$\{(\hat{T}_i, d_{1i}),\ i \in \mathbf{S}_{1c}\}$. The Horvitz-Thompson estimator of the population total T_y is given by $\hat{T}_{y\text{HT}} = \sum_{i \in \mathbf{S}_{1c}} d_{1i} \hat{T}_i$. The with-replacement bootstrap procedures for multi-stage cluster sampling are as follows.

1. Select a bootstrap sample $\mathbf{S}^*_{1c}$ of size k from the original sample $\mathbf{S}_{1c}$ using simple random sampling with replacement; denote the bootstrap sample data by $\{(\hat{T}^*_i, d^*_{1i}), i \in \mathbf{S}^*_{1c}\}$.
2. Compute the bootstrap version of $\hat{T}_{y\text{HT}}$ using the bootstrap sample data: $\hat{T}^*_{y\text{HT}} = \sum_{i \in \mathbf{S}^*_{1c}} d^*_{1i} \hat{T}^*_i$.
3. Repeat Steps 1 and 2 a large number B times, independently, to obtain $\hat{T}^{*(b)}_{y\text{HT}}$, $b = 1, \cdots, B$.

It can be shown by using the results from Example 4.5 and Problem 10.1 that the bootstrap variance estimator given in (10.2) is consistent if $(k-1)/k \doteq 1$. If k is small, the bootstrap variance estimator $v_{\text{boot}}\big(\hat{T}_{y\text{HT}}\big)$ should be adjusted by multiplying the factor $k/(k-1)$.

Multi-stage cluster sampling designs with large sampling fractions for first stage clusters are difficult to handle by bootstrap methods. Further discussions and solutions can be found in Rao and Wu (1988), Rao et al. (1992), and Sitter (1992a,b).

The bootstrap methods are applicable to parameters that can be expressed as functions of several population totals. A simple example is the Hájek estimator of the population mean, $\hat{\mu}_{y\text{H}} = \sum_{i \in \mathbf{S}} d_i y_i / \sum_{i \in \mathbf{S}} d_i$, which is a ratio of two Horvitz-Thompson estimators. For each bootstrap sample $\mathbf{S}^*$, the bootstrap version of $\hat{\mu}_{y\text{H}}$ is computed as $\hat{\mu}^*_{y\text{H}} = \sum_{i \in \mathbf{S}^*} d^*_i y^*_i / \sum_{i \in \mathbf{S}^*} d^*_i$. The bootstrap variance estimator $v_{\text{boot}}\big(\hat{\mu}_{y\text{H}}\big)$ can be computed in the same way as (10.2). The with-replacement bootstrap methods can also be extended to stratified multi-stage sampling with unequal probability selection at different stages under conditions on the sample sizes and sampling fractions as specified in Sects. 10.1.1, 10.1.2 and 10.1.3 on the three commonly used survey designs. Saigo (2010) provided comparisons among four bootstrap methods for stratified three-stage sampling designs.

10.2 The Pseudo-Population Bootstrap

The with-replacement bootstrap follows directly from the original bootstrap proposed by Efron (1979) for independent and identically distributed sample data, where the resampling process mimics the initial sample selection from an assumed infinite population. In survey practice, however, samples are typically selected without replacement to avoid units being included in the sample more than once. Theoretical results on with-replacement sampling procedures are for the purpose of comparisons and approximations for variance estimation under without-replacement survey designs.

It is natural to develop *without-replacement bootstrap* methods to mimic the survey sampling process from finite populations. The basic idea is to first create a pseudo-population using the observed sample data and then select a bootstrap sample from the pseudo-population using the original survey design. This is also called the *pseudo-population bootstrap*. Consider the simple scenario where the survey sample $\mathbf{S}$ of size n is selected from the population of size N by SRSWOR. Let $\{y_1, \cdots, y_n\}$ be the values of y in the sample. To illustrate the basic idea, we focus on estimating the variance of $\hat{T}_{y\text{HT}} = N\bar{y}$, where $\bar{y}$ is the sample mean. Assuming that $N = kn$ for some integer k, the without-replacement bootstrap method proposed by Gross (1980) consists of the following steps:

1. Construct a pseudo-population of size N by replicating the sample k times.
2. Draw a bootstrap sample of size n from the pseudo-population using SRSWOR to get $\{y_1^*, \cdots, y_n^*\}$. Compute the bootstrap version of $\hat{T}_{y\text{HT}}$ as $\hat{T}_{y\text{HT}}^* = N\bar{y}^*$, where $\bar{y}^*$ is the mean of the bootstrap sample.
3. Repeat Step 2 a large number B times, independently, to obtain $\hat{T}_{y\text{HT}}^{*(b)}$, $b = 1, \cdots, B$.

It is shown in Problem 10.3 that $V_*\big(\hat{T}_{y\text{HT}}^*\big) = N^2\{(n-1)/(n-f)\}(1-f)s_y^2/n$, where s_y^2 is the sample variance and $f = n/N$ is the sampling fraction. It follows that the bootstrap variance estimator computed by (10.2) is valid when the sample size n is not too small, i.e., $(n-1)/(n-f) \doteq 1$, irrespective of the sampling fractions $f \in (0, 1)$.

For more general cases where $k = N/n$ is not an integer, Step 1 on constructing the pseudo-population of size N can be modified by replicating the sample $\lfloor k \rfloor$ times and supplemented by a sample of size $N - n\lfloor k \rfloor$ $(< n)$ selected from the original sample using SRSWOR, where $\lfloor k \rfloor$ denotes the integer part of k. The without-replacement bootstrap variance estimator remains valid (Booth et al. 1994).

Let θ_N be a general finite population parameter and $\hat{\theta}$ be a survey weighted estimator using data $\{(y_i, \mathbf{x}_i, d_i), i \in \mathbf{S}\}$ collected by a single-stage unequal probability sampling design. Recall that the survey weight d_i can be interpreted as "the number of units in the survey population which are represented by unit i selected for the survey sample". The pseudo-population can be constructed using the survey weights and the pseudo-population bootstrap variance estimator for $\hat{\theta}$ can be computed by mimicking the original sample design. This has been discussed by Holmberg (1998), Chauvet (2007) and Wang and Thompson (2012). The following procedures are presented by Chen et al. (2019).

1. Replicate the triplet $(y_i, \mathbf{x}_i, d_i)$ a total of $\lfloor d_i \rfloor$ times for all i in $\mathbf{S}$ to create $\mathbf{U}_{\text{fix}}$, the fixed part of the pseudo-population.
2. Create $\mathbf{U}_{\text{sup}}$, the supplementary part of the pseudo-population by drawing the required number $(< n)$ of units from $\{(y_i, \mathbf{x}_i, d_i), i \in \mathbf{S}\}$ using the original sampling design that selected $\mathbf{S}$, leading to the completed pseudo-population $\mathbf{U}^* = \mathbf{U}_{\text{fix}} \cup \mathbf{U}_{\text{sup}}$ with size N.
3. Take a bootstrap sample $\mathbf{S}^*$ of size n from $\mathbf{U}^*$ using the same sampling design that selected $\mathbf{S}$ from the survey population $\mathbf{U}$.

4. Compute the bootstrap version of the estimator $\hat{\theta}$, denoted as $\hat{\theta}^*$, using the bootstrap sample $\mathbf{S}^*$.
5. Repeat Steps 3 and 4 a large number B times to obtain $\hat{\theta}^{*(b)}$, $b = 1, \cdots, B$. Define $v(\hat{\theta}) = (B-1)^{-1} \sum_{b=1}^{B} \left(\hat{\theta}^{*(b)} - \bar{\theta}^*\right)^2$, where $\bar{\theta}^* = B^{-1} \sum_{b=1}^{B} \hat{\theta}^{*(b)}$.
6. Repeat Steps 2–5 to obtain J independent copies of $v(\hat{\theta})$, denoted as v_j: $j = 1, \cdots, J$. Compute the final bootstrap variance estimator as $v_{\text{boot}}(\hat{\theta}) = J^{-1} \sum_{j=1}^{J} v_j$.

The last step is to mitigate the impact of random selections for the supplementary part of the pseudo-population. Step 2 presented above is different from the method proposed by Wang and Thompson (2012), which includes unit i in the supplementary part of the pseudo-population with probability $r_i = d_i - \lfloor d_i \rfloor$, leading to a random size for $\mathbf{U}_{\text{sup}}$. Chauvet (2007) and Chen et al. (2019) showed that the pseudo-population bootstrap variance estimator $v_{\text{boot}}(\hat{\theta})$ is consistent if the original sample $\mathbf{S}$ is selected by Poisson sampling or a high-entropy fixed sample size PPS sampling design. Extension of the pseudo-population bootstrap to stratified sampling is straightforward but it is less obvious for multi-stage cluster sampling. Wang and Thompson (2012) described an extension to two-stage cluster sampling.

The idea of mimicking the original sampling method in selecting bootstrap samples can also be implemented using the *mirror-match bootstrap* of Sitter (1992a). For single-stage sampling designs with sampling fraction $f = n/N$, a bootstrap sample consists of $k = \lfloor 1/f \rfloor$ independent samples, each has the size $m = \lfloor nf \rfloor$ and is selected from the original sample $\mathbf{S}$ using the same sampling method that selects $\mathbf{S}$ from the survey population. A supplementary part may also be needed to make the size of the bootstrap sample as n. The method was initially proposed for stratified sampling and works well for scenarios where the (stratum) sampling fraction f and the sample size n satisfy $nf \geq 1$. Sitter (1992a) also proposed a modified version that works for scenarios with $nf < 1$. A simplified version of the mirror-match bootstrap was presented in Saigo et al. (2001) using the repeated half-sample bootstrap.

10.3 The Multiplier Bootstrap

The with-replacement bootstrap and the pseudo-population bootstrap are developed with the main focus on variance estimation for the Horvitz-Thompson estimator. The methods are applicable to parameters which are functions of several population totals. A more general class of finite population parameters is defined as the solution to the census estimating equations (7.7) as discussed in Sect. 7.1.2, where the vector-valued estimator $\hat{\theta}$ solves the survey weighted estimating equations (7.8). Direct applications of bootstrap methods for estimating the variance-covariance matrix of $\hat{\theta}$ would require solving (7.8) repeatedly for each bootstrap sample. An alternative approach is to use the results of Theorem 7.1 and estimate the variance-covariance of $\hat{\theta}$ through the variance-covariance of $\mathbf{G}_n(\theta_N) = N^{-1}\sum_{i\in\mathbf{S}} d_i\mathbf{g}_i(\theta_N)$, where $\mathbf{g}_i(\theta) = \mathbf{g}(y_i, \mathbf{x}_i; \theta)$. This is the so-called *estimating function bootstrap* (Hu and Kalbfleisch 2000). Noting that $\mathbf{G}_n(\theta_N)$ is also a Horvitz-Thompson estimator, the results from Sects. 10.1 and 10.2 can be applied. In this section, we present an alternative estimating function bootstrap method which takes advantage of the unique features of estimating equations.

10.3.1 The Multiplier Bootstrap with Estimating Functions

Consider a single-stage unequal probability sampling design with negligible sampling fractions. A valid variance-covariance estimator $v_p\big(\mathbf{G}_n(\theta_N)\big)$ is given in Part (c) of Problem 7.1. Let n be the sample size. Without loss of generality, we drop the factor N^{-1} and let $\mathbf{G}_n(\theta) = \sum_{i\in\mathbf{S}} d_i\mathbf{g}_i(\theta)$ in this section. The multiplier bootstrap consists of the following three steps (Zhao et al. 2020b):

1. Generate $(\omega_1^*, \cdots, \omega_n^*)$ as an independent and identically distributed sample from a probability distribution with mean $\mu = 1$ and variance $\sigma^2 = 1$.
2. Compute the bootstrap version of $\mathbf{G}_n(\theta_N)$ as $\mathbf{G}_n^* = \sum_{i\in\mathbf{S}} \omega_i^* d_i\mathbf{g}_i(\hat{\theta})$.
3. Repeat Steps 1 and 2 a large number B times, independently, to obtain $\mathbf{G}_n^{*(b)}$, $b = 1, \cdots, B$.

One can view $\omega_i^* d_i$ as the bootstrap weight which is obtained by multiplying the initial weight d_i by a randomly generated ω_i^*. The multiplier bootstrap estimator of the variance-covariance matrix $V_p\big\{\mathbf{G}_n(\theta_N)\big\}$ is computed as

$$v_{\text{boot}}\big\{\mathbf{G}_n(\theta_N)\big\} = \frac{1}{B-1}\sum_{b=1}^{B}\Big(\mathbf{G}_n^{*(b)} - \bar{\mathbf{G}}_n^*\Big)\Big(\mathbf{G}_n^{*(b)} - \bar{\mathbf{G}}_n^*\Big)',$$

where $\bar{\mathbf{G}}_n^* = B^{-1}\sum_{b=1}^{B}\mathbf{G}_n^{*(b)}$, which can be replaced by $\mathbf{G}_n(\hat{\theta}) = \mathbf{0}$. Noting that $E_*(\omega_i^*) = 1$ and $V_*(\omega_i^*) = 1$, we have $E_*\big(\mathbf{G}_n^*\big) = \mathbf{G}_n(\hat{\theta}) = \mathbf{0}$ and

$$V_*\big(\mathbf{G}_n^*\big) = \sum_{i\in\mathbf{S}} d_i^2\big\{\mathbf{g}_i(\hat{\theta})\big\}\big\{\mathbf{g}_i(\hat{\theta})\big\}',$$

which is the estimator for the variance-covariance matrix $V_p\big\{\mathbf{G}_n(\theta_N)\big\}$ given in Part (c) of Problem 7.1 under PPS sampling with replacement.

The multiplier bootstrap can be extended to stratified sampling where $\mathbf{G}_n(\theta) = \sum_{h=1}^{H}\sum_{i\in\mathbf{S}_h} d_{hi}\mathbf{g}_{hi}(\theta)$. The bootstrap weights are given by $\omega_{hi}^* d_{hi}$ where the ω_{hi}^* are generated as independent and identically distributed random variables with mean 1 and variance 1. The multiplier bootstrap variance estimator is valid when the stratum sampling fractions are small. The method can also be extended to multi-stage sampling with the first stage cluster sampling designs having negligible sampling fractions.

The multiplier bootstrap does not provide valid variance estimator for arbitrary Horvitz-Thompson estimators, and it does not seem to add any practical value since the scenarios discussed above all have estimators available for the variance-covariance matrix. The method, however, can find valuable applications for non-standard scenarios as shown in the next section.

10.3.2 Variance Estimation in the Presence of Nuisance Functionals

The *Gini coefficient* and the *Lorenz curve* are two important measures of inequalities in income and wealth. Let Y be the non-negative random variable representing income with cumulative distribution function $F(y)$ and mean $\mu_0 = E(Y)$. Let Y_1 and Y_2 be the income of two randomly selected individuals. The Gini coefficient is defined as $\theta = E|Y_1 - Y_2|/(2\mu_0)$, which represents the normalized mean difference of income for the two individuals. The Gini coefficient can be alternatively expressed as (David 1968)

$$\theta = \frac{1}{\mu_0}\int_0^\infty \big\{2F(y) - 1\big\} y dF(y)\,.$$

Let $t(\tau) = F^{-1}(\tau)$ be the population quantile for $\tau \in (0, 1)$. The Lorenz curve (Lorenz 1905; Gastwirth 1971) based on the income distribution $F(y)$ is defined as

$$\theta(\tau) = \frac{1}{\mu_0}\int_0^{t(\tau)} y dF(y)\,, \qquad \tau \in (0, 1)\,.$$

The curve provides a graphical representation of the distribution of income in the population. The finite population Gini coefficient θ_N is the solution to

$$G_N(\theta) = \sum_{i=1}^N g(y_i, \theta, F_N(y_i)) = 0\,,$$

where $g(y, \theta, F_N(y)) = \{2F_N(y) - 1\}y - \theta y$ and $F_N(y)$ is the finite population distribution function. The finite population Lorenz curve θ at a given $\tau \in (0, 1)$ is the solution to

$$G_N(\theta) = \sum_{i=1}^N g(y_i, \theta, t_N(\tau)) = 0\,,$$

where $g(y, \theta, t) = y\{I(y \leq t) - \theta\}$ and $t_N(\tau) = F_N^{-1}(\tau) = \inf\{t \mid F_N(t) \geq \tau\}$ is the finite population quantile.

The two examples have one common feature: the estimating function g involves a nuisance functional. It is the finite population distribution function $F_N(y)$ for the Gini coefficient and the finite population quantile function for the Lorenz curve. Unlike nuisance parameters for which profile analysis is often used, nuisance functionals are typically handled by a plug-in estimator. Let $\mathbf{g}(y, \mathbf{x}, \theta; \psi)$ be the estimating function for defining θ, where ψ is a nuisance function. The vector of the finite population parameters θ_N is the solution to

$$G_N(\theta) = \sum_{i=1}^{N} \mathbf{g}(y_i, \mathbf{x}_i; \theta, \psi_N) = \mathbf{0}\,,$$

where ψ_N is the finite population version of ψ defined through another census estimating equation. Let $\hat{\psi}$ be a survey weighted estimator of ψ_N. The estimator $\hat{\theta}$ for the parameter of interest, θ_N, is the solution to

$$G_n(\theta) = \sum_{i \in \mathbf{S}} d_i \mathbf{g}(y_i, \mathbf{x}_i; \theta, \hat{\psi}) = \mathbf{0}\,.$$

We assume that θ is $k \times 1$ and $\mathbf{g}$ is $r \times 1$ and $r = k$. The over-identified cases with $r > k$ and other technical details are discussed in Zhao et al. (2020b). It can be shown that the asymptotic variance-covariance matrix of $\hat{\theta}$ has the same structure as the one given in (7.9):

$$\big\{\mathbf{H}_N(\theta_N)\big\}^{-1} V_p\big\{\mathbf{G}_n(\theta_N)\big\}\big\{\mathbf{H}_N'(\theta_N)\big\}^{-1}\,,$$

where $\mathbf{H}_N(\theta) = \sum_{i=1}^N \partial\mathbf{g}(y_i, \mathbf{x}_i; \theta, \psi_N)/\partial\theta$, assuming $\mathbf{g}$ is smooth in θ. Note that the factor N^{-1} is dropped from all three terms in (7.9).

The most crucial difference under the current setting is that $\mathbf{G}_n(\theta)$ contains a plug-in estimator $\hat{\psi}$ for the nuisance functional. It is shown by Zhao et al. (2020b) that the linearization method is very difficult to use to estimate $V_p\big\{\mathbf{G}_n(\theta_N)\big\}$. The multiplier bootstrap method described in Sect. 10.3.1 can be used to handle the extra variation induced by the uncertainty in the estimation of ψ by $\hat{\psi}$ through a simple modification to Step 2. With the $(\omega_1^*, \cdots, \omega_n^*)$ generated from Step 1, first compute the bootstrap version of the plug-in estimator $\hat{\psi}^*$ using bootstrap weights $(\omega_1^* d_1, \cdots, \omega_n^* d_n)$, and then compute the bootstrap version of $\mathbf{G}_n(\theta_N)$ using the bootstrap weight and the plug-in estimator $\hat{\psi}^*$. Repeat Step 1 and the modified Step 2 independently to obtain bootstrap copies $\mathbf{G}_n^{*(b)}$, $b = 1, \cdots, B$. The variance $V_p\big\{\mathbf{G}_n(\theta_N)\big\}$ can then be estimated using the standard bootstrap variance formula.

10.4 Replication Weights for Public-Use Survey Data

Analysis of survey data requires detailed design information such as stratification and clustering indicator variables and sample inclusion probabilities. In practice, public-use survey data are created and released to users and such datasets often report only the variables of interest and the final survey weights $\{w_i, i \in \mathbf{S}\}$ obtained by adjusting for unit nonresponse and calibration on auxiliary variables selected by the producer of the data when the population means or totals of those variables are known. Furthermore, the data file provides B columns of replication weights $\{w_i^{(b)}, i \in \mathbf{S}\}$, $b = 1, \cdots, B$ designed for variance estimation. Note that issues with item nonresponse and imputation for missing values are not discussed in this section. Table 10.1 shows a typical format of public-use survey data files as seen by the users. The labeling of y or x on variables is arbitrary and is for the purpose of illustration only.

The set of final survey weights is used to compute point estimates and the additional columns of replication weights are used to produce valid variance estimates. For instance, the survey weighted estimate for the population total T_y for the first variable y_1 in the table is computed as $\hat{T}_y = \sum_{i \in \mathbf{S}} w_i y_{i1}$. To compute the estimated variance, each column of the replication weights is used to obtain replicated copies of $\hat{T}_y$:

$$\hat{T}_y^{(b)} = \sum_{i \in \mathbf{S}} w_i^{(b)} y_{i1}, \qquad b = 1, \cdots, B.$$

The replication variance estimate is then computed in the same way as the bootstrap variance and is given by

$$v_{\text{rep}}\big(\hat{T}_y\big) = \frac{1}{B-1} \sum_{b=1}^{B} \Big(\hat{T}_y^{(b)} - \bar{T}_y\Big)^2, \qquad (10.3)$$

where $\bar{T}_y = B^{-1} \sum_{b=1}^{B} \hat{T}_y^{(b)}$. Information required for the computation is contained in the dataset and advanced inferential problems may be handled through survey weighted estimating equations.

Table 10.1 Public-use survey data file with final survey weights and replication weights

i	y_{i1}	y_{i2}	x_{i1}	x_{i2}	x_{i3}	w_i	$w_i^{(1)}$	$\cdots$	$w_i^{(B)}$
1	y_{11}	y_{12}	x_{11}	x_{12}	x_{13}	w_1	$w_1^{(1)}$	$\cdots$	$w_1^{(B)}$
2	y_{21}	y_{22}	x_{21}	x_{22}	x_{23}	w_2	$w_2^{(1)}$	$\cdots$	$w_2^{(B)}$
3	y_{31}	y_{32}	x_{31}	x_{32}	x_{33}	w_3	$w_3^{(1)}$	$\cdots$	$w_3^{(B)}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	y_{n1}	y_{n2}	x_{n1}	x_{n2}	x_{n3}	w_n	$w_n^{(1)}$	$\cdots$	$w_n^{(B)}$

The creation of additional columns of replication weights for valid variance estimation is an important but difficult task for the producer of the data files. It depends on the original survey design and how the final survey weights are produced. We illustrate the process for two standard scenarios.

1. Basic Design Weights and With-Replacement Bootstrap This is the ideal scenario where there is no unit nonresponse and the final survey weights are the basic design weights, i.e., $w_i = d_i = 1/\pi_i$. In addition, it is assumed that the survey design satisfies conditions such that the with-replacement bootstrap is valid for variance estimation. Let $\mathbf{S}^*$ be the bootstrap sample of size n selected from the original sample $\mathbf{S}$ by simple random sampling with replacement. The bootstrap version of $\hat{T}_y$ can be written as

$$\hat{T}_y^* = \sum_{i \in \mathbf{S}^*} w_i^* y_i^* = \sum_{i \in \mathbf{S}} h_i w_i y_i \,,$$

where h_i is the number of times that unit i is selected in the bootstrap sample, $i \in \mathbf{S}$. It is apparent that $0 \le h_i \le n$ and $\sum_{i \in \mathbf{S}} h_i = n$. The $(h_1, \cdots, h_n)$ are also called the bootstrap frequencies and follow the multinomial distribution with parameters $(p_1, \cdots, p_n; n)$ and $p_i = n^{-1}$ for all i under the with-replacement bootstrap procedure. A set of replication weights $(w_1^{(b)}, \cdots, w_n^{(b)})$ can be constructed by letting $w_i^{(b)} = h_i w_i$. The process can then be carried out independently for $b = 1, \cdots, B$.

Each set of bootstrap weights may contain some zero components corresponding to the zero bootstrap frequencies. This is an undesirable scenario in practice and can be avoided by taking the average of J independent sets of bootstrap weights to create a single set of replication weights. The strategy has been used in some of the public-use datasets created by Statistics Canada. The replication variance estimator $v_{\text{rep}}\big(\hat{T}_y\big)$ given in (10.3) needs to be modified by multiplying a correction factor J.

2. Calibration Weighting Using Auxiliary Information Let $\mathbf{d} = (d_1, \cdots, d_n)$ be the basic design weights and $T_{\mathbf{x}}$ be the known population totals of the auxiliary variables $\mathbf{x}$. The final survey weights $\mathbf{w} = (w_1, \cdots, w_n)$ are the calibration weights such that $\mathbf{w}$ minimizes the chisquare distance $\chi^2(\mathbf{w}, \mathbf{d})$ discussed in Sect. 6.1.2 subject to the calibration constraints (6.1), i.e., $\sum_{i \in \mathbf{S}} w_i \mathbf{x}_i = T_{\mathbf{x}}$. The replications weights can be produced through a bootstrap version of the calibration weighting process as follows:

(i) Select a bootstrap sample $\mathbf{S}^*$ of size n from the original sample $\mathbf{S}$ using simple random sampling with replacement. Denote the bootstrap sample data by $\{(y_i^*, \mathbf{x}_i^*, d_i^*), i \in \mathbf{S}^*\}$. Note that some of the units in $\mathbf{S}^*$ may be duplicates of the same unit in $\mathbf{S}$.

(ii) Compute the bootstrap version of the calibration weights $\mathbf{w}^* = (w_1^*, \cdots, w_n^*)$ by minimizing the chisquare distance $\chi^2(\mathbf{w}^*, \mathbf{d}^*)$ between $\mathbf{w}^*$ and $\mathbf{d}^* = (d_1^*, \cdots, d_n^*)$ subject to the bootstrap version of the calibration constraints $\sum_{i \in \mathbf{S}^*} w_i^* \mathbf{x}_i^* = T_{\mathbf{x}}$.

If the control totals $T_{\mathbf{x}}$ are estimated from another survey, and the bootstrap samples are calibrated to those estimated benchmarks, the uncertainty in the benchmarks will not be account for by the replication weights. When the size of the sample used to estimate $T_{\mathbf{X}}$ is very large, the uncertainty in the benchmarks can be ignored.

There are two key observations from the bootstrap version of the calibration weighting process. First, the calibrated weights w_i^* are undefined for units not selected in the bootstrap sample $\mathbf{S}^*$. Second, if $i, j \in \mathbf{S}^*$ are two duplicated units from $\mathbf{S}$, the corresponding calibrated weights w_i^* and w_j^* must be equal according to the general expression of calibration weights given in (6.6). Let h_i be the number of times that unit i is selected in the bootstrap sample, $i \in \mathbf{S}$. The replication weights $(w_1^{(b)}, \cdots, w_n^{(b)})$ can be constructed by letting $w_i^{(b)} = h_i w_i^*$. if w_i^* is undefined, we must have $h_i = 0$ and $w_i^{(b)} = 0$. The process can be repeated independently for $b = 1, \cdots, B$. The issue of zero components in the weights can be mitigated by averaging J independent sets of calibrated bootstrap weights to produce a single set of replication weights, and adjusting the final replication variance estimator by the factor J.

In survey practice, unit nonresponses are handled by deleting the missing units and adjusting the original design weights using methods described in Sect. 9.2 for the final sample. The adjusted weights are typically treated as the design weights for the final sample and used as the starting point for constructing replication weights. Imputation for missing item nonresponses is a more complicated problem and there is no simple solution to the construction of replication weights that can take into account of imputation, especially when the survey data file contains several variables with missing values.

Example 10.1 (Inference with Estimating Equations and Public-Use Data) Consider the survey weighted estimating equations described in Sect. 7.2 and the public-use dataset in the form presented in Table 10.1. The point estimator $\hat{\theta}$ is obtained as the solution to

$$\mathbf{G}_n(\theta) = \sum_{i \in \mathbf{S}} w_i \mathbf{g}(y_i, \mathbf{x}_i; \theta) = \mathbf{0},$$

where the w_i are the final survey weights included in the dataset. The variance-covariance matrix of $\hat{\theta}$ can be estimated using the variance formula given in (7.9), Part (b) of Theorem 7.1 as

$$v_p\big(\hat{\theta}\big) = \big\{\mathbf{H}_n(\hat{\theta})\big\}^{-1} v_p\big\{\mathbf{G}_n(\hat{\theta})\big\}\big\{\mathbf{H}_n'(\hat{\theta})\big\}^{-1},$$

where

$$\mathbf{H}_n(\hat{\theta}) = \Big\{\sum_{i \in \mathbf{S}} w_i \frac{\partial}{\partial \theta} \mathbf{g}(y_i, \mathbf{x}_i; \theta)\Big\}\Big|_{\theta = \hat{\theta}},$$

$$v_p\{\mathbf{G}_n(\hat{\theta})\} = \frac{1}{B-1}\sum_{b=1}^{B}\Big(\mathbf{G}_n^{(b)} - \bar{\mathbf{G}}_n\Big)\Big(\mathbf{G}_n^{(b)} - \bar{\mathbf{G}}_n\Big)',$$

and the B columns of replication weights are used to obtain

$$\mathbf{G}_n^{(b)} = \sum_{i\in\mathbf{S}} w_i^{(b)}\mathbf{g}(y_i, \mathbf{x}_i; \hat{\theta}), \quad b = 1, \cdots, B \quad \text{and} \quad \bar{\mathbf{G}}_n = \frac{1}{B}\sum_{b=1}^{B}\mathbf{G}_n^{(b)}.$$

◇

Example 10.2 (Pseudo Empirical Likelihood Ratio Confidence Intervals) The replication weights are constructed primarily for variance estimation. They do not necessarily provide approximation to the sampling distribution of a test statistic. The with-replacement bootstrap replication weights with or without calibration weighting described in this section are indeed valid for confidence intervals and hypothesis tests if the original survey design has large sample size and small sampling fractions. Consider the pseudo empirical likelihood method discussed in Sect. 8.2. With the public-use survey dataset, the pseudo EL function of (8.5) is re-defined as

$$\ell(\mathbf{p}) = n\sum_{i\in\mathbf{S}}\tilde{w}_i(\mathbf{S})\log(p_i),$$

where $\mathbf{p} = (p_1, \cdots, p_n)$ and $\tilde{w}_i(\mathbf{S}) = w_i/\sum_{j\in\mathbf{S}} w_j$ are the normalized final survey weights. Maximizing $\ell_{\mathrm{WR}}(\mathbf{p})$ subject to $\sum_{i\in\mathbf{S}} p_i = 1$ leads to $\hat{p}_i = \tilde{w}_i(\mathbf{S})$. Let $\theta_N = \mu_y$ be the survey population mean. Let $\hat{p}(\theta)$ be the maximizer of $\ell_{\mathrm{WR}}(\mathbf{p})$ under the normalization constraint $\sum_{i\in\mathbf{S}} p_i = 1$ and the parameter constraint

$$\sum_{i\in\mathbf{S}} p_i(y_i - \theta) = 0 \tag{10.4}$$

for a fixed θ. The pseudo empirical log-likelihood ratio function at the given θ is computed as

$$r(\theta) = -2\big\{\ell(\hat{\mathbf{p}}(\theta)) - \ell(\hat{\mathbf{p}})\big\}.$$

Note that the factor -2 is included here but not in (8.7). We have $r(\theta) \geq 0$ for any θ. The $(1-\alpha)$-level confidence interval for θ_N can be constructed as $\{\theta \mid r(\theta) \leq b_\alpha\}$, where b_α is the upper 100αth-quantile of the sampling distribution of $r(\theta)$ at $\theta = \theta_N$. The value of b_α can be approximated by the quantile of the empirical distribution of $\{r^{(b)}(\hat{\theta}), b = 1, \cdots, B\}$. The $r^{(b)}(\hat{\theta})$ is computed in the same way as $r(\theta)$, with $\tilde{w}_i(\mathbf{S})$ used in $\ell(\mathbf{p})$ replaced by the normalized replication weights $\tilde{w}_i^{(b)}(\mathbf{S}) = w_i^{(b)}/\sum_{j\in\mathbf{S}} w_j^{(b)}$ and θ in (10.4) replaced by $\hat{\theta} = \sum_{i\in\mathbf{S}}\tilde{w}_i(\mathbf{S})y_i$. Wu and Rao (2010) contains a theoretical justification of the validity of the bootstrap calibrated pseudo EL ratio confidence intervals. A similar result can also be found in Tan and Wu (2015). A key feature of the approach presented here is that the scaling constant specified in Theorem 8.1 is no longer required. The approach is readily applicable to more general parameters θ_N defined through estimating functions $\mathbf{g}(y, \mathbf{x}; \theta)$, which amounts to replacing (10.4) by $\sum_{i\in\mathbf{S}} p_i \mathbf{g}(y_i, \mathbf{x}_i; \theta) = \mathbf{0}$. ◇

Replication weights are a key feature of public-use data files. Constructing them is a challenging task for data file producers but the released datasets facilitate statistical inferences for data users. Replication weights also provide an effective way not to disclose confidential information, including potentially sensitive design information such as stratum and cluster indicators (Lu and Sitter 2008).

10.5 An Algebraic Construction of Replication Weights

Public-use survey data files producers usually have access to detailed design information such as second order inclusion probabilities. It is possible to achieve the first major objective of replication weights on valid variance estimation without using any resampling methods. The following idea of algebraic construction of replication weights was first outlined in Fay (1984) and Fay and Dippo (1989) and was later examined in detail by Kim and Wu (2013).

We consider $\hat{T}_{y\text{HT}} = \sum_{i\in\mathbf{S}} d_i y_i$ and assume that π_i and π_{ij} are all available for the final sample $\mathbf{S}$ of size n. The design-unbiased variance estimator is given by

$$v_p\big(\hat{T}_{y\text{HT}}\big) = \sum_{i\in\mathbf{S}}\sum_{j\in\mathbf{S}} \frac{\pi_{ij} - \pi_i\pi_j}{\pi_{ij}\pi_i\pi_j} y_i y_j \,. \tag{10.5}$$

The variance estimator can be written as a quadratic form $v_p\big(\hat{T}_{y\text{HT}}\big) = \mathbf{y}\varDelta\mathbf{y}'$, where $\varDelta = (\varOmega_{ij})$ is an $n \times n$ matrix with $\varOmega_{ij} = (\pi_{ij} - \pi_i\pi_j)/(\pi_{ij}\pi_i\pi_j)$ and $\mathbf{y} = (y_1, \cdots, y_n)'$. Suppose that we want to re-express the variance estimator in the form of

$$v_p\big(\hat{T}_{y\text{HT}}\big) = \sum_{k=1}^{L} c_k\Big(\hat{T}_y^{(k)} - \hat{T}_{y\text{HT}}\Big)^2 , \tag{10.6}$$

where the c_k are pre-specified constants and $\hat{T}_y^{(k)}$ is the kth replicated version of $\hat{T}_{y\text{HT}}$. For the form of the bootstrap variance estimator given in (10.2), we would have $c_k = (B-1)^{-1}$, $L = B$ and $\bar{T}^*_{y\text{HT}}$ replaced by $\hat{T}_{y\text{HT}}$. The matrix $\varDelta$ is nonnegative definite and has spectral decomposition

$$\varDelta = \sum_{k=1}^{L_0} \lambda_k \mathbf{e}_k \mathbf{e}_k' \,,$$

where the $\lambda_k > 0$ are the eigenvalues associated with the eigenvectors $\mathbf{e}_k$ and L_0 $(\leq n)$ is the number of non-zero eigenvalues. Let $\mathbf{w}^{(k)} = (w_1^{(k)}, \cdots, w_n^{(k)})'$ be the kth set of replication weights and $\mathbf{d} = (d_1, \cdots, d_n)'$ be the set of basic design weights. If we let $L = L_0$ and

$$\mathbf{w}^{(k)} = \mathbf{d} + \big(\lambda_k / c_k\big)^{1/2}\mathbf{e}_k\,, \quad k = 1, \cdots, L\,,$$

then the replication variance estimator given by (10.6) becomes identical to the design-unbiased variance estimator given in (10.5). The replication method therefore is fully efficient.

The first major issue for the algebraic construction method is the eigenvalue-eigenvector decomposition of the $n \times n$ matrix Δ. Most regular computers can handle $n = 5000$ but samples with very large sizes may need more powerful computing machines. Noting that $L_0 = \text{rank}(\Delta)$ which is very close to n, the second major issue is how to reduce the number of sets of replication weights when n is very large. Kim and Wu (2013) proposed to take a random sample of a smaller B sets of replication weights from the initial L_0 constructed sets. The replication variance estimator using the selected B sets of replication weights, multiplying by a scaling factor L/B, remains consistent. Kim and Wu (2013) further proposed a calibration method to improve the efficiency of the replication variance estimator based on the smaller number of replication weights.

When the exact second order inclusion probabilities cannot be computed but the survey design allows for approximate variance estimation such as the Hájek variance estimator discussed in Sect. 4.2.2, the algebraic construction method is still applicable, as long as the approximate variance estimator can be expressed in a quadratic form and the matrix Δ can be clearly defined. If the final survey weights w_i are the calibration weights, the variance of the calibration estimator $\hat{T}_{y\text{C}} = \sum_{i \in \mathbf{S}} w_i y_i$ is asymptotically equivalent to the variance of the Horvitz-Thompson estimator for the residual variable $r_i = y_i - \mathbf{x}_i'\hat{\mathbf{B}}_\text{C}$ (Sect. 6.1.2). It can be shown that the variance estimator $v_p\big(\hat{T}_{y\text{C}}\big)$ can be written in a quadratic form, and consequently replication weights can be constructed using the same method based on the spectral decomposition (Problem 10.6).

10.6 Bootstrap Methods for Imputed Survey Data

Variance estimation with imputed survey data does not follow standard procedures for complete data. This is apparent from discussions presented in Sect. 9.6. Replication weights developed for complete data do not produce valid variance estimators when they are naively applied to imputed datasets. The naive replication method underestimates the variance since the extra variation induced by imputation for missing values is not properly taken into account.

Bootstrap methods for complete survey data can be modified to produce valid variance estimators for imputed datasets. The key step for the modified bootstrap method is to re-impute the missing values in the bootstrap sample using the same imputation method which was used to produce the original imputed dataset (Efron 1994; Shao and Sitter 1996). Suppose that the initial sample $\mathbf{S}$ is selected by single-stage PPS sampling and the sampling fraction n/N is small. The with-replacement bootstrap procedure described in Sect. 10.1.1 is valid if there are no missing values. Suppose that the variable y is subject to missingness and the missing y in the final dataset is imputed by the random regression imputation method discussed in Sect. 9.4.1 (Scenario III). Using the same notation from Chap. 9 but changing the imputed value y_i^* to y_i° to avoid notational confusions with the bootstrap sample, we have $\mathbf{S} = \mathbf{S}_R \cup \mathbf{S}_M$ and the imputed dataset is denoted as $\{(\tilde{y}_i, \mathbf{x}_i, \delta_i, d_i), i \in \mathbf{S}\}$, where $\tilde{y}_i = \delta_i y_i + (1 - \delta_i) y_i^\circ$ and $y_i^\circ = \mathbf{x}_i' \hat{\beta} + \hat{\varepsilon}_j$ is the imputed value. Let $\tilde{T}_{y\mathrm{HT}} = \sum_{i \in \mathbf{S}} d_i \tilde{y}_i$ be the Horvitz-Thompson estimator of T_y based on the imputed dataset. The following modified bootstrap method can be used to estimate the variance of $\tilde{T}_{y\mathrm{HT}}$.

1. Select a bootstrap sample $\mathbf{S}^*$ of size n from the original sample $\mathbf{S}$ using simple random sampling with replacement; denote the bootstrap sample data by $\{(\tilde{y}_i^*, \mathbf{x}_i^*, \delta_i^*, d_i^*), i \in \mathbf{S}^*\}$.
2. Let $\mathbf{S}^* = \mathbf{S}_R^* \cup \mathbf{S}_M^*$, where $\mathbf{S}_R^*$ consists of i with $\delta_i^* = 1$ and $\mathbf{S}_M^*$ includes all i with $\delta_i^* = 0$. Compute the estimated regression coefficients $\hat{\beta}^*$ using the bootstrap data $\{(\tilde{y}_i^*, \mathbf{x}_i^*, \delta_i^*, d_i^*), i \in \mathbf{S}_R^*\}$ from the respondents and obtain the residuals $\{\hat{\varepsilon}_i^*, i \in \mathbf{S}_R^*\}$; Re-impute $\tilde{y}_i^*$ for $i \in \mathbf{S}_M^*$ using the random regression imputation method to get $\tilde{y}_i^* = y_i^{\circ *} = \mathbf{x}_i^* \hat{\beta}^* + \hat{\varepsilon}_j^*$.
3. Compute the bootstrap version of $\tilde{T}_{y\mathrm{HT}}$ using the re-imputed bootstrap sample data: $\tilde{T}_{y\mathrm{HT}}^* = \sum_{i \in \mathbf{S}^*} d_i^* \tilde{y}_i^*$.
4. Repeat Steps 1–3 a large number B times, independently, to obtain $\tilde{T}_{y\mathrm{HT}}^{*(b)}$, $b = 1, \cdots, B$.

The bootstrap variance estimator $v_{\text{boot}}\big(\tilde{T}_{y\mathrm{HT}}\big)$ is computed using the same formula as (10.2) and works well as long as the original bootstrap procedure is valid for complete data (Shao and Sitter 1996). Extensions of the procedures described above to stratified sampling are straightforward when the stratum sample sizes are large and sampling fractions are small.

The with-replacement bootstrap procedure tends to overestimate the variance when the sampling fractions are not small. This is also true when the procedure is modified for imputed survey datasets. The following modified pseudo-population (without replacement) bootstrap provides valid variance estimates for $\tilde{\theta}$, the estimator of θ based on the imputed dataset, and works well for any sampling fractions (Chen et al., 2019). Once again, we assume that the random regression imputation method is used to create the original imputed dataset $\{(\tilde{y}_i, \mathbf{x}_i, \delta_i, d_i), i \in \mathbf{S}\}$.

1. Replicate the quadruplet $(\tilde{y}_i, \mathbf{x}_i, \delta_i, d_i)$ a total of $\lfloor d_i \rfloor$ times for all i in $\mathbf{S}$ to create $\mathbf{U}_{\text{fix}}$, the fixed part of the pseudo-population.

2. Create $\mathbf{U}_{\text{sup}}$, the supplementary part of the pseudo-population by drawing the required number ($< n$) of units from $\{(\tilde{y}_i, \mathbf{x}_i, \delta_i, d_i), i \in \mathbf{S}\}$ using the original sampling design that selected $\mathbf{S}$, leading to the completed pseudo-population $\mathbf{U}^* = \mathbf{U}_{\text{fix}} \cup \mathbf{U}_{\text{sup}}$ with size N.
3. Take a bootstrap sample $\mathbf{S}^*$ of size n from $\mathbf{U}^*$ using the same sampling design that selected $\mathbf{S}$ from the survey population $\mathbf{U}$.
4. Let $\mathbf{S}^* = \mathbf{S}_R^* \cup \mathbf{S}_M^*$, where $\mathbf{S}_R^*$ consists of i with $\delta_i^* = 1$ and $\mathbf{S}_M^*$ includes all i with $\delta_i^* = 0$. Compute the estimated regression coefficients $\hat{\beta}^*$ using the bootstrap data $\{(\tilde{y}_i^*, \mathbf{x}_i^*, \delta_i^*, d_i^*), i \in \mathbf{S}_R^*\}$ from the respondents and obtain the residuals $\{\hat{\varepsilon}_i^*, i \in \mathbf{S}_R^*\}$; Re-impute $\tilde{y}_i^*$ for $i \in \mathbf{S}_M^*$ using the random regression imputation method to get $\tilde{y}_i^* = y_i^{\circ *} = \mathbf{x}_i^* \hat{\beta}^* + \hat{\varepsilon}_j^*$.
5. Compute the bootstrap version of the estimator $\tilde{\theta}$, denoted as $\tilde{\theta}^*$, using the bootstrap sample $\mathbf{S}^*$.
6. Repeat Steps 3–5 a large number B times to obtain $\tilde{\theta}^{*(b)}$, $b = 1, \cdots, B$. Define $v(\tilde{\theta}) = (B-1)^{-1} \sum_{b=1}^{B} \left(\tilde{\theta}^{*(b)} - \bar{\theta}^*\right)^2$, where $\bar{\theta}^* = B^{-1} \sum_{b=1}^{B} \tilde{\theta}^{*(b)}$.
7. Repeat Steps 2–6 to obtain J independent copies of $v(\hat{\theta})$, denoted as v_j: $j = 1, \cdots, J$. Compute the final bootstrap variance estimator as $v_{\text{boot}}(\tilde{\theta}) = J^{-1} \sum_{j=1}^{J} v_j$.

A major practical issue of the bootstrap methods for imputed data is the requirement for knowledge of δ_i, i.e., the imputed dataset needs to carry identification flags for imputed values. Practical convenience and privacy concerns for public-use survey data are not in favor of including such flags for imputed datasets but most theoretical developments on variance estimation under imputation for missing values require information on the original response status. This is a scenario where theory and practice collide with respect to what is needed and what can be offered. Creation of replication weights for valid variance estimation with imputed public-use survey data remains an open research topic, and the problem becomes even more complicated when imputation is done for multiple variables with missing values.

If the public-use survey dataset contains many variables but only a few are subject to imputation for missing values, it is still desirable to include replication weights which provide valid variance estimation for complete data. When a particular analysis involves variables with imputed values, it might be possible to use a hybrid method for variance estimation. The regular replication method with the given replication weights in the data file can be used to estimate the design component of the variance, such as $V_p E_\xi \left\{\hat{\mathbf{U}}^*(\theta)\right\}$ in equation (9.16), and the imputation component of the variance, which depends on the missing mechanism and the imputation model, may be estimated through an analytic method. The hybrid method for variance estimation was discussed in a different context by Wu and Sitter (2001b).

10.7 Additional Remarks

Bootstrap variance estimators and confidence intervals are computed through a Monte Carlo simulation process. An important question in practice is the number of bootstrap samples needed in order to have a stable bootstrap variance estimator or confidence interval. While the simple answer is "the more the better", the most commonly accepted choice is $B = 2000$. This is an issue of Monte Carlo simulation errors, and the choice of B can be justified using the example of 95% confidence intervals. Consider a confidence interval $(\hat{L}, \hat{U})$ such that $p = P(\hat{L} < \theta < \hat{U}) = 0.95$. If the coverage probability is computed based on B simulated intervals $(\hat{L}^{(b)}, \hat{U}^{(b)})$, $b = 1, \cdots, B$, we have $\hat{p} = B^{-1}\sum_{b=1}^{B} I(\hat{L}^{(b)} < \theta < \hat{U}^{(b)}) \doteq 0.95$ and the margin of error is controlled approximately by $2\{V(\hat{p})\}^{1/2} = 2\{0.05 \times 0.95/B\}^{1/2}$, which is less than one percentage point for $B = 2000$. Most statistical agencies and survey organizations use $B = 500$ sets of replication weights for public-use survey datasets to reduce the burden of managing excessively large data files, and the margin of error for coverage of 95% confidence intervals is controlled by two percentage points.

The bootstrap was first introduced by Efron (1979) for independent and identically distributed sample data. There exists a large body of literature on bootstrap methods for complex survey data. Many key references can be found in the review paper by Mashreghi et al. (2016) on bootstrap methods for finite population sampling. In addition to bootstrap, two other commonly used resampling methods are jackknife and balanced repeated replications (BRR). The concept of jackknife was first used for bias correction and was much earlier than the advent of the bootstrap (Quenouille 1956). The jackknife methods create systematic subsamples from the original sample and use replicated copies of the estimator based on the subsamples of smaller size to assess the variability. For instance, the delete-1 jackknife creates n subsamples of size $n - 1$ by deleting one unit from the original sample. The jackknife method unfortunately does not work for parameters which are non-smooth functions of population means or totals such as population quantiles. Shao and Tu (1995) contains general theory of the jackknife and its applications to survey data. Rao and Shao (1992) proposed an adjusted jackknife for variance estimation with survey data under hot deck imputation. The BRR method was first proposed by McCarthy (1966, 1969) under stratified sampling where the number of strata is large and the number of primary sampling units (psu) within each stratum is two (i.e., $n_h = 2$). Replicated copies of the estimator can be created based on a number of half-samples formed by deleting one psu from the sample in each stratum. The set of all possible balanced half-samples may be defined by a design matrix. See Krewski and Rao (1981) and Rao and Wu (1985) for further details on jackknife and BRR, and Rao and Shao (1996, 1999) on BRR. The jackknife and the BRR are primarily for variance estimation of non-linear statistics which are smooth functions of population means or totals, to bypass analytic expansions required for the linearization method. The two methods are less useful and less flexible than the bootstrap in terms of dealing with non-smooth statistics or approximations to the sampling distributions of point estimators and test statistics.

10.8 Problems

10.1 (The With-Replacement Bootstrap Under Single-Stage PPS Sampling) Consider the with-replacement bootstrap procedures of McCarthy and Snowden (1985) as described in Sect. 10.1. Show that

$$E_*\big(\hat{T}^*_{y\mathrm{HT}}\big) = \hat{T}_{y\mathrm{HT}} \qquad \text{and} \qquad V_*\big(\hat{T}^*_{y\mathrm{HT}}\big) = \frac{n-1}{n} v\big(\hat{T}_{y\mathrm{HH}}\big)\,,$$

where $E_*(\cdot)$ and $V_*(\cdot)$ denote respectively the expectation and the variance under the bootstrap sampling procedures and $v\big(\hat{T}_{y\mathrm{HH}}\big)$ is the unbiased variance estimator for the Hansen-Hurwitz estimator given in (10.1).

10.2 (The Modified With-Replacement Bootstrap) Suppose that the with-replacement bootstrap procedures of McCarthy and Snowden (1985) as described in Sect. 10.1 are modified by changing the bootstrap sample size to m instead of using the original sample size n. The value of m is chosen such that the bootstrap variance $V_*\big(\hat{T}^*_{y\mathrm{HT}}\big)$ matches exactly the design-unbiased variance estimator $v\big(\hat{T}_{y\mathrm{HT}}\big)$ when the original sample $\mathbf{S}$ is selected by SRSWOR.

(a) Show that the mathematical expression of m is given by $m = n(1-f)/(1-n^{-1})$, where $f = n/N$.

(b) Suppose that m is not an integer. A randomization strategy can be developed as follows: Select a bootstrap sample $\mathbf{S}^*$ of size $\lfloor m \rfloor$ with probability γ or of size $\lfloor m \rfloor + 1$ with probability $1-\gamma$, where $\lfloor m \rfloor$ is the integer part of m, such that the bootstrap variance matches exactly the design-unbiased variance estimator when the original sample $\mathbf{S}$ is selected by SRSWOR. Find the expression for the probability γ.

10.3 (The Without-Replacement Bootstrap of Gross (1980)) Suppose that $\mathbf{S}$ is a sample of size n selected from the population of size N by SRSWOR. Let $\{y_1, \cdots, y_n\}$ be the values of y in the sample. Let $\bar{y}$ and s_y^2 be the sample mean and the sample variance, respectively, and assume that $N = kn$ for some integer k. Show that under the without-replacement bootstrap method of Gross (1980) described in Sect. 10.2,

$$E_*(\bar{y}^*) = \bar{y} \qquad \text{and} \qquad V_*(\bar{y}^*) = \frac{n-1}{n-f}(1-f)\frac{s_y^2}{n}\,,$$

where $E_*(\cdot)$ and $V_*(\cdot)$ denote respectively the expectation and the variance under the bootstrap sampling procedures, and $f = n/N$.

10.4 (Variance Estimation for the Ratio and the Regression Estimators) Conduct a simulation study to evaluate the with-replacement bootstrap variance estimator with comparisons to the linearization variance estimator for the ratio and the regression estimator of the population mean.

(a) Generate the finite population of $N = 5000$ under settings suitable for ratio and regression estimation.
(b) Consider two sampling designs: (i) SRSWOR; and (ii) PPS sampling without replacement.
(c) Choose n such that $n/N = 0.01$ and $n/N = 0.10$.
(d) Assess the performance of the variance estimator using Relative Bias and Mean Squared Errors.

Note: The true values of the variance of the ratio and the regression estimators under a given sampling design can be simulated by a separate set of $B = 2000$ simulated samples.

10.5 (Variance Estimation for the Gini Coefficient and the Lorenz Curve) Conduct a simulation study to evaluate the performance of bootstrap variance estimators for the Gini coefficient and the Lorenz curve at selected quantile levels (i.e., values of τ). The variables for the finite population should include y, a non-negative income variable following a continuous probability distribution and at least one x variable which is related to y and can be used as the size variable for single-stage PPS sampling designs.

(a) Provide detailed expressions for the finite population Gini coefficient and the Lorenz curve and the survey weighted estimators.
(b) Specify the detailed procedures for the with-replacement bootstrap (Sect. 10.1.1) and the multiplier bootstrap (Sect. 10.3.1) under the current setting.
(c) Consider two sampling designs: (i) Simple random sampling without replacement, and (ii) Single-stage PPS sampling without replacement; and conduct the simulation study separately under each design.

10.6 (Construction of Replication Weights for Calibration Estimators) Suppose that the final survey weights w_i are the calibration weights under the chisquare distance measure as discussed in Sect. 6.1.2. Let $\hat{T}_{y\mathrm{C}} = \sum_{i \in \mathbf{S}} w_i y_i$ be the calibration estimator and $v_p\big(\hat{T}_{y\mathrm{C}}\big)$ be the variance estimator.

(a) Show that $v_p\big(\hat{T}_{y\mathrm{C}}\big)$ can be written in a quadratic form $\mathbf{y}\Delta_{\mathrm{C}}\mathbf{y}'$.
(b) Show in detail how to construct replication weights using the spectral decomposition of Δ_{C}.
(c) Argue that the replication weights constructed in (a) and (b) also provide a valid variance estimator when the initial calibration weights are obtained using other distance measures such as those discussed in Sects. 6.2.1 and 6.2.2.

10.7 (Bootstrap Variance Estimation with Imputed Survey Data) Conduct a simulation study to evaluate the performance of bootstrap variance estimators under imputation for missing values. The finite population of size $N = 5000$ is generated from the regression model $y_i = 1 + x_{1i} + x_{2i} + x_{3i} + \varepsilon_i$, where $x_{1i} \sim 0.5 + \text{Exp}(1)$, $x_{2i} \sim \chi^2(1)$, $x_{3i} \sim$ Bernoulli(0.5) and $\varepsilon_i \sim N(0, \sigma^2)$, all independent of each other. Consider a single-stage randomized systematic PPS sampling method with $\pi_i \propto x_{1i}$. Conduct two separate simulation studies with sample sizes $n = 100$ and 500, corresponding to $f = n/N = 2\%$ and 10%.

(a) Choose σ^2 such that the finite population correlation coefficient between y_i and the linear predictor $\eta_i = 1 + x_{1i} + x_{2i} + x_{3i}$ is 0.70.
(b) Describe how to generate the response status indicator variables δ_i for the study variable y_i under the MAR assumption such that the average missing rate is controlled at 0.30.
(c) Describe in detail the random regression imputation method.
(d) Describe in detail the three bootstrap methods for imputed survey data: (i) The naive with-replacement bootstrap; (ii) The modified with-replacement bootstrap; and (iii) The modified pseudo-population bootstrap.

Evaluate the performances of the three bootstrap variance estimators using the coverage probabilities of the 95% confidence intervals for the population mean μ_y based on the normal approximation to $(\tilde{\mu}_{y\text{HT}} - \mu_y)/\{v_{\text{boot}}(\tilde{\mu}_{y\text{HT}})\}^{1/2}$, where $\tilde{\mu}_{y\text{HT}}$ is the Horvitz-Thompson estimator of μ_y based on the imputed sample data.

Chapter 11
Bayesian Empirical Likelihood Methods

Bayesian inference involves three main phases: the formulation of a likelihood function for the observed sample data, the choice of a prior distribution for the parameter of interest, and the inferential procedures based on the posterior distribution of the parameter given the observed sample data. Bayesian methods for inference on finite population parameters based on complex survey data face hurdles in all three phases. The difficulties of formulating a likelihood function with complex survey data were documented by Rao and Wu (2009), and the choices of a prior distribution for finite population parameters are conceptually challenging. Moreover, posterior inferences with valid frequentist interpretation under the design-based framework turn out to be even harder to achieve.

In this chapter, we first provide a brief review of Bayesian approaches to finite population inference. We then present Bayesian empirical likelihood methods for the finite population mean as well as general parameters defined through estimating functions. The focus is on how to formulate the inferential procedures under the Bayesian framework, and discussions of design-based frequentist properties of Bayesian point estimators and credible intervals. Some of the technical details on the pseudo empirical likelihood or the sample empirical likelihood can be found in Chap. 8.

11.1 Bayesian Inference with Survey Data

Probability sampling designs and design-based inference, where the randomization is under repeated sampling from a fixed finite population, are widely accepted as the standard approach in sample surveys (Hansen et al. 1983). The design-based approach has played a dominant role, especially in the production of official statistics, and has had profound impact on survey sampling research and practice (Rao 2011). Design-based inference can be viewed as nonparametric and has the

C. Wu, M. E. Thompson, *Sampling Theory and Practice*, ICSA Book
Series in Statistics, https://doi.org/10.1007/978-3-030-44246-0_11

ability to handle complex survey design features. Design-based point estimators and the associated design-consistent variance estimators are used to construct confidence intervals or conduct statistical hypotheses based on normal approximations. Such intervals or tests are asymptotically valid "whatever the unknown properties of the population" (Neyman 1934).

Parametric Bayesian inference requires specification of a parametric model for the sample data. If the sampling probabilities have no dependence on the variables in the sample data, the Bayes posterior distribution of the population mean corresponding to any prior is independent of the sampling design, as shown by Godambe (1966, 1968) and Ericson (1969). Royall and Pfeffermann (1982) studied Bayesian inference for the finite population mean assuming normality of the residuals and using flat priors on the parameters of a linear regression model. Results on the posterior mean and the posterior variance of the finite population mean are similar to those of the model-based method of Royall (1970). The parametric Bayesian approach has found useful applications in the context of small area estimation where the sample data are sparse at certain domain levels and strong model assumptions become necessary (Rao and Molina 2015).

A nonparametric Bayesian approach seems to be more attractive for estimation of finite population parameters, but it requires the specification of a nonparametric likelihood function based on the full sample data $\{y_i, i \in \mathbf{S}\}$ and a prior distribution on the vector of population parameters $\tilde{\mathbf{y}} = (\tilde{y}_1, \tilde{y}_2, \cdots, \tilde{y}_N)'$. The nonparametric likelihood induced by the probability sampling design is the "flat" Godambe likelihood $L(\tilde{\mathbf{y}})$ defined at the beginning of Chap. 8, and it is noninformative for the population parameters $\tilde{\mathbf{y}}$. One way out of this difficulty is to assume an informative (exchangeable) prior on the N-dimensional parameter vector and combine it with the noninformative likelihood (Ericson 1969) to get an informative posterior, but Bayesian inferences based on the posterior distribution do not depend on the sampling design. Meeden and Vardeman (1991) proposed a simulation based noninformative Bayesian approach to interval estimation, assuming that "the unseen are like the seen", which is equivalent to the exchangeability assumption. The approach leads to a Polya urn distribution as a pseudo-posterior distribution over the unobserved population values y_i. The Polya posterior is based on a stepwise Bayesian argument and it is not a true Bayesian posterior based on a single prior. The exchangeability assumption prevents possible extensions of the method to complex survey designs with valid frequentist interpretations under the design-based framework.

The scale-load approach of Hartley and Rao (1968) provides a non-trivial likelihood function similar to the empirical likelihood of Owen (1988) under simple random sampling. Hartley and Rao (1968) studied Bayesian inference using a noninformative compound-multinomial prior (Hoadley 1969) with the scale-load likelihood and obtained a posterior mean and posterior variance similar to Ericson's (1969). The scale-load approach is promising but has limited applicability because the scale-load likelihoods cannot be obtained easily for complex sampling designs. Bayesian inference using the pseudo empirical likelihood, which is discussed in the next section, is applicable to general survey designs with desirable frequentist properties. It provides similar results to the scale-load based approach under simple random sampling.

11.2 Bayesian Inference Based on Pseudo Empirical Likelihood

The pseudo empirical likelihood function discussed in Chap. 8 can be used for Bayesian inferences on the population mean $\theta = \mu_y$ or any functions of θ. We consider general unequal probability sampling designs with the first order inclusion probabilities π_i. Using the notation from Chap. 8, we let $\tilde{d}_i(\mathbf{S}) = d_i / \sum_{j\in\mathbf{S}} d_j$ be the normalized weights where $d_i = 1/\pi_i$ are the basic design weights.

Let $(p_1, \ldots, p_n)$ be a discrete probability measure over the n sampled units in $\mathbf{S}$ satisfying $p_i > 0$ and $\sum_{i\in\mathbf{S}} p_i = 1$. Recall that the pseudo empirical log-likelihood function defined in Sect. 8.2 is given by

$$\ell_{\mathrm{WR}}(\mathbf{p}) = n \sum_{i\in\mathbf{S}} \tilde{d}_i(\mathbf{S}) \log(p_i)\,.$$

We replace the sample size n by the effective sample size n^* which is defined as $n^* = n/\mathrm{deff}_{\mathrm{H}}$, where the design effect $\mathrm{deff}_{\mathrm{H}}$ is defined in Theorem 8.1 of Sect. 8.2.1. The pseudo empirical likelihood function with respect to $(p_1, \ldots, p_n)$ can be written as

$$L(p_1, \ldots, p_n) = \exp\Big\{n^* \sum_{i\in\mathbf{S}} \tilde{d}_i(\mathbf{S}) \log(p_i)\Big\} = \prod_{i\in\mathbf{S}} p_i^{\gamma_i}\,. \tag{11.1}$$

where $\gamma_i = n^* \tilde{d}_i(\mathbf{S})$. The expression given by (11.1) is not a genuine likelihood function but can be used to derive a posterior distribution for $(p_1, \ldots, p_n)$. Consider the Dirichlet prior $D(\alpha_1, \ldots, \alpha_n)$ on $(p_1, \ldots, p_n)$:

$$\pi(p_1, \ldots, p_n) = c(\alpha_1, \ldots, \alpha_n) \prod_{i\in\mathbf{S}} p_i^{\alpha_i - 1}\,,$$

where $c(\alpha_1, \ldots, \alpha_n) = \Gamma\big(\sum_{i\in\mathbf{S}} \alpha_i\big) / \prod_{i\in\mathbf{S}} \Gamma(\alpha_i)$ and $\Gamma(\cdot)$ is the gamma function. The parameters $(\alpha_1, \ldots, \alpha_n)$ are for the prior distribution. The posterior distribution of $(p_1, \ldots, p_n)$ given the sample data is Dirichlet $D(\gamma_1 + \alpha_1, \ldots, \gamma_n + \alpha_n)$ and is given by

$$\pi(p_1, \ldots, p_n \mid \mathbf{S}) = c(\gamma_1 + \alpha_1, \ldots, \gamma_n + \alpha_n) \prod_{i\in\mathbf{S}} p_i^{\gamma_i + \alpha_i - 1}\,.$$

The notations $\pi(\cdot)$ and $\pi(\cdot \mid \mathbf{S})$ for the prior and the posterior distribution follow the convention in Bayesian analysis and should not be confused with the inclusion probabilities π_i.

Note that the posterior distribution depends on the sample and the survey design through the definition of γ_i. A particular choice of the Dirichlet prior is the improper Dirichlet-Haldane prior corresponding to $\alpha_i = 0$ for all $i \in \mathbf{S}$ (Aitkin 2008). The posterior distribution of the population mean $\theta = \mu_y$ is determined through the relationship $\theta = \sum_{i \in \mathbf{S}} p_i y_i$ and the posterior distribution of $(p_1, \ldots, p_n)$. With the Dirichlet-Haldane prior, we have $\pi(p_1, \ldots, p_n \mid \mathbf{S}) \sim D(\gamma_1, \ldots, \gamma_n)$, which further leads to (Problem 11.1)

$$E(\theta \mid \mathbf{S}) = \hat{\mu}_{y\mathrm{H}} \quad \text{and} \quad Var(\theta \mid \mathbf{S}) = V_p(\hat{\mu}_{y\mathrm{H}}) + o(n^{-1}),$$

where $V_p(\hat{\mu}_{y\mathrm{H}})$ denotes the design-based variance for the Hájek estimator $\hat{\mu}_{y\mathrm{H}}$. It follows that the Bayesian point estimator using the posterior mean coincides with the design-based estimator and Bayesian credible intervals $(\hat{\theta}_1, \hat{\theta}_2)$ based on the posterior distribution of θ have valid frequentist properties under the design-based framework.

The Bayesian interval $(\hat{\theta}_1, \hat{\theta}_2)$ with balanced tail errors and the targeted $(1-\alpha)$-level coverage can be computed through the simulation-based Bayesian bootstrap method. Suppose that $\big(p_1^{[1]}, \ldots, p_n^{[1]}\big)$ is randomly generated from the Dirichlet distribution $D(\gamma_1, \ldots, \gamma_n)$. This can be done with most statistical software, including the free software R, either directly or indirectly. For the indirect method, we first generate $X_i \sim Gamma(\gamma_i)$, the gamma distribution with density function

$$f_i(x) = \Gamma(\gamma_i)^{-1} x^{\gamma_i - 1} \exp(-x)\,, x > 0\,,$$

independently for $i = 1, \ldots, n$, and then let $p_i^{[1]} = X_i / \sum_{j=1}^{n} X_j$, $i = 1, \ldots, n$. It can be shown (Problem 11.2) that $\big(p_1^{[1]}, \ldots, p_n^{[1]}\big) \sim D(\gamma_1, \ldots, \gamma_n)$. Let $\theta^{[1]} = \sum_{i \in \mathbf{S}} p_i^{[1]} y_i$. Repeat the process a large number B times to obtain $\theta^{[1]}, \ldots, \theta^{[B]}$. The lower and upper bounds of the interval $(\hat{\theta}_1, \hat{\theta}_2)$ can be approximated by the lower and upper $100(\alpha/2)$th sample quantiles of the simulated sequence for θ.

When the auxiliary population information $\mu_{\mathbf{x}}$ is known and is used for calibration weighting, we can replace the basic design weights d_i by the calibrated weights w_i in defining the pseudo empirical likelihood function $L(p_1, \cdots, p_n)$ given in (11.1). The only other modification is to replace the design effect $\mathrm{deff}_{\mathrm{H}}$ for the Hájek estimator by $\mathrm{deff}_{\mathrm{GREG}}$, which is the design effect for the generalized regression estimator defined in Sect. 8.2.2, and use $n^* = n/\mathrm{deff}_{\mathrm{GREG}}$ in the formulation of $L(p_1, \cdots, p_n)$. The resulting posterior distribution provides Bayesian point estimator and Bayesian credible intervals with valid design-based frequentist interpretations.

11.3 Bayesian Inference Based on Profile Pseudo Empirical Likelihood

The Bayesian approach described in Sect. 11.2 treats the nonparametric distribution $(p_1, \cdots, p_n)$ as unknown parameters and relates them to the parameter of interest $\theta = \mu_y$ through $\theta = \sum_{i\in\mathbf{S}} p_i y_i$. Another route for Bayesian inference is to use the profile pseudo empirical likelihood function. We consider general scenarios where the known population information $\mu_{\mathbf{x}}$ is used in the calibration constraints. Let d_i be the basic design weights. Let $\tilde{d}_i$, $n^* = n/\text{deff}_{\text{GREG}}$ and $\text{deff}_{\text{GREG}}$ be defined as in the previous section. The profile pseudo empirical log-likelihood function for $\theta = \mu_y$ is given by

$$\ell(\theta) = n^* \sum_{i\in\mathbf{S}} \tilde{d}_i(\mathbf{S}) \log\big\{\hat{p}_i(\theta)\big\}\,, \tag{11.2}$$

where the $\hat{p}_i(\theta)$ maximize $\ell_{\text{WR}} = n \sum_{i\in\mathbf{S}} \tilde{d}_i(\mathbf{S}) \log(p_i)$ subject to constraints

$$\sum_{i\in\mathbf{S}} p_i = 1\,, \qquad \sum_{i\in\mathbf{S}} p_i \mathbf{x}_i = \mu_{\mathbf{x}} \qquad \text{and} \qquad \sum_{i\in\mathbf{S}} p_i y_i = \theta \tag{11.3}$$

for a given θ. It is shown in Sect. 8.2.2 that $\hat{p}_i(\theta) = \tilde{d}_i(\mathbf{S})/\big\{1 + \lambda'(\theta)\mathbf{u}_i\big\}$, where $\mathbf{u}_i = \big(y_i - \theta, (\mathbf{x}_i - \mu_{\mathbf{x}})'\big)'$ and $\lambda = \lambda(\theta)$ is the solution to (8.18). It should be noted that the sampling design features are taken into account in the definition of the profile pseudo empirical log-likelihood function $\ell(\theta)$ given by (11.2) because it depends on the survey weights d_i and the design effect through the effective sample size n^*.

The profile pseudo empirical likelihood function is given by $L(\theta) = \exp\{\ell(\theta)\}$. Under the non-informative prior distribution $\pi(\theta)$ on θ, i.e., $\pi(\theta) \propto 1$, the posterior density function of θ is given by

$$\pi(\theta \mid \mathbf{y}, \mathbf{x}) = c(\mathbf{y}, \mathbf{x}) \exp\Big[-n^* \sum_{i\in\mathbf{S}} \tilde{d}_i(\mathbf{S}) \log\big\{1 + \lambda'(\theta)\mathbf{u}_i\big\}\Big]\,, \tag{11.4}$$

where $(\mathbf{y}, \mathbf{x})$ represents the observed sample data and $c(\mathbf{y}, \mathbf{x})$ is the normalizing constant such that $\int \pi(\theta \mid \mathbf{y}, \mathbf{x})d\theta = 1$. For the scalar θ considered here, the posterior density function $\pi(\theta \mid \mathbf{y}, \mathbf{x})$ can easily be computed by using a grid approximation method. A representative shape of the posterior density is shown in Fig. 11.1 (Problem 11.3).

Let $\hat{\theta}_{\text{PEL}} = \hat{\mu}_{y\text{PEL}}$ be the maximum pseudo EL estimator of μ_y given by (8.14). It can be shown that the posterior distribution of θ is asymptotically normal with mean $\hat{\theta}_{\text{PEL}}$ and variance matching the design-based variance of $\hat{\theta}_{\text{PEL}}$ (Rao and Wu 2010a, Theorem 1). This leads to the crucial statement that the $(1 - \alpha)$-level Bayesian

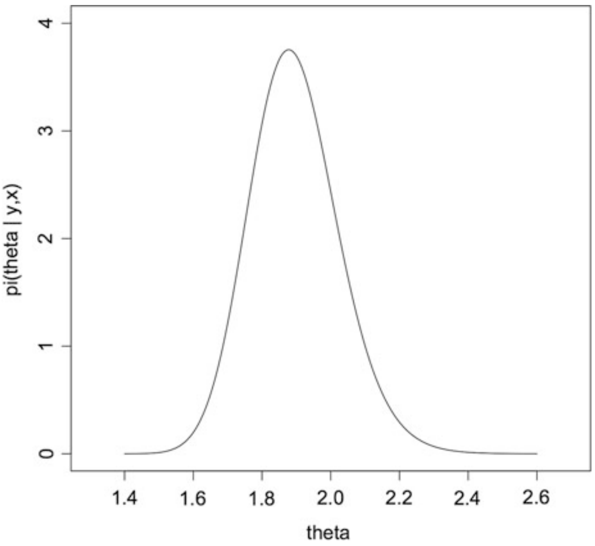

Fig. 11.1 A graphic representation of the posterior density function $\pi(\theta \mid \mathbf{y}, \mathbf{x})$ for the finite population mean $\theta = \mu_y$

credible intervals $(\hat{\theta}_1, \hat{\theta}_2)$ for $\theta = \mu_y$ have valid frequentist interpretation under the design-based framework.

There are two versions of Bayesian credible intervals: (i) the shortest interval $(\hat{\theta}_1, \hat{\theta}_2)$, where the lower and upper bounds satisfy $\pi(\hat{\theta}_1 \mid \mathbf{y}, \mathbf{x}) = \pi(\hat{\theta}_2 \mid \mathbf{y}, \mathbf{x})$ and $P(\hat{\theta}_1 < \theta < \hat{\theta}_2 \mid \mathbf{y}, \mathbf{x}) = 1-\alpha$; (ii) the interval $(\hat{\theta}_1, \hat{\theta}_2)$ with balanced tail error rates: $P(\theta \le \hat{\theta}_1 \mid \mathbf{y}, \mathbf{x}) = P(\theta \ge \hat{\theta}_2 \mid \mathbf{y}, \mathbf{x}) = \alpha/2$. Note that we have $\hat{\theta}_1 < \hat{\theta}_{\text{PEL}} < \hat{\theta}_2$ for both cases. The two intervals may have noticeable differences when the sample size n is small but behave similarly for large samples.

The basic design weights d_i may be replaced by the calibration weights w_i in defining the profile pseudo empirical likelihood function $\ell(\theta)$ given in (11.2). Under such scenarios the calibration constraints $\sum_{i \in \mathbf{S}} p_i \mathbf{x}_i = \mu_{\mathbf{x}}$ need to be removed from (11.3), and the resulting Bayesian posterior distribution of θ has similar properties.

For the special case of simple random sampling without replacement (SRSWOR), flat prior $\pi(\theta) \propto 1$ and no auxiliary information, the posterior density $\pi(\theta \mid \mathbf{y}, \mathbf{x})$ given by (11.4) reduces to

$$\pi(\theta \mid \mathbf{y}) = c(\mathbf{y}) \exp\Big[- \Big(1 - \frac{n}{N}\Big)^{-1} \sum_{i \in \mathbf{S}} \log\big\{1 + \lambda(\theta)(y_i - \theta)\big\}\Big]. \tag{11.5}$$

It differs from the Bayesian empirical likelihood posterior density function with independent and identically distributed (IID) sample data (Lazar 2003) by the finite population correction factor $1 - n/N$. When the sampling fraction n/N is not negligible, treating the sample selected by SRSWOR as if it is an IID sample, which amounts to replacing $1 - n/N$ by 1 in the posterior density given by (11.5), leads to over-coverage problems for the Bayesian credible intervals under the design-based framework (Problem 11.4).

11.4 Bayesian Inference for General Parameters

The two different Bayesian formulations presented in Sects. 11.2 and 11.3 are easy to implement, but they both involve the design effect which depends on the specific parameter of interest. Extensions to cases with a vector of population parameters seem to be difficult. In this section, we discuss Bayesian inference based on the profile sample empirical likelihood function (Sect. 8.7) for parameters defined through estimating equations.

Let $\mathbf{g}(y_i, \mathbf{x}_i; \theta)$ be the $r \times 1$ estimating functions for defining the $k \times 1$ vector of parameters θ_N. The estimating equations system can be over-identified (i.e., $r \geq k$) and the estimating functions can be smooth or non-differentiable with respect to θ. It is shown in this section that the estimating equations approach provides flexibilities in dealing with parameters such as population quantiles or parameters of quantile regression analysis without creating any additional difficulties.

Recall that the profile sample empirical log-likelihood function for θ is given by (Sect. 8.7.1)

$$\ell(\theta) = -\sum_{i \in \mathbf{S}} \log\Big[1 + \lambda'\{d_i \mathbf{g}(y_i, \mathbf{x}_i; \theta)\}\Big],$$

where the Lagrange multiplier $\lambda = \lambda(\theta)$ with the given θ is the solution to

$$\frac{1}{n}\sum_{i \in \mathbf{S}} \frac{d_i \mathbf{g}(y_i, \mathbf{x}_i; \theta)}{1 + \lambda'\{d_i \mathbf{g}(y_i, \mathbf{x}_i; \theta)\}} = \mathbf{0},$$

which corresponds to the set of constraints $\sum_{i \in \mathbf{S}} p_i\{d_i \mathbf{g}(y_i, \mathbf{x}_i; \theta)\} = \mathbf{0}$. The maximum sample empirical likelihood estimator of θ is denoted as $\hat{\theta}_{\text{SEL}}$, which is the maximum point of $\ell(\theta)$.

11.4.1 Bayesian Inference with a Fixed Prior

For notational simplicity, let $\mathbf{g}_i(\theta) = \mathbf{g}(y_i, \mathbf{x}_i; \theta)$. Let $\pi(\theta)$ be a fixed prior distribution which is independent of the sample size n. The posterior distribution of θ for the given sample $\mathbf{S}$ has the form $\pi(\theta \mid \mathbf{S}) \propto \pi(\theta)\exp\{\ell(\theta)\}$ and is given by

$$\pi(\theta \mid \mathbf{S}) = c(\mathbf{S})\exp\Big[\log\big\{\pi(\theta)\big\} - \sum_{i \in \mathbf{S}} \log\big\{1 + \lambda' d_i \mathbf{g}_i(\theta)\big\}\Big], \tag{11.6}$$

where $c(\mathbf{S})$ is the normalizing constant depending on the sample data $\{(y_i, \mathbf{x}_i, d_i), i \in \mathbf{S}\}$ such that $\int \pi(\theta \mid \mathbf{S})d\theta = 1$.

Theoretical properties of Bayesian inference based on (11.6) with large samples were investigated by Zhao et al. (2020a). Certain regularity conditions on the survey design, the finite population and the estimating functions are required in order to examine the asymptotic behavior of the Bayesian point estimators and credible intervals. In particular, the prior distribution is assumed to satisfy that the first derivative of $\log\{\pi(\theta)\}$ is bounded in a neighborhood of θ_N. Cases with non-differentiable estimating functions are accommodated by defining

$$\mathbf{H}(\theta) = \frac{\partial}{\partial\theta}\mathbf{U}(\theta) \quad \text{and} \quad \mathbf{U}(\theta) = \lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} \mathbf{g}(y_i, \mathbf{x}_i; \theta)\,.$$

The exact form of $\mathbf{H}(\theta)$ is not needed for Bayesian inference, as shown in Sect. 11.4.3. It follows from Theorem 1 of Zhao et al. (2020a) that the posterior density function given in (11.6) with a fixed prior has the following asymptotic expansion:

$$\pi(\theta \mid \mathbf{S}) \propto \exp\Big[-\frac{1}{2}\big(\theta - \hat{\theta}_{\text{SEL}}\big)'\mathbf{J}_n\big(\theta - \hat{\theta}_{\text{SEL}}\big) + R_n \Big]\,, \tag{11.7}$$

where $\mathbf{J}_n = n\mathbf{H}'\mathbf{W}^{-1}\mathbf{H}$, $R_n = o_p(1)$, $\mathbf{H} = \mathbf{H}(\theta_N)$, $\mathbf{W} = nN^{-2}\sum_{i=1}^{N} d_i\mathbf{g}_i\mathbf{g}_i'$, and $\mathbf{g}_i = \mathbf{g}(y_i, \mathbf{x}_i; \theta_N)$. Note that $\mathbf{W}$ is the same as $\mathbf{W}_2$ defined in Sect. 8.7.1. The posterior distribution of θ is asymptotically equivalent to a multivariate normal distribution with mean $\hat{\theta}_{\text{SEL}}$ and variance-covariance matrix $\mathbf{J}_n^{-1}$. The fixed prior distribution $\pi(\theta)$ has no impact on the posterior distribution under large samples.

One important implication from the asymptotic expansion of the posterior density function is that the posterior variance of θ matches the design-based variance of the posterior mean under single-stage PPS sampling without replacement with negligible sampling fractions (Problem 11.5). Consequently, Bayesian inference with any fixed prior has valid design-based frequentist properties under survey designs with large sample sizes and negligible sampling fractions.

11.4.2 Bayesian Inference with an n-Dependent Prior

The choice of a prior for Bayesian inference reflects the amount of prior information available to the data analyst. A fixed prior has impact on the analysis when the sample size is small or moderate but the influence diminishes under large samples. A stronger version of prior distributions is the so-called n-dependent prior, denoted as $\pi_n(\theta)$, for which the variance of the prior distribution shrinks as n gets large. In the literature on Bayesian inference, Zellner and Siow (1980) and Zellner (1986) studied the g-prior distributions which are n-dependent. Other examples include the priors used by Wasserman (2000), Yang and He (2012) and Narisetty and He (2014). For analysis of survey data, it is possible to obtain a point estimate with an estimated variance from existing surveys for the parameters of interest, and use the estimates to form a prior distribution. This technique was used by Rao and Ghangurde (1972) for Bayesian optimization in sampling finite populations.

The n-dependent prior $\pi_n(\theta)$ is assumed to satisfy the following regularity conditions: the function $\log\{\pi_n(\theta)\}$ is twice continuously differentiable, the prior density has bounded mode $\mathbf{m}_0 = \arg\max_\theta \pi_n(\theta)$, and the information matrix satisfies

$$\mathbf{H}_0 = -\Big[\frac{\partial^2}{\partial\theta\partial\theta'}\log\big\{\pi_n(\theta)\big\}\Big]\Big|_{\theta=\mathbf{m}_0} = O(n)\,.$$

It is shown in Theorem 3 of Zhao et al. (2020a) that the posterior density $\pi(\theta \mid \mathbf{S})$ given in (11.6) with the n-dependent prior $\pi_n(\theta)$ has the following asymptotic expansion:

$$\pi(\theta \mid \mathbf{S}) \propto \exp\Big[-\frac{1}{2}\big(\theta - \mathbf{m}_n\big)'\mathbf{K}_n\big(\theta - \mathbf{m}_n\big) + R_n\Big],$$

where $\mathbf{K}_n = \mathbf{H}_0 + \mathbf{J}_n^{-1}$, $\mathbf{m}_n = \mathbf{K}_n^{-1}\big(\mathbf{H}_0\mathbf{m}_0 + \mathbf{J}_n^{-1}\hat{\theta}_{\text{SEL}}\big)$, $R_n = 0_p(1)$, and $\mathbf{J}_n$ is defined in (11.7). The posterior distribution of θ is asymptotically equivalent to a multivariate normal distribution with the mean as a convex combination of the prior mode $\mathbf{m}_0$ and the maximum sample EL estimator $\hat{\theta}_{\text{SEL}}$ and the variance which is inversely related to the sum of the information matrix of the prior and the posterior variance under the noninformative prior.

The asymptotic expansion of the posterior density with an n-dependent prior shows that the impact of the prior distribution is asymptotically negligible if the information matrix of the prior satisfies $\mathbf{H}_0 = o(n)$. This leads to another crucial observation: The condition $\mathbf{m}_0 = \theta_N + O_p(n^{-1/2})$ on the prior distribution is necessary for the validity of design-based frequentist interpretation for Bayesian inference if the variance of the prior distribution is chosen with the order $O(n^{-1})$. For finite samples, the impact of the n-dependent prior $\pi_n(\theta)$ depends largely on the mode $\mathbf{m}_0$ of the distribution and to a lesser extent on the variance of the distribution or the information matrix $\mathbf{H}_0$.

11.4.3 Computational Procedures

The sample empirical likelihood based posterior distribution given in (11.6) for any pre-specified prior distribution can be efficiently simulated by an MCMC procedure using a Metropolis-Hastings algorithm. Bayesian inferences on θ_N or any functions of θ_N can therefore be carried out through a simulation based approach. The full posterior distribution of θ_N can be simulated as follows.

Let $h(\theta_2 \mid \theta_1)$ be a known convenient instrumental density, which is a Markov transition density describing how to go from state θ_1 to θ_2. For instance, one can use a multivariate normal distribution $MVN(\theta_1, \sigma_0^2 \mathbf{I}_k)$ for the distribution of θ_2 given θ_1, with σ_0^2 chosen to reflect the possible range of the parameters. Let $\theta^{(0)}$ be an initial value. A sample of draws from the posterior density $\pi(\theta \mid \mathbf{S})$ can be obtained by repeating the following two steps for $t = 0, 1, \cdots, T-1$:

(i) Given the current state $\theta^{(t)}$, generate θ^* from $h(\theta \mid \theta^{(t)})$ and compute

$$\lambda(\theta^{(t)}) = \arg\max_{\lambda \in \Lambda_n(\theta^{(t)})} \sum_{i \in \mathbf{S}} \log\left\{1 + \lambda' d_i \mathbf{g}_i(\theta^{(t)})\right\},$$

where $\Lambda_n(\theta) = \left\{\lambda \;\middle|\; 1 + \lambda' d_i \mathbf{g}_i(\theta) > 0,\ i \in \mathbf{S}\right\}$ is the feasible range for λ.

(ii) Generate $u \sim U(0, 1)$, the uniform distribution over $(0, 1)$, and compute the updated value $\theta^{(t+1)}$ using the following acceptance-rejection rule:

$$\theta^{(t+1)} = \begin{cases} \theta^* & \text{if} \quad u \le \alpha(\theta^{(t)}, \theta^* \mid \mathbf{S}), \\ \theta^{(t)} & \text{otherwise}, \end{cases}$$

where

$$\alpha(\theta_2, \theta_1 \mid \mathbf{S}) = \min\left\{1, \ \frac{\pi(\theta_1 \mid \mathbf{S})\, h(\theta_2 \mid \theta_1)}{\pi(\theta_2 \mid \mathbf{S})\, h(\theta_1 \mid \theta_2)}\right\}.$$

The value $\alpha(\theta_2, \theta_1 \mid \mathbf{S})$ is called the acceptance probability. It is practically important to notice that the normalizing constant $c(\mathbf{S})$ in the posterior density $\pi(\theta \mid \mathbf{S})$ is not required for the MCMC procedure due to the ratio of $h(\theta_2 \mid \theta_1)$ and $h(\theta_1 \mid \theta_2)$ in defining $\alpha(\theta_2, \theta_1 \mid \mathbf{S})$. Bayesian inference can be carried out based on the simulated sequence $\{\theta^{(t)} : t = T_0 + 1, \cdots, T\}$ with a large T after a suitable size of burn-in period T_0.

11.5 Additional Remarks

Bayesian inference is based on the posterior distribution of the parameters, which is conditional on the given sample. Conventional design-based inference has frequentist interpretations over repeated sampling. Godambe and Thompson (1971) were among the first to point out that Bayesian considerations could be mixed with frequency criteria, and such a mixing renders the robustness of Bayesian inference through its frequentist validation. While the use of Bayesian methods for survey data analysis, especially for official statistics, has long faced serious obstacles, there have been interesting discussions in recent years to revitalize the approach. Little (2011) advocated the use of "calibrated Bayes" to capitalize on the strength of both Bayesian and frequentist approaches. Under the calibrated Bayesian approach, inferences under a particular model are Bayesian, but frequentist methods are used for model development and model checking. Bayesian inference with valid design-based frequentist interpretation is not explicitly required by the calibrated Bayesian approach. Rao (2011) took a more pragmatic view towards Bayesian methods for survey sampling. Nonparametric Bayesian inferences with good design-based frequentist properties should be pursued for large samples and hierarchical Bayes methods based on parametric models can be employed for small area estimation.

The development of modern MCMC sampling techniques for Bayesian inference provides opportunities for survey data analysis which might be difficult under the traditional design-based framework. For instance, design-based estimation for parameters defined through non-differentiable estimating functions, such as population quantiles or parameters of quantile regression models, has been known to pose challenging problems. Under the Bayesian framework outlined in Sect. 11.4, they can be handled with the same MCMC procedure. In addition, estimation of parameters which are functions of the initial parameters or hypothesis testing problems can easily be handled with the same computational procedure. Other more advanced problems, such as model selection with complex survey data, can also be addressed under the Bayesian framework (Zhao et al. 2020a).

11.6 Problems

11.1 (Bayesian Pseudo Empirical Likelihood Based on $(p_1, \ldots, p_n)$) Suppose that $\pi(p_1, \ldots, p_n \mid \mathbf{S}) \sim D(\gamma_1, \ldots, \gamma_n)$, where $\gamma_i = n^* \tilde{d}_i(\mathbf{S})$, $n^* = n/\mathrm{deff}_{\mathrm{H}}$ and the design effect ($\mathrm{deff}_{\mathrm{H}}$) are defined in Sect. 7.1.1. Let $\theta = \sum_{i \in \mathbf{S}} p_i y_i$. Show that

$$E(\theta \mid \mathbf{S}) = \hat{\mu}_{y\mathrm{H}} \quad \text{and} \quad Var(\theta \mid \mathbf{S}) = V_p(\hat{\mu}_{y\mathrm{H}}) + o(n^{-1})\,.$$

Hint: If $(p_1, \ldots, p_n) \sim D(\gamma_1, \ldots, \gamma_n)$, then $E(p_i) = \gamma_i/a$, $Var(p_i) = (\gamma_i/a^2)\{1 - (\gamma_i + 1)/(a+1)\}$ and $Cov(p_i, p_j) = -\gamma_i\gamma_j/\{a^2(a+1)\}$ for $i \neq j$, where $a = \sum_{i=1}^n \gamma_i$.

11.2 (Computational Implementation of Bayesian Intervals Based on $(p_1, \ldots, p_n)$)

(a) Let $(\gamma_1, \cdots, \gamma_n)$ be pre-specified positive numbers. Let $X_i \sim Gamma(\gamma_i)$, $i = 1, \cdots, n,$ independent of each other. Let

$$Y_i = \frac{X_i}{\sum_{j=1}^n X_j}\,, \qquad i = 1, \cdots, n\,.$$

Show that $(Y_1, \cdots, Y_n) \sim D(\gamma_1, \cdots, \gamma_n)$, the Dirichlet distribution with parameters $(\gamma_1, \cdots, \gamma_n)$.

(b) Write an R program for generating random samples from a Dirichlet distribution using the indirect approach outlined in part (a).

(c) Write an R program to compute the $(1 - \alpha)$-level Bayesian interval $(\hat{\theta}_1, \hat{\theta}_2)$ described in Sect. 11.2 with the given α, $(\gamma_1, \cdots, \gamma_n)$ and the sample data $\{y_i, i \in \mathbf{S}\}$.

11.3 (Concavity of the Posterior Density Function) Show that the posterior density function $\pi(\theta \mid \mathbf{y}, \mathbf{x})$ given by (11.4) is a concave function of θ with the maximum value attained at $\theta = \hat{\theta}_{\text{PEL}}$, which is the maximum pseudo empirical likelihood estimator of μ_y given by (8.14).

Hint: Show that $d^2K(\theta)/d\theta^2 < 0$ for $\theta \in (y_{[1]}, y_{[n]})$, where $y_{[1]}$ and $y_{[n]}$ are the minimum and maximum order statistics and $K(\theta) = -\sum_{i \in \mathbf{S}} d_i \log\{1 + \lambda'(\theta)\mathbf{u}_i\}$.

11.4 (Bayesian Credible Intervals Based on the Profile Pseudo EL) Let $\{y_i, i \in \mathbf{S}\}$ be the sample of size n selected from a finite population of size N with SRSWOR. Let $\theta = \mu_y$ be the parameter of interest.

(a) Write an R program to compute the Bayesian credible interval $(\hat{\theta}_1, \hat{\theta}_2)$ with balanced tail error rates based on the posterior density function given by (11.5).

(b) Conduct a simulation study to investigate the effect of ignoring the finite population correction factor in terms of design-based coverage probabilities of the Bayesian credible intervals. Consider cases with $n/N = 0.01$, 0.05, 0.10 and 0.20.

11.5 (Posterior Variance Under Single-Stage PPS Sampling) Show that the posterior variance $\mathbf{J}_n^{-1}$ given in the asymptotic expansion (11.7) matches the design-based variance of the maximum sample EL estimator $\hat{\theta}_{\text{SEL}}$ under single-stage PPS sampling with replacement or single-stage PPS sampling without replacement with negligible sampling fractions.

Hint: The general variance expression for $\hat{\theta}_{\text{SEL}}$ is given in equation (8.39) of Sect. 8.7. An approximate design-based variance formula under single-stage PPS sampling is given in Problem 7.1.

Part III
Practical Issues and Special Topics in Survey Sampling

Chapter 12
Area Frame Household Surveys

This chapter discusses sampling designs and related issues for conducting household surveys using area frames and face-to-face interviews, through a case study from the International Tobacco Control (ITC) China Survey (Wu et al. 2010, 2015). It reveals some of the challenges faced by large scale household surveys with stratified multi-stage sampling designs and ambitious research plans with lengthy questionnaires, and provides some useful tips and tools for conducting such surveys.

12.1 Household Surveys

Household surveys are the backbone of information gathering on human populations in the modern era. The central role of the family in one's life is indisputable, and for most individuals, their activities and attributes are closely attached to their households. Although population censuses through listing and enumerating households have been used by many countries to collect essential information on social and demographic statistics, they can only be conducted every 5, 10 or 15 years and require a massive budget. Household surveys, on the other hand, are one of the most flexible sources of data for researchers and policy makers in social and economic studies, educational and community programs, medical and health sciences, etc. It is a powerful tool for collecting comprehensive, diverse and pertinent data on almost any population-based subject for scientific investigation and decision making.

Sampling strategies for household surveys depend largely on methods used for data collection. There are three traditional data collection methods: (i) mailed questionnaires; (ii) telephone interviews; and (iii) face-to-face interviews. The mode of data collection method often dictates the required sampling frames and survey operations as well as levels of cost for the survey.

C. Wu, M. E. Thompson, *Sampling Theory and Practice*, ICSA Book
Series in Statistics, https://doi.org/10.1007/978-3-030-44246-0_12

- Survey designs for mailed questionnaires require list frames of mailing addresses of all households in the target population. Mailed questionnaires are a less expensive mode of data collection and were widely used for surveys conducted from 1950s to 1980s. Their use in recent years has dramatically decreased due to low response rates, poor data quality and limitations on the types of variables that can be measured.
- Conducting household surveys through telephone interviews has been a popular approach for the past several decades. Telephone interviews are widely used in developed countries where landline telephones have become household commodities in the past century. Traditional telephone surveys rely on published phone directories as sampling frames. The development of random digit dialing (RDD) techniques allows more flexible frames and follow-up strategies. Computer assisted telephone interviewing (CATI) provides efficient tools for an integrated approach to data collection, data entry and data management. Some of the emerging issues with telephone surveys are connected with the movement toward the use of mobile phones, resulting in multiple phone numbers per household and multiple sampling frames. More detailed discussions can be found in Chap. 13 on telephone surveys and Chap. 16 on multiple frame surveys.
- Face-to-face interviews have much higher response rates than other household survey methods. They also allow the use of longer and more sophisticated questionnaires and better observations of respondent behavior. List frames of residential addresses obtained from administrative records or censuses are ideal for sample selection but may impose challenges for the fieldwork, which requires locating and contacting selected households. Area frames and area sampling are frequently used in practice due to better coverage of the target population and syncretized procedures of sampling selection and household visits. See Valliant et al. (2013) for further details on area sampling. The ITC China Survey to be discussed in this chapter uses area frames which are easy to construct for the survey population. Household surveys using face-to-face interviews, however, are more expensive and more time-consuming due to the requirements for personnel and travel and the demanding tasks of contacting and visiting selected households.

Household surveys have evolved from using a single mode of data collection to combining multiple methods. Survey data for health and medical studies sometimes require that measurements be taken at a clinical office or a special data collection center. As well, many household surveys are ongoing, with data collection being repeated periodically. There are three possible scenarios when a household survey is repeated over time, in what are often called *waves* or *cycles*: (i) cross-sectional, with each wave (or cycle) starting a fresh new sample; (ii) longitudinal, with the sampled units from the initial wave followed in subsequent waves; and (iii) rotational, with each wave adding a portion of newly sampled units while keeping a portion of the units from the previous wave on a rotating pattern.

There are many well-known household surveys which provide rich and continuing data sources for official statistics and scientific investigations. The General

Social Survey (GSS) and the Survey of Labour and Income Dynamics (SLID) of Canada, the Canadian Community Health Survey (CCHS), the Current Population Survey (CPS) and the National Health and Nutrition Examination Survey (NHANES) in the United States, and the General Household Survey (GHS) in Great Britain are just a few examples of highly impactful household surveys.

12.2 The ITC China Survey Design

The International Tobacco Control (ITC) Policy Evaluation Project began in 2002 with surveys conducted in an initial group of four countries: Australia, Canada, the UK and the USA. The ITC country surveys are longitudinal, and their aim is to evaluate the effects of policy measures being introduced in various countries under the WHO Framework Convention on Tobacco Control (FCTC). The initial ITC Four Country Survey used stratified sampling designs, with eligible households randomly selected by the RDD methods and data collected through telephone interviews. See Thompson et al. (2006) for methods of the ITC Four Country Survey.

12.2.1 The Target Population and Method for Data Collection

The ITC China Survey was launched in 2006. It was clear at the beginning of the planning stage that a national representative probability sample was not feasible due to the difficulties in covering the vast rural areas in China and a lack of resources and staffing. It was also clear that implementations of new tobacco control policies in China would almost surely start in big cities. The ITC China research team decided to carry out the survey in six cities: Beijing, Shanghai, Guangzhou, Shenyang, Changsha and Yinchuan. There was a seventh city under consideration but it was dropped from the study due to operational and data quality issues.

The six cities do not constitute a random sample of the entire population of China. They were judiciously selected based on geographical representation and levels of economic development. Members of the mobile population in these cities are not eligible for recruitment to the study due to the requirement of follow-ups in subsequent waves. The target population of the ITC China Survey consists of smokers and non-smokers who are 18 years or older, are permanent residents and live in residential buildings in each of the six cities. Smokers are defined as those who have smoked at least 100 cigarettes in their lifetime and are currently smoking at least once a week. Ex-smokers were not considered as a separate category at the initial wave of the survey.

The ITC China Survey was carried out through face-to-face interviews. The primary motivation behind the choice of this data collection method was not a lack of coverage of landline residential telephones but rather a cultural consideration. Most Chinese people are not used to accepting long interviews by telephone. Given the complexity and the sophistication of the survey questions and the longitudinal nature of the survey, it was decided that face-to-face interviews would be the best method for data collection.

12.2.2 Sample Size Determination

There were three main factors to be considered in sample size determinations. The first was the total cost. The survey had to operate within the available budget. The second was the requirement of not only obtaining reliable statistics at the aggregated level but also producing meaningful estimates for each city. The third was the primary focus on smokers as outlined by the main objective of the study.

Under stratified simple random sampling and for estimating population proportions at the stratum level, the most conservative choices of stratum sample sizes are $n_h = 1067$ in order to obtain estimates which are "accurate within three percentage points, 19 times out of 20" (Example 2.1 of Sect. 2.5.1). The overall sample size of the ITC China Survey for the baseline Wave 1 was $n = 6000$, with $n_h = 1000$ to be surveyed in each city, among them 800 adult smokers and 200 non-smokers. The size of the samples of non-smokers was constrained by available resources, but it nonetheless provided opportunities to examine differences between smokers and non-smokers in some of the key psychosocial and behavioral measures.

12.2.3 Frame Construction and Sample Selection

The ITC China Survey employed a stratified multistage cluster sampling design. Each city was treated as a stratum and within each city, there was a natural and well established hierarchical administrative system which provided excellent area coverage of the target population. Each city was geographically divided into first level administrative units, namely the Street Districts (Jie Dao), and each district was further divided into second level units, the Residential Blocks (Ju Wei Hui), all with clearly defined boundaries.

The sampling plan for the initial Wave 1 survey was to randomly select 10 Street Districts with probability proportional to the population size of the district. Within each of the ten sampled Street Districts, two Residential Blocks were selected, again with probability proportional to the population size of the block. The total number of residential blocks selected for each city was 20. The randomized systematic PPS sampling method (Sect. 4.4.2) was used for the selection of the districts and blocks. The final sample for each city consisted of 40 adult smokers and 10 non-smokers from each of the selected residential blocks, using the sampling method described below.

Frame construction for the top level clusters was relatively easy and straightforward. A list of Street Districts with information on population sizes was readily available for each city. For each of the ten selected districts, a list of Residential Blocks with population sizes was also available. The third level sampling frames were the lists of households within each of the selected Residential Blocks. The lists of households were available for most blocks from administrative records. If such a list did not exist, an area map was constructed to serve as the frame, with a convenient route and sequence of households that could be used for sample selection and field trips for data collection.

The lists or maps of households for the 20 selected Residential Blocks were not used directly for the selection of individuals. Instead, a sample of 300 households was drawn from each list by simple random sampling without replacement, resulting in a first phase sample of $300 \times 20 = 6000$ households in each city. The large first phase sample served as the sampling frame for the second phase (i.e., the final) sample of households and individuals. The use of PPS sampling at each of the first two stages (Street Districts and Residential Blocks), and a simple random sample of an equal number (300) of households in each selected block, ensured that each eligible household in the city had approximately the same chance of being included in the frame of 6000 households. See discussions on self-weighting survey designs in Problem 4.11.

An important part of the frame construction of the Wave 1 ITC China Survey was the complete enumeration of the 6000 households conducted prior to the selection of individuals. In the process, information on age, gender and smoking status for all adults living in these households was collected. The enumerated 300 households within each residential block were randomly ordered, and adult smokers and non-smokers were then approached following the randomized order until 40 adult smokers and 10 adult non- smokers were surveyed. Because of low smoking prevalence among women, one male smoker and one female smoker from each selected household were surveyed whenever possible to increase the sample size for women smokers. At most one non-smoker was interviewed per household. Where there was more than one person in a sampling category to choose from in a household, the next-birthday method was used to select the individual to be interviewed, and the selection was done prior to the household visit. Proxy interviews were not allowed in the ITC China Survey.

The frame of 6000 enumerated households in each city was also used in subsequent waves for selecting replenishment samples to replace those respondents from the previous wave who were lost to attrition. The same sampling procedure was used to select households from the list which were not sampled in the previous wave. If the list of 300 enumerated households in a residential block was exhausted, the required portion of the replenishment sample was either selected from an adjacent block or a newly enumerated list from the same block.

12.2.4 Survey Measures and Questionnaire Development

Surveys of the ITC Project were designed to measure (1) important smoking and smoking-related behaviours; (2) important psychosocial precursors to smoking and to cessation (e.g., intention to quit smoking, self-efficacy for quitting, beliefs about smoking and about quitting, perceived risk, societal and subjective norms, attitudes, denormalization beliefs); (3) important policy-relevant measures for each of the demand reduction policy domains of the FCTC, including those relevant to health warnings (e.g., salience, perceived effectiveness, behaviours relating to reactions to the warnings such as forgoing a cigarette because of the warnings), advertising/promotion (overall salience of pro-tobacco messages and anti-tobacco messages, noticing of tobacco sponsorships), purchasing and price-relevant behaviour, smoke-free laws, cessation, education. The survey also included key psychosocial mediators and (possible) moderators (e.g., time perspective, depression) of policy impact.

The main questionnaire for the adult smoker ITC China Survey included measures of the demand reduction policies of the FCTC, such as labelling, price/taxation, advertising/promotion, smoke-free, cessation, education, measures of behaviour and psycho-social characteristics. Most of these measures were common for all ITC surveys but some are specifically designed for the ITC China Survey. For example, the Wave 1 surveys (for both smokers and non-smokers) included a set of questions on the International Quit-and-Win Competition, an ongoing event organized by the Office of Tobacco Control of the Chinese Center for Disease Control and Prevention. The Wave 2 smoker survey included questions on alcohol consumption, intended to bring statistical evidence to bear on hypothesized psychological and behavioural linkages between drinking and smoking.

The development of the ITC China Survey questionnaires was a long, exacting and collaborative process. The questionnaires were first created in English, and included standard ITC survey questions used in other countries; they were then translated into Chinese through a system of multiple translators and with discussion of differences and resolution of those differences.

12.3 The ITC China Survey Procedures

The operational procedures of the ITC China Survey had to take into account several challenges faced by the team. The survey questionnaires were long and bulky, requiring more than half an hour to complete for smokers, and many psychosocial and behavioral questions were difficult to comprehend. It was also difficult to contact and access selected households due to the existence of several layers of security measures in most residential areas.

12.3.1 Survey Team

The ITC China team consisted of members from the Chinese Center for Disease Control and Prevention (China CDC) and international members from the ITC Project. In each city, a project coordinator was appointed at the provincial or city CDC, and the project coordinator subsequently assembled a team consisting of one or two deputy team leaders, one data manager, one quality controller and 20 interviewers. Most of these people were staff members at the local CDC, street districts or residential blocks, who were associated with the China CDC system. Some of the interviewers were recruited from students at local medical schools. Team members at the national China CDC as well as international team members oversaw all major steps in the survey execution.

The use of local staff members for the survey team turned out to be very crucial for the project. These people were trusted by the local residents, and were able to make phone calls to set up interview times and go through security checks to enter household dwelling units.

12.3.2 Training

All survey-related materials, including questionnaires, training and quality control manuals, were fully discussed and finalized at a pre-survey workshop. Participants of the workshop included the international team members, members from the China National CDC and representatives from each of the cities. The workshop provided a platform for key team members to have some commonality on the ITC China Survey project, to work out details for the training and fieldwork organization, to foresee potential problems and to suggest possible solutions. There were two training manuals developed, one for the enumeration process and one for the survey interview.

The complete enumeration for basic demographic information and smoking status of all adults living in the 300 randomly selected households within each selected residential block was the first crucial step of the survey. The enumeration data not only served as a basis for the final stage sample selection of individuals but also provided a rich source for the estimation of prevalence for different age-gender groups. This task was carried out by local residential block staff members, with training provided by each city. Training of interviewers was also organized at the city level, with support and supervision from the ITC China team members both at the China National CDC and at the ITC Project Data Management Centre at the University of Waterloo.

12.3.3 Quality Control

Several quality control procedures were put in place. One was a three-level checking of finished questionnaires. The ITC China team established an efficient reporting and communication system among the interviewers, the data manager and the quality controller of each city, and the central team members at the National CDC. A standard checklist was created for each of the three levels: the interviewer, the city quality controller and the designated central team member.

Another major quality control procedure was the practice of making MP3 recordings for each of the 800 smoker interviews in each of the six cities. These recordings were valuable not only in monitoring the quality of each interviewer's work, but also in alerting the research team to ways of improving the interview script for the survey and in identifying and correcting errors occurred during the data entry process.

12.3.4 Measures of Data Quality

High quality data are the ultimate goal of survey designs and procedures. A perfectly designed survey does not necessarily lead to data with desirable quality. Practical constraints and respondent behavior always inject noise and sometimes non-negligible biases into the final survey sample. It is imperative that large scale surveys report key measures of data quality so that users of the data are aware of potential issues with analysis.

Cooperation and Response Rates These are two key measures of data quality for cross-sectional and longitudinal surveys. For longitudinal surveys, the rates are computed at the baseline wave. The cooperation rate is calculated as the ratio of the number of completed interviews and the total number of successful contacts which include both completed interviews and refusals. The response rate is computed as the ratio of the number of completed interviews and the total number of individuals selected in the initial sample, including those who are unable to be contacted.

Retention Rate This is a key measure of quality for longitudinal survey data. The retention rate is calculated as the ratio of the number of individuals who have been successfully followed and interviewed in the current wave of the survey and the number of individuals who were interviewed in the previous wave. Retention rates can be similarly computed between any two waves of the survey.

12.4 Weight Calculation

Survey weights are calculated based on the sampling design and changes and modifications made during the survey process. We demonstrate how to compute the cross-sectional survey weights at the baseline wave and longitudinal survey weights between any two waves using the example of the ITC China Survey.

12.4.1 Dealing with Cluster Level Unit Nonresponse

The ITC China Survey used a stratified multi-stage cluster sampling design. The first stage and second stage clusters (Street Districts and Residential Blocks) were selected by randomized systematic PPS sampling. The sampling design, however, was altered in Guangzhou, where one residential block was replaced by a substitute unit, and also in Shenyang, where one street district was replaced by another one, because of unforeseeable changes in these two cities.

Cluster level unit nonresponse and substitution of units can also occur in other surveys. A question of both practical and theoretical interest is as follows: when the original sample units were selected by a randomized systematic PPS sampling method, and some units were later replaced by substitute units, selected from units not included in the original sample by the randomized systematic PPS sampling method, how should the inclusion probabilities for the final sample be computed?

Motivated by the real problem from the ITC China Survey, Thompson and Wu (2008) proposed a simulation-based approach to assessing the effect of substitution of units for the randomized systematic PPS sampling method. When all design information is available, which is the case for the ITC China Survey and many other multi-stage survey designs, the inclusion probabilities for the final modified design can be approximated through Monte Carlo simulations. Thompson and Wu (2008) also reported the following interesting observations: (i) when a PPS sampling procedure is modified owing to substitution of units, the resulting inclusion probabilities are no longer proportional to the size measure, even if the substitute units are selected by the same PPS sampling method; (ii) the impact of substitution of units on the final inclusion probabilities depends on the sizes of the units being replaced. If the units being replaced are of average size, the final inclusion probabilities under the modified sampling design are nearly proportional to the unit size.

12.4.2 Cross-Sectional Weights

Cross-sectional survey weights are calculated as the reciprocal of the inclusion probabilities. While the inclusion probabilities under a multistage sampling design are usually calculated as a product of the sequence of conditional inclusion probabilities from top to bottom, the weights are most conveniently constructed from bottom to top at the four levels of sample selection: individual, household, residential block and street district. Weights for the ITC China Survey were computed separately for adult male smokers and female smokers. Weight calculations demonstrated below are for adult smokers treated as a single sample.

- *Household level weights*. Each surveyed smoker has a household level weight W_1, which is the number of adult smokers in the household represented by the surveyed smoker.
- *Residential block level weights*. Each surveyed smoker has a block level weight W_2. This is the number of adult smokers in the block represented by the surveyed smoker:

$$W_2 = \frac{N_1}{N_2} \times \frac{M_1}{M_a} \times W_1 \,,$$

  where N_1 is the total number of households in the block, N_2 is the number of households enumerated (i.e., $N_2 = 300$ under the current design), M_1 is the number of smoking households (i.e., household with at least one adult smoker) among the N_2 enumerated households, and M_a is the number of smoking households surveyed to reach the quota of 40 smokers.
- *Street district level weights*. Each surveyed smoker has a district level weight W_3. This is the number of smokers in the same district represented by the surveyed smoker:

$$W_3 = \frac{P_b}{2P_c} \times W_2 \,,$$

  where P_b is the population size of the district and P_c is the population size of the block from which the smoker is selected. The factor 2 represents the number of blocks selected within the district and $2P_c/P_b$ are the inclusion probabilities for the second stage clusters (i.e., blocks) under PPS sampling.
- *City level weights*. Each surveyed smoker has a final city level weight W_4. This is the number of smokers in the city represented by the surveyed smoker:

$$W_4 = \frac{P_a}{10P_b} \times W_3 \,,$$

  where P_a is the population size of the city, the factor 10 is the number of street districts selected within the city, and $10P_b/P_a$ are the inclusion probabilities for the first stage clusters (i.e., districts) under PPS sampling.

The city level weights W_4 are the final survey weights. Weight calculations are carried out separately for each of the six cities. Similar calculations can be done for the sample of non-smokers.

12.4.3 Longitudinal Weights

Longitudinal surveys follow the initially surveyed individuals over time. Statistical analysis of longitudinal survey data focuses more on changes over time at the individual level and uses data from multiple waves (Sect. 7.4). Attrition is always the main issue with longitudinal surveys when some of the individuals surveyed at the baseline wave are lost in subsequent waves. Let $\mathbf{S}_1$ be the set of sampled individuals at the baseline Wave 1 and $\mathbf{S}_2$ be the set of individuals with successful follow-up at Wave 2 such that $\mathbf{S}_2 \subset \mathbf{S}_1$. The longitudinal sample consists of individuals from $\mathbf{S}_2$ with data available from both waves.

Calculation of longitudinal survey weights follows the same procedures of unit nonresponse adjustment as discussed in Sect. 9.2. The individuals lost in subsequent waves need to be removed from the initial sample since they do not provide longitudinal observations, and survey weights for individuals which remain in the longitudinal sample need to be adjusted. Methods for weight adjustment depend on mechanisms of lost to follow-up or refusals to stay in the sample. If the mechanism is similar to MCAR for unit non-response, then the ratio adjustment can be used. Let $w_i^{[1]}$ be the cross-sectional survey weight at Wave 1 for individual $i \in \mathbf{S}_1$. The Waves 1–2 longitudinal weights are computed as

$$w_i = w_i^{[1]} \frac{\sum_{j \in \mathbf{S}_1} w_j^{[1]}}{\sum_{j \in \mathbf{S}_2} w_j^{[1]}} , \quad i \in \mathbf{S}_2 .$$

The ratio adjusted longitudinal weights w_i, $i \in \mathbf{S}_2$ satisfy $\sum_{i \in \mathbf{S}_2} w_i = \sum_{j \in \mathbf{S}_1} w_j^{[1]}$. The ITC China longitudinal survey datasets imposed the MCAR assumption within each residential block. The longitudinal weights were computed by first using the ratio adjustment within each residential block and then rescaled at the city level so the total longitudinal weight remained unchanged from the initial cross-sectional weights.

12.5 Additional Remarks

Sample selection for household surveys is typically done first at the household level. Within each selected household, one eligible individual is then selected for the final survey sample. The next-birthday method was used by the ITC China Survey to select a respondent where there was more than one person in a sampling category to choose from in a household. Two other existing methods for selecting individuals within a household are the Kish method and the last-birthday method. Binson et al. (2000) compared the effectiveness of the three methods using data from a national telephone survey and showed that the next-birthday method had a higher rate of retaining respondents in subsequent waves, although the differences between the last-birthday method and the next-birthday method were not statistically significant.

Incentives are often used for household surveys to increase response rates. The ITC China Survey provided a small gift to each respondent who completed the survey. There are two related issues with the use of incentives. One is the increased cost, which can be substantial for large scale surveys. The other is the potential to introduce biases. Those who complete the survey because the availability of incentives may display different characteristics to those who do not care about incentives. The impact of using incentives needs to be assessed for the survey project at hand.

Data quality needs to be assessed and reported through suitable measures, such as cooperation rate, response rate or retention rate, as discussed in Sect. 12.3.4. It is therefore important to record the so-called *paradata* required for computing these measures during the survey process.

The ITC Project has had surveys conducted in more than 25 countries and regions since its inception in 2002. There are many challenges in the organization, data collection and analysis of international surveys. Analysis is an increasingly important part of the motivation for large scale cross-cultural surveys. The fundamental challenge for analysis is to discern the real response (or lack of response) to policy change, separating it from the effects of data collection mode, differential nonresponse, external events, time-in-sample, culture, and language. Thompson (2008) contains detailed discussions on issues, challenges and analysis strategies related to large scale international surveys.

Traditional household surveys using probability sampling methods face challenges in high cost, low response rate and slow operational process, which have forced statistical agencies and survey firms to seek cheaper and quicker alternative approaches. Web-based household surveys using volunteer or commercial panels have been used in recent years. This is part of the emerging topic on non-probability survey samples and is discussed further in Chap. 17.

Chapter 13
Telephone and Web Surveys

Although traditional area sampling designs and face-to-face data collection methods are still in use for some important surveys, the data in many surveys are collected over the telephone, either automatically or by interviewers, or over the internet (web)with self-administration. With new data collection technologies, capturing responses has become easier, but reaching potential respondents has become more difficult.

Examples of telephone surveys include opinion polling and election polling, now often carried out using automatic dialling and touch-tone or key pad responses, or social or health surveys using live interviewers. Web surveys are commonly used in marketing and customer surveys, but also increasingly for establishment surveys and surveys of individuals.

The purposes of this chapter are to:

- describe the sampling frames and sampling methods of modern social or health surveys being carried out using the telephone or the web
- discuss the components of survey error to be taken into account in the analysis of the data
- outline the assumptions usually employed in the construction of survey weights for such data

13.1 Telephone Survey Frames and Sampling Methods

The frames for telephone surveys have traditionally been directories, random digit dial (RDD) frames, or a combination of these. Landline telephone numbers, associated with household addresses, may or may not be listed in directories, depending on the preference of the holders. They present a way of reaching randomly selected households; typically, one respondent or a small number of respondents is selected from a household where telephone contact has been made.

C. Wu, M. E. Thompson, *Sampling Theory and Practice*, ICSA Book Series in Statistics, https://doi.org/10.1007/978-3-030-44246-0_13

RDD techniques evolved in order to ensure that in the random samples of telephone numbers, unlisted numbers would be called as well as listed numbers. Most RDD methods in use for official or scientific surveys are *list-assisted*, meaning that the numbers are taken randomly from the union of banks of numbers known from the lists to contain some residential telephone numbers. Some such methods use *hundreds banks* of telephone numbers of length k, each bank corresponding to a sequence of $k - 2$ digits which are the first $k - 2$ digits of a listed number. See Norris and Paton (1991) for an example.

With greater use of mobile phones, the population coverage of landlines has decreased. Although mobile phone penetration is very high in many countries, barriers to reaching and recruiting respondents through mobile phones have meant that telephone surveying is increasingly difficult to carry out. Comprehensive lists of mobile phone numbers are often not available, making it necessary for an RDD method to call many numbers from a very large random number frame to reach potential respondents; depending on their contracts, respondents may have to pay to receive calls; and often mobile phone holders are reached at times and places where it is not convenient to respond to a survey.

Recently there has been a trend to recruit respondents for telephone surveys through other modes of approach. Prominent among these is Address Based Sampling (Link et al. 2008; Shook-Sa et al. 2016), where residences are sampled from an exhaustive list of postal addresses, and households are recruited through mailed or face-to-face invitations. Selection of respondents can then be made within participating households. Respondents can be asked to mail in responses, or can be sent to an online survey (Dillman 2017); alternatively, interviews may be carried out face-to-face with Computer Aided Personal Interviewing (CAPI), or (relevant to this section) by telephone, at appointed times.

When it is the household that is reached initially, as occurs with telephoning a landline number or mailing to an address, a respondent or respondents within the household must be selected, preferably according to a probability sampling design. Ideally, the person contacted initially would *roster* the household (i.e. list the household members with their relevant characteristics) and make a selection at random from among those eligible to be respondents. However, when the initial contact is by “cold calling”, there is seldom time to carry out this exercise. Typically, the interviewer will ask for the eligible household member with the next birthday, or with the most recent birthday. It is not difficult to see that this method will not effect a random selection within a household, or even generate a sample representative with respect to (say) age and marital status (see Problem 13.1), but the hope is that it will remove the kind of bias that might result from the survey always being answered by the person who picks up the phone, or by the person in the household most willing to respond. In practice, there is seldom a way of confirming that the selected respondent really is the one with the next birthday, and the composition of a telephone sample recruited this way is often skewed toward women and people in older age groups. Thus sometimes *quotas* (target numbers of respondents) for sex and age groups are also imposed for the overall sample.

13.2 Components of Survey Error for Telephone Surveys

For descriptive aims of a telephone survey, such as estimating a population mean, total or proportion for all adults in the population, there several sources of non-sampling error, in addition to the sampling error due to interviewing only a sample of the population.

If recruitment is done through cold calling of telephone numbers, the frame of landline numbers will suffer from *under-coverage* if not every member of the target population lives in a household with a landline, and *over-coverage* if some members of the target population are reachable at more than one landline number. If the landline frame is supplemented with a mobile phone frame, under-coverage will be less severe, but substantial over-coverage is highly likely, as many people will be reachable from both frames. The frame overlap will have to be accounted for carefully.

Non-response in surveys with telephone recruitment is increasingly prevalent, and can occur at several stages of the attempt to recruit, obtain consent, and interview. Most commonly, the household's telephone number is not answered in several attempts to call it, and contact is never made with the household (landline) or the individual (mobile phone). Less commonly, but still frequently, the person who answers the telephone hangs up during the attempt by the interviewer to introduce the survey, or refuses to bring the intended respondent to the telephone. Finally, the intended respondent may speak with the interviewer, but refuse to consent to respond—or may consent to respond, and begin the interview, but refuse to complete it. Non-response introduces bias into the ultimate sample selection, as individuals who do respond tend to be those who perceive that they have the time to do so, or have more interest in the subject of the survey. Goyder (1987) and Tourangeau et al. (2000) explore the motivations of sampled individuals to respond or to refuse.

A closely related problem for landline recruitment is the inadvertent failure to interview the intended respondent due to mis-application of the rule for selecting a respondent from the household. This too introduces bias into the ultimate sample selection.

Measurement error is present when the interview responses are untrue or when the responses fail to measure the intended quantity or construct adequately. When the data collection mode (in this case the telephone interview) influences the response through cognitive effects or through the presence of an interviewer, there may exist measurement biases relative to what might have been obtained through another mode. For example, there is evidence that reported physical activity tends to be lower in face-to-face interviews than in telephone interviews (Béland and St-Pierre 2008). There may be a so-called *social desirability* bias (Kreuter et al. 2008) in the telephone interview in this case. For another example, if a question has a Likert scale of responses such as *Strongly agree, Agree, Neither agree nor disagree, Disagree, Strongly disagree*, telephone respondents, who are hearing the options, are more likely than web respondents, who are seeing the options, to pick an extreme option (Dillman et al. 2009).

13.3 Web Survey Frames and Sampling Methods

In the most straightforward cases, the frames of potential respondents to web surveys are email lists. An invitation to respond is sent by email with a link to the online questionnaire. An email list may be institutional, such as a list of employees of an organization or students at a university. Otherwise, it may be a list constructed and maintained by a government agency, or a *panel* managed by a survey firm. Typically, for the latter, the panel members, once recruited, are invited to surveys from time to time, and are paid in some manner for each survey questionnaire completed.

The recruitment methods of survey firms vary. In some cases, the panel members are recruited from a probability sample, possibly Address Based, or from respondents to RDD telephone surveys. These *probability based* methods tend to be relatively costly. For many panels, recruitment occurs through websites in which potential respondents "opt in" rather than being invited in a systematic manner. The firms monitor respondent behaviour, and will try to weed out those who show a tendency to respond too speedily or respond inattentively.

There are important advantages to surveying online:

- the ability to reach large numbers of potential respondents very cheaply
- the ease of sending reminders of invitations
- the "richness" of the frames: typically, the institutional list or the commercial panel will have information about its members that will allow invitations to be targeted to specific groups, such as cigarette smokers; this information will also facilitate integration of the survey data with data from other sources, as described in Chap. 17
- the ability to draw stratified random samples from the frames with ease
- the potential for increase in quality of responses through careful programming of the questionnaire
- the absence of an interviewer, and of interviewer effects such as social desirability bias
- the ease of data capture and preparation of data for analysis

The disadvantages are also formidable:

- the often prohibitive expense (to a firm) of constructing and maintaining a comprehensive or probability based frame or panel –
- – and the resulting necessity to use non-probability samples such as opt-in panels to obtain sufficient sample size
- the transitory nature of the contact information (email addresses)
- other technical reasons for invitations becoming lost
- the potential for respondent over-use, or respondent fatigue from large numbers of invitations –
- – and resulting low rates of retention of panel members
- the absence of an interviewer's guidance and ability to pace the interview

All traditional components of nonsampling error—coverage errors (under-coverage or over-coverage), nonresponse error, and measurement error—have a large presence in a web survey, when the target population is a subset of the general population. Most obviously, web panels tend to differ from the general population of those likely to respond to surveys on the dimensions of age (the age of web panellists has tended to be younger), internet access and frequency of internet use.

It is somewhat paradoxical that in many societies, comprehensive information exists in electronic form about most members of the population, yet it is increasingly difficult to ask new questions directly of a representative sample. Much research is in progress on how to exploit comprehensive auxiliary information to minimize the effects of nonsampling error, so that web surveys can reach their full potential in the economic, social and health sciences.

13.4 Mixed Mode Surveys

There are advantages to combining modes of data collection. For example, in a multiple frame survey, a variety of modes of recruitment may be natural; different modes may be used for recruitment of respondents and for the actual interviews; or respondents once recruited may be given a choice of their preferred interview mode.

An example of use of different modes to recruit and to collect data is the Connecting Health and Technology (CHAT) study (Cantrell et al. 2018), in which young respondents were recruited through an initial mailing and follow-up from households selected from a combination of an area sampling frame and a rich address frame; once recruited, the participants provided their responses by web.

The International Tobacco Control Survey has two examples of longitudinal surveys in which the interview mode has changed over time, from telephone to web (in Canada, the US, the UK and Australia) or from face-to-face to telephone (in Malaysia). The gradual nature of the changes has allowed data collection mode effects to be modeled and accounted for in cross-wave and cross-country comparisons. See for example Huang et al. (2010).

13.5 Construction of Survey Weights

For descriptive aims, the purpose of including weights with the dataset is to provide the user with a means of constructing design-based point estimates that are approximately unbiased for finite population quantities such as means and proportions.

When the sampling frame used to recruit the respondents is both comprehensive and a *rich frame*, such as some address-based frames for households, or registries (listings) with large amounts of information on individual units, and when the initially intended sample is a probability sample, the steps begin with the construction of *survey design weights* that are the reciprocals of inclusion probabilities, and proceed through adjustments of the weights for non-response and calibration, as described in Chap. 6. This is true regardless of the mode of data collection or interview.

However, when respondents are recruited through RDD calling, or through opt-in procedures on the web, these steps cannot be implemented fully. Survey weights are still typically provided, but with cautions about difficulties in assessing biases in the estimators.

13.5.1 Survey Weights for RDD Surveys

In the case of RDD telephone surveys, the sample of numbers called is typically a simple random sample or a geographically stratified random sample from the frame or frames of (to some extent) randomly generated telephone numbers.

1. Case of a Landline RDD Frame

If the frame consists of landline numbers, names and postal addresses may be available for the listed numbers, but otherwise the frame contains little information about the household. Some of the numbers reached will correspond to households where no one is eligible to respond to the survey. The final sample of numbers from which a respondent is identified thus constitutes an equal probability sample (within strata) from the population of eligible and potentially responding households with landlines. As the size of this population is typically unknown, the survey design weight for a household is usually taken to be an unknown constant A_{0Lh} within its geographic stratum h. The constant A_{0Lh} is the number H_{Lh} of landline numbers of eligible households divided by the number h_{Lh} of responding eligible households within stratum h. If the sampling design selects one eligible person at random in an eligible household, the individual's survey design weight is $k \times A_{0Lh}$ where k is the number of eligible persons in the household.

The calling process itself may add so-called *paradata*, to the list of numbers; these are data on characteristics that can be observed to be associated with response rates from numbers known to be residential. See West (2011) for an overview of paradata and their uses. They are potentially helpful in adjusting the design weights to reduce non-response bias, and would be implemented in inferences from surveys like election polls where minimizing bias is particularly important.

Finally, the weights will be calibrated within strata to known or approximate totals for the eligible population (from censuses or high quality official surveys) by demographic characteristics such as cross-tabulations of sex, age group and educational attainment. The calibrated weights do not depend on the constants A_{0Lh}.

2. Case of a Mobile Phone RDD Frame

The approach in this case is similar to that for a landline frame, although the frame itself is likely to be even less informative in this case. As mobile phone numbers are usually associated with individuals rather than households, the survey design weight for the individual would usually be taken to be a constant A_{0Mh} within the individual's geographic stratum h. The constant A_{0Mh} can be thought of as the number N_{Mh} of eligible people contactable at mobile phone numbers, divided by the number n_{Mh} of eligible individuals responding by mobile phone, within the geographic stratum h. The weights will be calibrated as in the case of the landline frame, and the calibrated weights will not depend on the constants A_{0Mh}.

3. Case of Frames of Landline and Mobile Numbers

If the RDD survey has both a landline and a mobile phone frame, with the true extent of both the covered populations and their overlap being unknown, calculating a survey design weight for an individual up to a constant multiple requires more information. As will be explained in detail in Chap. 16 on multiple frame surveys, it is useful for analysis to know the frame membership of each respondent in the sample. One possible approach to weighting is to ask each landline respondent "Would you have been available through a mobile phone during [survey period]?"; and to ask each mobile phone respondent "Would you have been available through a landline during [survey period]?" The probability of inclusion of someone in stratum h who is available only by landline would be given by $1/(k \times A_{0Lh})$, where k is the number of eligible people in the household. The probability of inclusion of someone in stratum h who is available only by mobile phone would be given by $1/A_{0Mh}$. If the sample sizes are small relative to the population size, the probability of inclusion of someone available through both would be approximately the sum of those two quantities. However, in order to know the inclusion probabilities of all respondents up to a constant multiple within stratum h, we need to be able to estimate the relative values of A_{0Lh} and A_{0Mh}.

Let $\mathbf{S}_{Lh}$ be the sample of households j in stratum h from which a landline respondent is obtained, and let k_j be the number of eligible people in the j-th household. Let $j(i)$ be the household from which landline respondent i is recruited. Let $\mathbf{S}_{Mh}$ be the sample of respondents i in stratum h who are recruited by mobile phone. Finally, let $\delta_i = 1$ if respondent i is available through both frames, and $\delta_i = 0$ otherwise. We then have two estimates of the size in h of the population of eligible people available through both frames, namely

$$\sum_{i \in \mathbf{S}_{Lh}} k_{j(i)} A_{0Lh} \delta_i \qquad \text{and} \qquad \sum_{i \in \mathbf{S}_{Mh}} A_{0Mh} \delta_i \,.$$

Equating these two estimates gives

$$A_{0Mh} = A_{0Lh} * \tilde{n}_{LMh} / n_{MLh}, \tag{13.1}$$

where n_{MLh} is the number of stratum h respondents reached by mobile phone who are also available by landline, and $\tilde{n}_{LMh} = \sum_{i \in \mathbf{S}_{Lh}} k_{j(i)} \delta_i$ is the number of eligible people in responding stratum h landline households whose chosen respondent is available by mobile phone. See Problem 13.2 for the use of this fact in construction of an approximate survey design weight for each member of the sample.

Calibration can then be carried out as in the two previous cases.

13.5.2 Survey Weights for Web Surveys

As noted at the beginning of this section, when the web survey sample is a probability sample from a well-defined frame, the construction of weights proceeds as described in Chap. 6. On the other hand, when the web survey sample is a non-probability sample, design inclusion probabilities are not available. Chapter 17 sets forth the problems of analysis and weights construction in such a case, and outlines current research on possible solutions.

13.6 Problems

13.1

(a) Consider a household with three adult members A, B and C. Their respective birthdays are March 1, June 1 and November 1. Suppose that in a household telephone survey, the interview is to be administered to the adult in the household with the next birthday. In a year which has 365 days, suppose the recruitment call to the household takes place on a randomly selected day, and the person answering the telephone follows the interviewer's instruction correctly. What is the probability that person A is selected for interview? Similarly for persons B and C. Why would practitioners analyse the data from a survey with this method of recruitment as though everyone in the household has the same chance of being selected?

(b) In a household survey population, suppose that the percentages of the five household types are: (I) One adult, 27%; (II) A couple of adults with no children, 25%; (III) A couple with children, 25%; (IV) A single parent with children, 9%; and (V) Exactly three adults, 14%. Suppose the proportions according to age of oldest adult (AOA) are as in the following table.

	HH Type				
AOA	I	II	III	IV	V
18–30	0.06	0.06	0.08	0.04	0.06
31–60	0.11	0.07	0.14	0.04	0.02
61+	0.10	0.12	0.03	0.01	0.06

Suppose an equal probability sample of households is taken, and that just one adult is selected for interview from each household, at random or using a next birthday method. Consider the composition of the sample, *unweighted*. Assuming there is no non-response, show that adults in households where the oldest adult is 18–30 years of age will be under-represented.

13.2 Starting from equation (13.1), indicate how to produce (up to an unknown constant) survey design weights for the members of the combined sample from the RDD frame and the mobile phone frame, as input to a calibration step.

Chapter 14
Natural Resource Inventory Surveys

We discuss sampling strategies for natural resource inventory surveys through the Fishery Products International (FPI) survey of fish abundance indices. The primary objective of the chapter is to formulate the problem under the general framework established in Parts I and II on survey design and estimation, and to provide useful tools for conducting surveys under similar circumstances.

14.1 Surveys of Non-human Populations

Human populations serve as the foundation for developing finite population sampling theory and methods. The concepts of finite populations, sampling frames, observational units, survey variables, etc., can be illustrated through concrete examples of surveys of human populations. The three basic features of complex survey designs, namely stratification, clustering and unequal probability selection, become inevitable parts of survey designs when the population structures dictate frame construction and sample selection. Other issues such as mode of data collection, questionnaire design, recruitment strategies, training of interviewers, and treatment of nonresponse arise naturally from surveying human populations. However, there are also types of surveys conducted with non-human populations, with each type having its own unique characteristics but all sharing common features of finite population sampling methods.

14.1.1 Non-human Population Surveys

Business and Establishment Surveys Establishment surveys seek to measure the structure, activities and outputs of organizations such as universities and hospitals

C. Wu, M. E. Thompson, *Sampling Theory and Practice*, ICSA Book Series in Statistics, https://doi.org/10.1007/978-3-030-44246-0_14

rather than individuals. Business surveys are a special type of establishment survey that collect information on trends in the activities and the well-being of business operations, which are critically important to economic policy makers. Sampling frames for business surveys are lists of registered firms and enterprises in various business sectors. The Canadian Survey of Labor and Income Dynamics (SLID), which gathers information on two sensitive barometers of the economy, namely employment and income, is not a business survey since it is conducted on the population of Canadians. The two books edited by Cox et al. (1995) and Snijkers et al. (2013) are rich in material on many aspects of designing and conducting business surveys.

Agricultural Surveys Agricultural surveys collect information on farming activities and products, including traditional crops, yields, livestock, land use and other agricultural resources. Although the agriculture sector has assumed a more and more marginal economic role in Western countries during the past century in terms of its percentage contribution to Gross National Product (GNP), the demand for agricultural information has never faded, due to the importance of issues related to sustainable development, the food industry, quality of life and public health. Agricultural activities are highly correlated to land use, recovery and protection of the environment, and the need for timely and accurate information has led to continued interest in the field. Agricultural surveys share certain common features with establishment and business surveys and have some similarities to natural resource inventory surveys. Benedetti et al. (2010) present a comprehensive treatment of sampling methodologies for agricultural surveys with extensive discussions of related practical issues.

Natural Resource Inventory Surveys Natural populations such as forests, landscapes, environments and ecosystems do not fit naturally into the finite population sampling framework. Many commonly used terms in survey sampling, from sampling units and sampling frames to study variables and population parameters, require precise definitions under suitable formulations of the survey problems. Natural resource inventory surveys may further be refined as or extended to environmental surveys, ecological surveys, forestry surveys, etc. The fish abundance index surveys over a large ocean area to be discussed in the rest of the chapter deal with the important natural resource of fish populations, and the sampling strategies are more akin to surveys of conventional natural resources than surveys of wild animals or human populations. Gregoire and Valentine (2007) is an excellent reference on sampling strategies for natural resources and the environment.

Surveys of Animal Populations Populations of birds and wild animals in a region or fish in an enclosed water area impose unique challenges for scientific investigations. Information gathering over such populations requires special sampling techniques. Distance sampling (Buckland 2004), for instance, is a widely used method for estimating the density or the abundance of animal populations. Capture-recapture techniques and their variations (Seber 1982) are effective tools for estimating the sizes of animal populations. Adaptive and network sampling methods to be described in the next chapter also find applications in animal population surveys.

14.1.2 Features of Natural Resource Inventory Surveys

Sampling strategies for establishment and business surveys as well as agricultural surveys are in line with classic finite population surveys where sampling units and sampling frames may be constructed through registered lists of firms, enterprises and farms. There are crucial components of human participation in the survey process and measurements of survey variables can be obtained through questionnaires or web-based reporting systems.

On the other hand, development of sampling methodologies for natural resource inventory surveys requires careful formulation of the problem and precise adaptations of the essential terms or even the basic concepts of finite population sampling methods, in part because the analogue of the sample unit may be a spatial location.

- The population along with a certain attribute such as a forest with its density or an ecosystem with its coverage should be treated as continuous and may not naturally divide into smaller discrete units. The parameter of interest is usually the total amount of the attribute over the region, which can be quantified as an integral of a continuous attribute density function.
- The absolute density of the attribute at a particular point is often not measurable and instead a surrogate measure reflecting the relative density is used for data collection.
- Sampling and estimation are carried out through suitable approximations and/or discretized methods using fixed area plots or artificially defined grids.
- Auxiliary information from aerial or satellite images or historical data may be used to help design the survey and to yield more efficient estimation.

The relative measure of the density serves the purpose of managing and monitoring natural resources if the same survey is repeatedly conducted for the same region over time. Natural resource inventory surveys and agricultural surveys may also have an interest in exploiting the spatial characteristics of the data to determine design and estimation strategies; see Benedetti et al. (2015) for further discussions.

14.2 Fish Abundance Surveys

The management of fishery resources in an open ocean region requires reliable estimation of fish abundance. The absolute fish abundance is impossible to obtain. Instead, fish abundance indices estimated from catch data are used in practice. The estimated indices are the major factors considered in the decision-making process for setting annual quotas for commercial fish species with high abundance, and maximum allowable percentages of by-catch (unintentional capture) of species under monitor for low abundance.

We present a case study on the design and analysis of the scientific research trawl surveys conducted by Fishery Products International (FPI) Ltd (Chen et al. 2004). We demonstrate how classic survey sampling methods can be adapted for fish abundance surveys and show how to use empirical likelihood methods to compute point estimates and confidence intervals of the fish abundance indices.

14.2.1 The Fish Abundance Index

A fish population in a large open area does not constitute a conventional finite population. It is a mobile population that changes over time due to migration, recruitment, natural mortality and fishing mortality. The fundamental principle for design-based analysis in survey sampling, namely that the finite population parameters are fixed and can be determined without error by conducting a census, is grossly violated. The underlying population dynamics are hard to observe and the true stock size may never be known. Fortunately, what is really important for stock assessment and management is to monitor the fluctuation of the fish population so that a major decline or boost in the population size or the total population biomass can be detected, and consequently, appropriate management strategies can be adopted.

Consider the fish abundance in an open ocean region denoted by $\mathbf{R}$. The fish abundance, i.e., the total stock size in the region, can be defined as $A(\mathbf{R}) = \int_{\mathbf{R}} \lambda(\mathbf{x})d\mathbf{x}$, where $\lambda(\mathbf{x})$ is the density of the fish stock at location $\mathbf{x}$. This definition, however, cannot be used in practice. In addition to the density function $\lambda(\mathbf{x})$ being unknown, $A(\mathbf{R})$ is not a finite population parameter. Its formulation does not admit the use of covariates other than the location variables. The problem can be reformulated through two modification steps.

The first step is to discretize the definition of fish abundance. We divide the region $\mathbf{R}$ into N equal-sized grid squares represented by $\mathbf{g}_i$, $i = 1, \cdots, N$. Let $\mathbf{x}_i$ be the central location of grid square $\mathbf{g}_i$. The fish abundance can be re-defined as $A(\mathbf{R}) = \sum_{i=1}^{N} \lambda(\mathbf{x}_i)$, where $\lambda(\mathbf{x}_i)$ represents the total stock size in grid square $\mathbf{g}_i$.

The second step is to define a response variable that can be observed for the survey. Historically, the catch-effort data reported by commercial fishing units are used to estimate the population abundance. Such information, however, is confounded with many uncontrolled factors and does not provide a reliable picture of fish abundance. Scientific research trawl surveys using a standard vessel and gear type and a probability sampling design have been adopted by many organizations since the 1970s. The response variable Y is defined as the number (or biomass) of fish caught in a given location (grid square) by the research vessel through a unit of fishing time, i.e., catch per unit effort (CPUE). It is apparent that Y is a random variable due to the constant dynamic movement of the fish population and other uncontrolled factors such as crew experience and the time point for data collection.

Let $\mu(\mathbf{x}_i) = E_\xi(Y_i \mid \mathbf{x}_i)$ be the expected CPUE, where ξ indicates the model for the catching process and $\mathbf{x}_i$ may include other covariates in addition to the location variables. The fish abundance index in the region $\mathbf{R}$ is defined as

$$I(\mathbf{R}) = \sum_{i=1}^{N} \mu(\mathbf{x}_i)\,. \tag{14.1}$$

In fisheries literature, it is often postulated that $\mu(\mathbf{x}_i) = c\lambda(\mathbf{x}_i)$, where c is the so-called catchability coefficient (Schnute 1994). If c can be determined from other sources, then $I(\mathbf{R})/c$ is the total population stock size in the region. The true relation between $\mu(\mathbf{x}_i)$ and $\lambda(\mathbf{x}_i)$, however, can be very complicated (Gunderson 1993). We focus on estimating the fish abundance index $I(\mathbf{R})$ defined by (14.1) using scientific research trawl survey data.

14.2.2 The FPI Survey Design

Fishery activities in the Grand Bank region on the east coast of Canada are governed by Northwest Atlantic Fisheries Organization (NAFO). The region is an important resource for Canadian fisheries industry. Surveys over part of the region have been conducted by the Fishery Products International (FPI) Ltd of Canada since 1996 to provide estimates of abundance indices for several important commercial fish species including yellowtail flounder, Atlantic cod, and American plaice. There have also been historical catch data available to help with the survey design and estimation.

The entire region is divided into $N = 626$ equal-sized grid squares, each 10×10 square miles. The fish abundance densities for yellowtail flounder based on the 2001 survey data are shown in Fig. 14.1 with quantiles of the population distribution. The finite population of 626 grid squares is stratified where the strata boundaries were determined based on practicality in terms of data collection using a scientific vessel as well as homogeneity of fish abundance within each stratum. The density plots for other fish species suggest slightly different boundaries for the stratification. The five strata shown in Fig. 14.1 are primarily based on the density for yellowtail flounder but work reasonably well for several other major species in the region.

An approximately optional sample size allocation scheme for the stratified sampling design can be derived as follows. The expected CPUE over the $N = 626$ sampling units has a clear non-uniform distribution, and the variations are typically large for areas with high densities and small for low density areas. It is therefore reasonable to assume for the purpose of sample size allocation that the stratum standard deviations are proportional to the stratum means of the expected CPUE. The current year CPUE should be highly correlated to CPUE from historical data. This leads to the allocation scheme of stratum sample sizes to be proportional to the estimated stratum totals of the CPUE using data from previous surveys or other

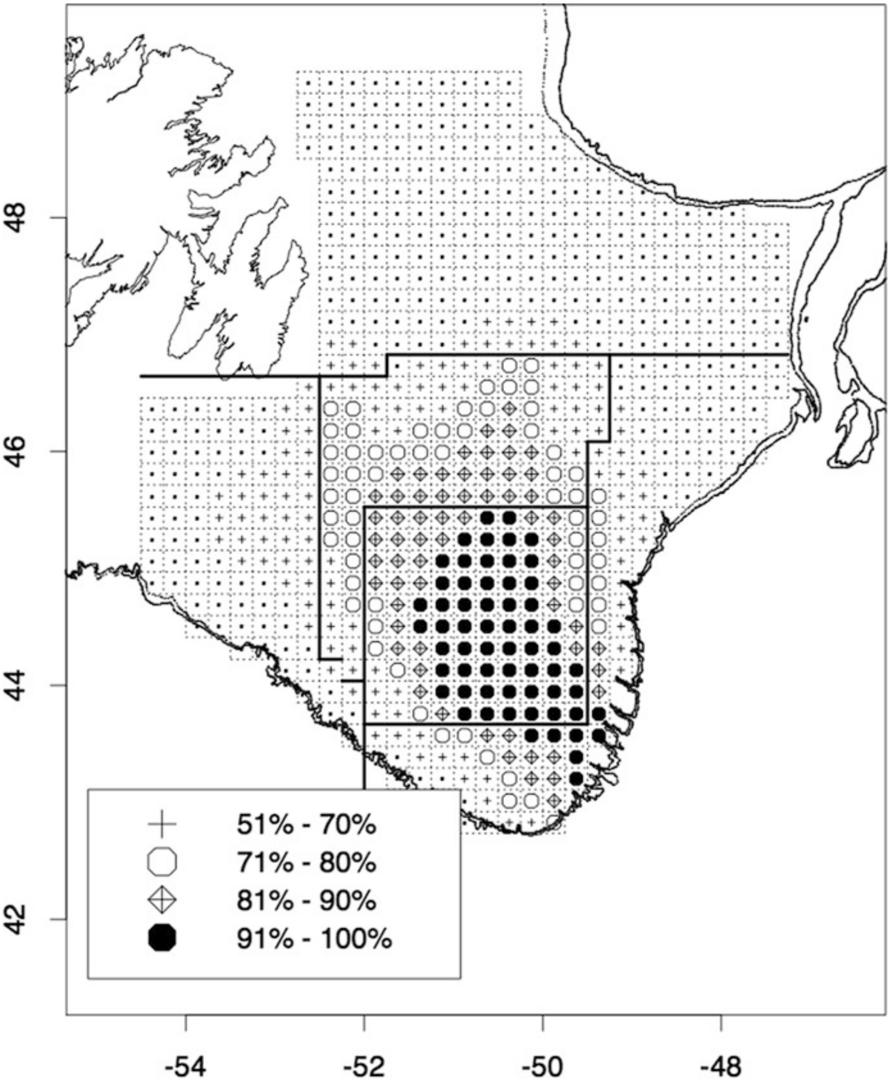

Fig. 14.1 Historical density of the fish population and sampling units for the Grand Bank area

sources. See Problem 3.5 for theoretical details on optimal sample size allocations based on an auxiliary variable.

Sampling costs are typically also considered in determining sample size allocation. The large stratum in the northern part of the region has low densities and is more expensive to sample in terms of distance to be travelled by the survey vessel. The optimal sample size allocation scheme described in Sect. 3.2.3 would recommend to assign much smaller sample sizes to those strata, where the initially assigned sample sizes are already small due to low abundance. The dynamics of the fish population and changes of environmental conditions over time, however, require that sufficient attention be given to less active areas and hence it is necessary to maintain certain minimum sample sizes for all strata. Further decrease of sample size in low abundance strata for cost considerations therefore is not recommended.

14.3 Estimation of Fish Abundance Indices

Let $\theta_N = I(\mathbf{R})$ be the fish abundance index defined by (14.1) for a particular fish species. Sample data from scientific trawl surveys have the format $\{(Y_i, \mathbf{x}_i, d_i), i \in \mathbf{S}\}$, where Y_i is the observed CPUE from grid square i and $\mathbf{x}_i$ is the vector of covariates such as the longitude and the latitude of the location, level of light during the tow, time of day, average depth of the tow, and other variables. The d_i are the survey weights. For stratified sampling, the stratum indicator variables are included as part of $\mathbf{x}_i$.

Estimation of $\theta_N = \sum_{i=1}^{N} \mu(\mathbf{x}_i)$ using survey data does not follow directly from standard design-based inference. It requires a model for the catching process and a joint randomization framework involving both the model and the survey design. Properties of point estimators and variance estimation need to be assessed under the joint framework. If the covariates $\mathbf{x}_i$ are available for all the grid squares, an alternative estimator of θ can be constructed through a model-based prediction approach.

14.3.1 Models for the Catching Process

The first required component of the joint framework is to specify a model for the catch data from the survey. The general form of the model for $(Y_i, \mathbf{x}_i)$ can be specified as

$$E_{\xi}\big(Y_i \mid \mathbf{x}_i\big) = \mu(\mathbf{x}_i) \quad \text{and} \quad V_{\xi}\big(Y_i \mid \mathbf{x}_i\big) = V_i\sigma^2 \,, \tag{14.2}$$

where σ^2 is a model parameter and V_i has a known form. Given all the $\mathbf{x}_i$, the Y_i's are assumed to be independent. This independence assumption is only required for variance estimation (Sect. 14.3.3). The mean function $\mu(\mathbf{x}_i)$, i.e., the expected CPUE, may be modeled parametrically or nonparametrically.

The observed catch per unit effort Y_i is measured either as the number of fish or the biomass of fish obtained per unit effort. The loglinear model is a suitable choice for this type of response which is nonnegative and has a right skewed distribution. We consider the following loglinear model (ξ):

$$E_\xi\big(Y_i \mid \mathbf{x}_i\big) = \mu(\mathbf{x}_i,\beta) = \exp\big(\mathbf{x}_i'\beta\big) \quad \text{and} \quad V_\xi\big(Y_i \mid \mathbf{x}_i\big) = \sigma^2\mu(\mathbf{x}_i,\beta)\,. \tag{14.3}$$

The mean function is specified by $\log\{\mu(\mathbf{x}_i,\beta)\} = \mathbf{x}_i'\beta$ with unknown model parameters β. It is assumed that $\mathbf{x}_i$ has 1 as its first component so that the model contains an intercept. The variance function $V_i = \mu(\mathbf{x}_i,\beta)$ works well for the loglinear model and σ^2 is called the overdispersion parameter. The survey weighted quasi maximum likelihood estimator $\hat{\beta}$ is the solution to the quasi score equations

$$\mathbf{Q}(\beta) = \sum_{i\in\mathbf{S}} d_i\mathbf{D}_i V_i^{-1}\big\{Y_i - \mu(\mathbf{x}_i,\beta\big\} = \mathbf{0}\,,$$

where $\mathbf{D}_i = \partial\mu(\mathbf{x}_i,\beta)/\partial\beta$. Noting that $\mathbf{D}_i = \mathbf{x}_i\{\mu(\mathbf{x}_i,\beta)\}$ and $V_i = \mu(\mathbf{x}_i,\beta)$, the quasi score equations can be re-written as

$$\mathbf{Q}(\beta) = \mathbf{X}'\mathbf{W}\big(\mathbf{Y} - \mu\big) = \mathbf{0}\,,$$

where $\mathbf{X} = (\mathbf{x}_1,\cdots,\mathbf{x}_n)'$, $\mathbf{W} = \text{diag}(d_1,\cdots,d_n)$, $\mathbf{Y} = (Y_1,\cdots,Y_n)'$, $\mu = (\mu_1,\cdots,\mu_n)'$, and $\mu_i = \mu(\mathbf{x}_i,\beta)$. The Newton-Raphson procedure given in (7.11) can be used to find the solution $\hat{\beta}$, with $\mathbf{G}_n(\theta)$ replaced by $\mathbf{Q}(\beta)$ and $\mathbf{H}_n(\theta)$ substituted by

$$\mathbf{H}(\beta) = -\mathbf{X}'\mathbf{W}\mathbf{V}\mathbf{X}\,,$$

where $\mathbf{V} = \text{diag}(\mu_1,\cdots,\mu_n)$. The fitted values for $\mu_i = \mu(\mathbf{x}_i,\beta)$ are given by $\hat{\mu}_i = \mu(\mathbf{x}_i,\hat{\beta}) = \exp(\mathbf{x}_i'\hat{\beta})$. A moment estimator for the overdispersion parameter σ^2 can then be obtained based on the fitted residuals and is given by $\hat{\sigma}^2 = (N - k)^{-1}\sum_{i\in\mathbf{S}} d_i(Y_i - \hat{\mu}_i)^2/\hat{\mu}_i$.

The spatial features of the trawl survey data over the open ocean region are shown to have the most significant impact on the model for the catching process. Let x_1 and x_2 be respectively the longitude and the latitude of the location of the grid square. It was shown by Chen et al. (2004) that x_1, x_2, x_1^2, x_2^2 and the interaction term $x_1 \times x_2$ are all significant under the loglinear model. Given that the spatial factors are included in the model, the temporal effect associated with the data collection becomes negligible. The only other variable found to be significant is x_3: level of light during the tow (recorded using a 0–9 scale). This observation from the model building process is quite crucial for some of the estimation methods to be discussed in the following sections, since both x_1 and x_2 are available for all the grid squares in the artificially created finite population.

Nonparametric smoothing techniques can be used to estimate the expected CPUE $\mu(\mathbf{x}_i)$. Chen et al. (2004) explored the use of a local linear smoothing estimator (Fan and Gijbels 1996) and showed that the method performs well with the FPI survey data.

14.3.2 Point Estimation

The fish abundance index $\theta_N = I(\mathbf{R}) = \sum_{i=1}^{N} \mu(\mathbf{x}_i)$ is the population total of the expected CPUE $\mu(\mathbf{x}_i)$ over the N grid squares. The observed response variables are the CPUE Y_i which are random variables under the catching process. The conventional Horvitz-Thompson estimator is computed as

$$\hat{\theta}_{\text{HT}} = \sum_{i \in \mathbf{S}} d_i Y_i \, . \tag{14.4}$$

The Horvitz-Thompson estimator is unbiased for θ_N under the joint randomization of the model (ξ) and the survey design (p) since

$$E\big(\hat{\theta}_{\text{HT}}\big) = E_p E_\xi \Big(\sum_{i \in \mathbf{S}} d_i Y_i \Big) = \sum_{i=1}^{N} \mu(\mathbf{x}_i) \, ,$$

and the statement holds for whatever the true model of the catching process.

A model-based prediction estimator of the abundance index can be constructed under an assumed model. The prediction approach, however, requires that the covariates $\mathbf{x}$ be available for all the grid squares in the finite population. Let $\mu_i = \mu(\mathbf{x}_i, \beta)$ be specified under the assumed model and $\hat{\beta}$ be an estimator based on the survey data $\{(Y_i, \mathbf{x}_i, d_i), i \in \mathbf{S}\}$. Let $\hat{\mu}_i = \mu(\mathbf{x}_i, \hat{\beta})$, $i = 1, \cdots, N$. The prediction estimator of θ_N is given by

$$\hat{\theta}_N = \sum_{i \in \mathbf{S}} Y_i + \sum_{i \notin \mathbf{S}} \hat{\mu}_i \, . \tag{14.5}$$

It is approximately unbiased under the assumed model since $E_\xi(\hat{\theta}_N - \theta_N) \doteq 0$, and the approximation amounts to replacing β by $\hat{\beta}$. The estimator $\hat{\theta}_N$ given in (14.5) becomes biased when the model ξ is misspecified. Chen et al. used a model involving the spatial covariates x_1, x_2, x_1^2, x_2^2, $x_1 \times x_2$ plus the variable x_3 on the level of light, and set $x_3 = 4.5$ (the median of the $0 - 9$ scale) for computing the predicted value $\hat{\mu}_i = \mu(\mathbf{x}_i, \hat{\beta})$ for $i \notin \mathbf{S}$.

The prediction estimator $\hat{\theta}_N$ given in (14.5) can alternatively be written as $\hat{\theta}_N = \sum_{i \in \mathbf{S}} Y_i - \sum_{i \in \mathbf{S}} \hat{\mu}_i + \sum_{i=1}^{N} \hat{\mu}_i$. A more efficient and robust estimator is the generalized difference estimator discussed in Sect. 5.4.3 and is given by

$$\hat{\theta}_{\text{GD}} = \sum_{i \in \mathbf{S}} d_i Y_i - \sum_{i \in \mathbf{S}} d_i \hat{\mu}_i + \sum_{i=1}^{N} \hat{\mu}_i \, . \tag{14.6}$$

The estimator $\hat{\theta}_{\text{GD}}$ is approximately unbiased under the assumed model for computing $\hat{\mu}_i$ and is also unbiased under the joint framework even if the model is misspecified. In addition, the estimator $\hat{\theta}_{\rm GD}$ is more efficient than the Horvitz-Thompson estimator $\hat{\theta}_{\rm HT}$ as shown in the next section on variance estimation.

14.3.3 Variance Estimation

The two sources of variation associated with the catching process and the sampling selection of grid squares induce two variance components for the estimated abundance indices. Let Var denote the overall variance under the joint ξp randomization. We have

$$\begin{aligned} Var\big(\hat{\theta}_{\rm HT}\big) &= V_\xi\big\{E_p\big(\hat{\theta}_{\rm HT}\big)\big\} + E_\xi\big\{V_p\big(\hat{\theta}_{\rm HT}\big)\big\} \\ &= V_\xi\Big(\sum_{i=1}^N Y_i\Big) + E_\xi\big\{V_p\big(\hat{\theta}_{\rm HT}\big)\big\}\,, \end{aligned}$$

where Y_i, $i \notin \mathbf{S}$ are conceptual values of the CPUE should the ith grid square be sampled. The second term, the design-based variance component, can be estimated by $v_p\big(\hat{\theta}_{\rm HT}\big)$ using a suitable design-based variance estimator for the Horvitz-Thompson estimator. The first term is the variance component due to the catching process, and estimation of the term requires an assumed model. Under the loglinear model (14.3),

$$V_\xi\Big(\sum_{i=1}^N Y_i\Big) = \sigma^2\sum_{i=1}^N \mu_i = \sigma^2\theta_N\,,$$

where $\theta_N = I(\mathbf{R})$ is the abundance index. An approximately unbiased variance estimator is given by

$$var\big(\hat{\theta}_{\rm HT}\big) = \hat{\sigma}^2\hat{\theta}_{\rm HT} + v_p\big(\hat{\theta}_{\rm HT}\big)\,.$$

It turns out that this is an example where the model component of the variance cannot be ignored, since the first term has order $O(N)$ and the second term is of order $O(N^2/n)$, and the sampling fraction satisfies $n/N = O(1)$ under the current survey design. The two terms have the same order of magnitude.

The variance for the generalized difference estimator $\hat{\theta}_{\rm GD}$ given in (14.6) can be derived using a similar decomposition of the total variance $Var(\hat{\theta}_{\rm GD}) = V_\xi\{E_p(\hat{\theta}_{\rm GD})\} + E_\xi\{V_p(\hat{\theta}_{\rm GD})\}$. Due to the difference structure, the estimation of the regression coefficients β has no impact on the asymptotic variance (Sect. 6.3.1). It follows that

$$Var(\hat{\theta}_{\rm GD}) \doteq V_\xi\Big(\sum_{i=1}^N Y_i\Big) + E_\xi\Big\{V_p\Big(\sum_{i\in\mathbf{S}} d_i\big(Y_i - \mu_i\big)\Big)\Big\}\,.$$

The model component of the variance is the same as the one in $Var(\hat{\theta}_{\text{HT}})$ but the design-based variance component is based on the residuals $\varepsilon_i = Y_i - \mu_i$ instead of the responses Y_i. Consequently, we have $Var(\hat{\theta}_{\text{GD}}) < Var(\hat{\theta}_{\text{HT}})$ if the model provides reasonable fit to the survey data. An approximately unbiased variance estimator for $\hat{\theta}_{\text{GD}}$ can be constructed as $var(\hat{\theta}_{\text{GD}}) = \hat{\sigma}^2\hat{\theta}_{\text{GD}} + v_p(\hat{\theta}_{\text{GD}})$, with $v_p(\hat{\theta}_{\text{GD}})$ computed based on the fitted residuals $\hat{\varepsilon}_i = Y_i - \hat{\mu}_i$ for the Horvitz-Thompson estimator.

14.3.4 Pseudo Empirical Likelihood Methods

The FPI fish abundance survey uses a stratified sampling design. For ease of computation, the pseudo empirical likelihood function can be formulated using the single combined sample through first order inclusion probabilities, as shown below. The stratified design feature is taken into account in computing the design effect or the estimated variance. One important observation is that the population size N (i.e., the total number of grid squares) is known for fish abundance surveys, and therefore inferences on the population total and the population mean become equivalent problems.

Let n be the overall sample size. Let $\mathbf{p} = (p_1, \cdots, p_n)$ be a discrete probability measure over the n sampled units. The pseudo empirical log-likelihood function is defined as

$$\ell(\mathbf{p}) = n \sum_{i \in \mathbf{S}} \tilde{d}_i(\mathbf{S}) \log(p_i) ,$$

where $\tilde{d}_i(\mathbf{S}) = d_i / \sum_{j \in \mathbf{S}} d_j$. The stratified structure of the sample is ignored at this stage and the normalization constraint is given by $\sum_{i \in \mathbf{S}} p_i = 1$.

Auxiliary information from the current year survey data can be incorporated through a suitable model on the response variable CPUE. Let $\hat{\mu}_i = \mu(\mathbf{x}_i, \hat{\beta})$ be the fitted values under the model (14.3), $i = 1, \cdots, N$. The model-calibration constraint is specified as

$$\sum_{i \in \mathbf{S}} p_i \hat{\mu}_i = \frac{1}{N} \sum_{i=1}^{N} \hat{\mu}_i . \tag{14.7}$$

Noting that the fish abundance index is defined as a population total, the model-calibrated maximum pseudo empirical likelihood estimator of $\theta_N = I(\mathbf{R})$ is computed as

$$\hat{\theta}_{\text{PEL}} = N \sum_{i \in \mathbf{S}} \hat{p}_i Y_i ,$$

where the $\hat{p}_i$ maximize $\ell(\mathbf{p})$ subject to constraints $\sum_{i\in\mathbf{S}} p_i = 1$ and (14.7). It can be shown that

$$\hat{\theta}_{\text{PEL}} = N\Big[\sum_{i\in\mathbf{S}} \tilde{d}_i(\mathbf{S})Y_i + B\Big\{\eta_N - \sum_{i\in\mathbf{S}} \tilde{d}_i(\mathbf{S})\mu_i\Big\}\Big] + o_p\big(Nn^{-1/2}\big),$$

where $\mu_i = \mu(\mathbf{x}_i, \beta)$, $\eta_N = \theta_N/N = N^{-1}\sum_{i=1}^N \mu_i$ is the mean value of grid abundance index, and

$$B = \Big\{\sum_{i=1}^{N} \big(\mu_i - \eta_N\big)Y_i\Big\} \Big/ \Big\{\sum_{i=1}^{N} \big(\mu_i - \eta_N\big)^2\Big\}. \tag{14.8}$$

The quantity B is the conceptual population regression coefficient (slope) when Y is regressed on μ, corresponding to the sample version

$$\hat{B} = \Big\{\sum_{i\in\mathbf{S}} \tilde{d}_i(\mathbf{S})\big(\hat{\mu}_i - \hat{\eta}_N\big)Y_i\Big\} \Big/ \Big\{\sum_{i\in\mathbf{S}} \tilde{d}_i(\mathbf{S})\big(\hat{\mu}_i - \hat{\eta}_N\big)^2\Big\},$$

where $\hat{\eta}_N = N^{-1}\sum_{i=1}^N \hat{\mu}_i$. Under stratified simple random sampling, we have $\sum_{i\in\mathbf{S}} d_i = N$, $\tilde{d}_i(\mathbf{S}) = d_i/N$, and

$$\hat{\theta}_{\text{PEL}} = \sum_{i\in\mathbf{S}} d_i Y_i + B\Big\{\sum_{i=1}^{N} \mu_i - \sum_{i\in\mathbf{S}} d_i\mu_i\Big\} + o_p\big(Nn^{-1/2}\big).$$

The model-calibrated maximum pseudo empirical likelihood estimator $\hat{\theta}_{\text{PEL}}$ is asymptotically equivalent to the generalized difference estimator $\hat{\theta}_{\text{GD}}$ given in (14.6) when the model is correctly specified but is more efficient otherwise. This can be seen from the fact that $E_\xi(B) = 1$ under the model and $\hat{\theta}_{\text{PEL}}$ is a regression-type estimator when the model is misspecified. See Problem 5.9 on the optimality of the generalized regression estimator and Sect. 6.3.1 on the optimality of the model-calibration estimator. The variance formula is given by

$$Var(\hat{\theta}_{\text{PEL}}) \doteq V_\xi\Big(\sum_{i=1}^{N} Y_i\Big) + E_\xi\Big\{V_p\Big(\sum_{i\in\mathbf{S}} d_i\big(Y_i - B\mu_i\big)\Big)\Big\}. \tag{14.9}$$

The pseudo empirical likelihood approach has two major advantages. The first is the flexibility in incorporating auxiliary information, including the current year survey data and historical catch data over the same region, through additional constraints. The second is the property that the pseudo empirical likelihood ratio confidence intervals for the abundance indices have more desirable properties for species with low abundance or where the observed CPUE contains many zeros. The arguments are similar to those in the examples presented in Sect. 8.2.3. We consider the mean value of the grid abundance index $\eta_N = \theta_N/N = N^{-1}\sum_{i=1}^N \mu_i$. Let $\hat{\mathbf{p}}$ be the maximizer of $\ell(\mathbf{p})$ under the normalization constraint $\sum_{i\in\mathbf{S}} p_i = 1$ and the model-calibration constraint (14.7); let $\hat{\mathbf{p}}(\eta)$ maximize $\ell(\mathbf{p})$ subject to $\sum_{i\in\mathbf{S}} p_i = 1$, the constraint (14.7) and the parameter constraint

$$\sum_{i\in\mathbf{S}} p_i Y_i = \eta$$

for a fixed η. The pseudo empirical log-likelihood ratio function is given by

$$r(\eta) = \ell(\hat{\mathbf{p}}(\eta)) - \ell(\hat{\mathbf{p}})\,.$$

It follows from technical arguments similar to those detailed in Problem 8.2 that the adjusted pseudo EL ratio function $-2r(\eta)/a$ has an asymptotic χ^2 distribution with one degree of freedom when $\eta = \eta_N$. The adjustment factor is defined as $a = Var(\hat{\eta}_{\text{PEL}})/(\sigma_\varepsilon^2/n)$, where $Var(\hat{\eta}_{\text{PEL}}) = N^{-2}Var(\hat{\theta}_{\text{PEL}})$, with $Var(\hat{\theta}_{\text{PEL}})$ given in (14.9),

$$\sigma_\varepsilon^2 = \frac{1}{N-1}\sum_{i=1}^N \varepsilon_i^2 \qquad \text{with} \qquad \varepsilon_i = Y_i - \mu_Y - B(\mu_i - \eta_N)\,,$$

the quantity B is defined by (14.8), and $\mu_Y = N^{-1}\sum_{i=1}^N Y_i$ is defined over the conceptual population values of CPUE.

Let $\hat{a}$ be a suitable estimate of the adjustment factor a. The $(1-\alpha)$-level confidence interval for η_N is computed as $(\hat{\eta}_1, \hat{\eta}_2) = \{\eta \mid -2r(\eta)/\hat{a} \le \chi_1^2(\alpha)\}$. The $(1-\alpha)$-level confidence interval for the abundance index θ_N is obtained as $(N\hat{\eta}_1, N\hat{\eta}_2)$.

Chapter 15
Adaptive and Network Surveys

In this chapter, the units of a finite survey population form a network, potentially linked to one another in a way that facilitates sampling and observation; this can be especially useful when the sampling frame exists only partially. For example, members of a population of intravenous drug users can be sampled through the acquaintanceship networks of users who come to a safe injection site; in this case, two members of the population are linked if they are acquainted. For another example, wildlife observations are often carried out in randomly selected areas together with neighbouring areas linked to the initially selected areas through proximity (Thompson and Seber 1996). Network surveys have much in common with indirect sampling procedures (Lavallée 2007), wherein members of a population without a current frame can be surveyed through sampling from a related listed population.

What these examples have in common is a *graph* structure, where the population units correspond to *vertices* and the links correspond to *edges*. The *neighbours* of a unit are the other units to which it is directly linked; they may be spatially near one another, in the sense of neighbouring areas, or only conceptually close. A sampling design that chooses an initial sample, and follows the links of its members to add to the sample, is termed a *link tracing design*.

In examples like the first one above, it may be challenging to make inferences to the population because the users who come to the safe injection site are not a random sample of users, and a sample from these users and their network neighbours is not a probability sample. However, if the initial sample IS selected randomly, the link-tracing design can be a probability sampling design with inclusion probabilities that are quantifiable in principle, given sufficient information about the network.

The design-based theory of network sampling is conceptually straightforward when the aim is to estimate a population mean or total of a variable y taking a value for each node. However, we may also interested in situations where the variable values for linked units are somehow related. For example, the values for close neighbours may be more closely correlated than the values for distant neighbours or

C. Wu, M. E. Thompson, *Sampling Theory and Practice*, ICSA Book Series in Statistics, https://doi.org/10.1007/978-3-030-44246-0_15

non-neighbours. The property that linkage is more likely between nodes with like characteristics is called *homophily*. In homophily studies we tend to be interested in characteristics of linked pairs, or pairs of neighbours.

15.1 Finite Populations with Network Structure

This section introduces the theoretical framework for network sampling.

15.1.1 Terminology and Notation

Formally, a *network* is a mathematical object that consists of a set of units referred to as *nodes*, and a set E of *links* (or *edges* in the terminology of graph theory) between the nodes. A finite population $\mathbf{U}$ of size N is usefully regarded as a network if the units are nodes, and an edge linking two nodes corresponds to a specific relationship of interest between the nodes. In this chapter we consider cases where the graph of the network is *undirected*, so that the collection E of edges can be represented as a set of pairs $\{i, j\}$ of unit labels. For example, a population of N scientists could form a network where an edge between researchers i and j would signify that they had been co-authors of a published paper. A spatial network might be represented by a square lattice in the plane, with edges of the form $\{(i_1, i_2), (j_1, j_2)\}$ where $\mid j_1 - i_1 \mid + \mid j_2 - i_2 \mid = 1$.

The *degree* k_i of a node i is the number of edges in which i participates. The population *mean degree*

$$\mu_K = \frac{1}{N} \sum_{i=1}^{N} k_i$$

is a simple measure of the connectivity of the network. More generally, the *degree distribution* $\{f(k) : k = 0, 1, \ldots\}$, where $f(k) =$ (number of nodes with degree $k)/N$, may be of interest to study. The *joint degree distributions* are the distributions of the degrees of randomly selected pairs and, more generally, r-tuples of nodes.

The *distance* $d(i, j)$ from node i to node j is the minimum number of edges in a connected path from i to j in the network. The d-th *order neighbourhood* of a node i is the set of nodes j within a distance d of i.

A network or graph is termed *connected* if for any pair of nodes i and j in the network there exists a connected path from i to j. A node is *isolated* if it participates in no edges. In general, a network is the union of disjoint *connected components* and isolated points.

15.1.2 Estimation Problems

Continuing to consider a population $\mathbf{U}$ which has a network structure, let y_i, x_i be values for unit i of a variable of interest y and covariate or auxiliary variable x respectively. The basic sampling estimation problem is to estimate the population total or mean of the variable y, given a probability sample from the nodes of the population, and auxiliary information. Estimating the mean degree or the degree distribution of the nodes themselves are special cases. The objects of estimation can be more complex connectivity measures, such as the number of *triangles* (where a triangle is a set of three nodes all connected by edges) existing in the population. It might also be of interest to estimate the population ratio of y to x, as in traditional sampling theory. There might be questions about network-related relationships to consider, such as a measure of homophily, the tendency of linking to depend on a characteristic of interest. A simple example would be the proportion of linked pairs among pairs $\{i, j\}$with both members belonging to a domain D, in comparison with what might be expected if population members were linked at random; the characteristic of interest y in this case would be the indicator variable for D. For quantities involving network relationships to be estimated, it is necessary to observe not only sampled y and x values, but also information on connections among sampled nodes.

In general, with a probability sampling design, the estimators of the quantities can be based on the construction of the Horvitz-Thompson estimator (for totals) and the Hájek estimator (for means). For example, we might expect the estimator of the mean degree from a probability sample to take the form

$$\frac{1}{N}\sum_{i\in\mathbf{S}}\frac{k_i}{\pi_i} \quad \text{or} \quad \sum_{i\in\mathbf{S}}\frac{k_i}{\pi_i}\Big/\sum_{i\in\mathbf{S}}\frac{1}{\pi_i}\,.$$

We might similarly expect estimators of sums of variables belonging to pairs of units to take the form of

$$\sum_{i\in\mathbf{S}}\ \sum_{j>i,j\in\mathbf{S}}\frac{z_{ij}}{\pi_{ij}}$$

where z_{ij} is defined appropriately. For link tracing designs, a theory more elaborate than traditional sampling theory is needed.

15.1.3 Probability Models for Networks

For some theoretical purposes, we assume a probability model for the finite population network. Usually, the nodes are considered to be given entities with names or *labels* (although in large sample theory the structure is assumed to be evolving), and the model being what generates the links. A simple model for links is the Erdös-Rényi random graph, in which each pair of vertices is independently connected (linked) with probability p, and not connected with probability $1 - p$. If the number of nodes is N, the degree of a node is binomial $(N - 1, p)$; if $N \to \infty$ and $p \to 0$ with Np tending to a finite positive λ, the degree distribution tends to a Poisson distribution with mean λ for every node. In the limit, the finite population mean degree also tends to λ. In this model, clustering of edges is minimal in the sense that edge existence is independent from pair to pair of nodes.

The Erdös-Rényi graph is *sparse* in the sense of exhibiting a vanishing fraction of all possible edges as $N \to \infty$ as long as $p \to 0$, a weaker condition than Np tending to finite positive λ. See Veitch and Roy (2019) for a treatment of sampling and estimation on sparse random graphs.

In approximating inclusion probabilities for nodes under link tracing designs, it is useful to be able to assume that, for large networks, short *cycles* (closed pathways) are rare in the d-th order neighbourhood of a collection of seeds. In an Erdös-Rényi model, this assumption requires a more stringent restriction than sparsity on the behaviour of the mean degree as $N \to \infty$; the restriction is satisfied when Np remains bounded as $N \to \infty$. The shortest cycle of a graph is termed its *girth*. Some theory of cycles in random graphs can be found in the monograph by Frieze and Karoński (2016).

Graph theorists have been interested in defining network probability models with non-Poisson degree distributions, because in many real-life networks, the degree distribution seems to have heavier tails than a Poisson distribution. The *configuration model* of Molloy and Reed (1995) is a construction algorithm that produces a finite graph with a given degree sequence. Like the Erdös-Rényi model, the configuration model is *exchangeable* in the sense that the probability of any resulting graph is invariant under permutation of the labels of the nodes. The configuration model shares other important properties of the Erdös-Rényi model.

Other network probability models relevant to the sampling of networks include models for how the connection probabilities depend on characteristics of the nodes or pairs or sets of nodes. An important class of models in this category is that of Exponential Random Graph Models, as described by Snijders et al. (2006) and Handcock and Gile (2010).

15.2 Link Tracing Sampling Designs

In this section, we consider several important examples of link tracing designs. We begin with designs for the sampling of individual units, in particular snowball sampling and the closely related Respondent Driven Sampling. We then look at two methods of sampling clusters of hard-to-reach or rare units, namely indirect sampling and adaptive cluster sampling.

15.2.1 Snowball Sampling

In the simplest probability form of *snowball sampling* (Frank and Snijders 1994), a simple random sample without replacement $\mathbf{S}_0$ of n_0 *seeds* is selected, and the links emanating from each seed are followed, with the y, x and k values being observed for every node visited in a d-th order neighbourhood of the seed, for a suitable pre-determined value of d. A node may be included in the sample as a seed or as a *non-seed*, and in principle may be included multiple times. However, if N is relatively large, and if the graph of the network has sufficiently large girth, we can neglect the multiple inclusion possibility. The probability that a node i is included as a seed is n_0/N, and if d is equal to 1, the probability that i is included as a non-seed is the probability that at least one of its neighbours is a seed, and is approximately equal to $k_i n_0/N$. Thus in a snowball sample, the inclusion probability π_i of node i is approximately $(1 + k_i)n_0/N$. If this approximation is close, we may construct a Horvitz- Thompson estimator of the population mean of y as follows:

$$\hat{\mu}_{y\text{HT}} = \sum_{i \in \mathbf{S}} \frac{y_i}{n_0(1 + k_i)} .$$

An alternative is an estimator suggested by Thompson et al. (2016), namely a ratio combination of the sample mean of y over the sample of seeds, and the Horvitz-Thompson mean estimator from the sample of non-seeds, adjusted for the fact that the latter estimates the mean degree over nodes that have at least one neighbour.

Still considering the case $d = 1$, and assuming the girth of the network graph is greater than 4, the joint inclusion probabilities of i and j can be approximated as follows. Problem 15.1 is to fill in the details.

(i) If i and j are linked,

$$\pi_{ij} \doteq \frac{2n_0}{N} + \frac{(k_i k_j - 1)n_0(n_0 - 1)}{N(N - 1)} .$$

(ii) If i and j have a neighbour ℓ in common,

$$\pi_{ij} \doteq \frac{n_0}{N} + \frac{(N-n_0)n_0(n_0-1)}{N(N-1)(N-2)} + \frac{(N-n_0)(N-n_0-1)n_0(n_0-1)(k_i+k_j-2)}{N(N-1)(N-2)(N-3)}$$
$$+ \frac{(N - n_0)(N - n_0 - 1)(N - n_0 - 2)(k_i - 1)(k_j - 1)}{N(N - 1)(N - 2)(N - 3)(N - 4)} .$$

(iii) if i and j are not linked and have no neighbours in common,

$$\pi_{ij} \doteq \frac{n_0(n_0 - 1)}{N(N - 1)} + \frac{n_0(N - n_0 - 1)(k_i + k_j)}{N(N - 1)(N - 2)} + \frac{(N - n_0)(N - n_0 - 1)k_i k_j}{N(N - 1)(N - 2)(N - 3)} .$$

15.2.2 Respondent Driven Sampling

Respondent Driven Sampling (RDS) was introduced by sociologist Heckathorn (1997). Respondents at each "wave" are asked to select the sample at the next wave by distributing coupons to a certain number of others (to whom they are linked by acquaintance) in the target population; these others can then choose to participate in the study, and to recruit further respondents in turn. The intuition was that after many waves of sampling the dependence of the final sample on the initial sample would be reduced, and that the final sample of respondents would approximate a probability sample. The paper by Gile (2011) provides a justification in terms of RDS on a random graph, modifying a method of Volz and Heckathorn (2008). In the latter paper, which considers that each respondent recruits just one other, the inclusion probability of a sampled unit is approximated as a constant times its degree k_i in the acquaintance network, based on treating the sample as independent draws from the stationary distribution of a random walk on the nodes of the network. The estimator of the mean of a variable y is then given by the Horvitz-Thompson type estimator

$$\hat{\mu}_{y\mathrm{VH}} = \sum_{i \in \mathbf{S}} \frac{y_i}{k_i} \Big/ \sum_{i \in \mathbf{S}} \frac{1}{k_i} \,.$$

Gile (2011) assumes the graph to be generated from the configuration model of Molloy and Reed (1995), and considers the transition probabilities of the corresponding walk in which the j-th node visited, $\mathbf{G}^*_j$ is selected from the distribution of possible edges from node $\mathbf{G}^*_{j-1}$. The transition probabilities, taken over the space of all possible configuration model graphs of given degree distribution, are very nearly proportional to the degree of the destination node. The analogue of the Volz-Heckathorn estimator is a Hansen and Hurwitz (1943) estimator. The corresponding self-avoiding random walk, corresponding to without replacement sampling, is also considered.

15.2.3 Indirect Sampling

Indirect sampling (Lavallée 2007) aims at producing an estimate for a population $\mathbf{U}^B$ by selecting a sample from a population $\mathbf{U}^A$ for which a sampling frame is available, and by using the links between the two populations. In the simplest setting, population $\mathbf{U}^B$ is assumed to consist of predefined clusters of units. For example, $\mathbf{U}^B$ may be a population of households (clusters of individuals) for which no detailed sampling frame exists, while $\mathbf{U}^A$ is a list of individuals with links to members of the households. We assume that in each of the clusters of $\mathbf{U}^B$, at least one member is linked to some individual in $\mathbf{U}^A$.

A sample $\mathbf{S}^A$ is taken from $\mathbf{U}^A$, in a probability sampling design with inclusion probabilities π^A_j. If a unit in cluster i of $\mathbf{U}^B$ is linked to a unit j of the $\mathbf{S}^A$ sample, every unit of cluster i in $\mathbf{U}^B$ will be surveyed.

Suppose we are interested in estimating the population total T_y^B, with an estimator of the form

$$\hat{T}_y^B = \sum_{i \in \mathbf{S}^B} w_i \sum_{r \in \mathbf{U}_i^B} y_{ir}, \tag{15.1}$$

where $\mathbf{U}_i^B$ denotes the i-th cluster in $\mathbf{U}^B$ and $\mathbf{S}^B$ is the sample of clusters that have members linked to units in $\mathbf{S}^A$. A solution is to use what is called the *Generalized Weight Share Method (GWSM)*, as follows.

Define the indicator $I_{j;i,r} = 1$ if there is a link between unit j of $\mathbf{U}^A$ and unit r of cluster i of $\mathbf{U}^B$, and $I_{j;i,r} = 0$ otherwise. Assuming there may be more than one unit r of cluster i linked to unit j of $\mathbf{U}^A$, let

$$L_{j;i}^B = \sum_{r \in \mathbf{U}_i^B} I_{j;i,r}$$

be the number of ways of accessing cluster i through j. Let

$$L_i^B = \sum_{j \in \mathbf{S}^A} L_{j;i}^B$$

be the number of ways of accessing cluster i through $\mathbf{S}^A$. Thus the *share* of the access to i carried by j can be regarded as

$$\alpha_{j;i} = \frac{L_{j;i}^B}{L_i^B}.$$

An unbiased estimator $\hat{T}_y^B$ from the perspective of the sampling design in $\mathbf{U}^A$ is

$$\hat{T}_y^B = \sum_{j \in \mathbf{S}^A} \frac{1}{\pi_j^A} \sum_{i \in \mathbf{S}^B} \alpha_{j;i} \sum_{r \in \mathbf{U}_i^B} y_{ir}\,; \tag{15.2}$$

this is in the form (15.1), where

$$w_i = \sum_{j \in \mathbf{S}^A} \frac{1}{\pi_j^A} \alpha_{j;i}\,.$$

However, it is also a Horvitz-Thompson type estimator in terms of the $\mathbf{U}^A$ sampling design.

As shown by Lavallée (2007), the theory of Horvitz-Thompson estimation can be used to develop estimators of variance for the weight share estimator 15.2 of a population total. As well, the concept of sharing the weight of an initial sample member among members of another population that are linked to it is useful in many contexts, such as longitudinal household surveys where the composition of a household can change over time (Lavallée 1995).

15.2.4 Adaptive Cluster Sampling

For another kind of network cluster sampling, consider a network of size N with a definition of *neighbourhood* in terms of a linkage relationship, e.g. units i, i' are neighbours of each other if they are linked by a path. Suppose the y values of main impact on the mean belong to units in a domain D of the population network, of originally unknown membership; perhaps D is the set of units for which y satisfies a certain condition, such as y having a value above a certain threshold. Suppose that there is a kind of homophily present, in that members of D tend to be close together or linked, so that following links allows them to be more readily located. We want to estimate the overall population mean μ_y.

Adaptive cluster sampling (Thompson 1990, 1991; Thompson and Seber 1996) is a link tracing design akin to snowball sampling. An initial sample $\mathbf{S}_0$ of n_0 seeds is taken. If a seed i belongs to D, its neighbours are added to the sample, and if any of the neighbours belongs to D, its neighbours are also added, and so on. The sample generated from the seed i is the set of all units that can be reached by a path from i, where the path stops at the first node that does not belong to D. If a seed i does not belong to D, it remains a single point in the sample. We can then think of the connected components of D as disjoint clusters; the set of sampled members of D is the union of those clusters that each contain at least one seed. A seed that does not belong to D does not generate further sample membership. The sample contains also some non-seeds that do not belong to D, those that are the stopping points of paths followed from members of D.

If the seeds are sampled using simple random sampling with replacement from the population network, the mean of y over the sample of seeds will be an unbiased estimator of μ_y. However, we would like to make use also of the values of y in all the points of the sampled clusters. Consider a seed k; let C_k be the (connected) cluster that includes that seed, and let m_k be the size of the cluster. (If the seed k is not in the domain D, C_k will have size 1.) Let $\bar{y}_k^*$ be the sample average of y in C_k, that is,

$$\bar{y}_k^* = \frac{\sum_{i \in C_k} y_i}{m_k}.$$

Then a *modified Hansen-Hurwitz estimator* for the mean μ_y is

$$\hat{\mu}_{y\text{HH}} = \frac{1}{n_0} \sum_{k=1}^{n_0} \bar{y}_k^*,$$

and an unbiased estimator of its variance is

$$\frac{1}{n_0(n_0-1)}\sum_{k=1}^{n_0}(\bar{y}_k^* - \hat{\mu}_{y\mathrm{HH}})^2 \,.$$

There are analogous forms for when the initial sample is selected without replacement. Note that the estimator $\hat{\mu}_{y\mathrm{HH}}$ does not use units in the sample that are non-seeds that do not belong to D. The estimator is simple but suboptimal (Thompson 1990).

15.3 Problem

15.1 Verify the formulae at the end of Sect. 15.2.1 for the joint inclusion probabilities in a scheme of snowball sampling with order $d = 1$. Derive an approximation for this scheme replacing $N - a$ by N for small values of a.

Chapter 16
Dual Frame and Multiple Frame Surveys

Most survey samples are selected from a single sampling frame presumably covering all of the units in the target population. *Multiple frame sampling* refers to surveys where two or more frames are used and independent samples are taken respectively from each of the frames. Inferences about the target population are based on the combined sample data. The method is referred to as *dual frame sampling* when the survey uses two frames. One of the assumptions with dual frame or multiple frame surveys is that one or several of the frames are incomplete but together they cover the entire target population.

Sampling designs are often dictated by several key factors, including the target population and parameters of interest, the population frame or frames for sampling selection of units, the mode of data collection, inference tools available for analyzing data under the chosen design, and the total cost. There are two major motivations behind the use of dual frame or multiple frame sampling methods: (i) to achieve a desired level of accuracy with reduced cost, and (ii) to have a better coverage of the target population and hence to reduce biases due to coverage errors. Even if a complete frame such as a household address list is available, it is often more cost effective to take a sample of reduced size from the complete frame and supplement the sample by additional data taken from other frames such as telephone directories or institutional lists which might be incomplete but less expensive to sample from. For surveys of human populations where the goal is to study special characteristics of individuals, such as persons with certain rare diseases, a sample taken from the frame for general population health surveys is usually not very informative. Other frames, such as lists of general hospitals and/or special treatment centers, often provide more informative data as well as extended coverage of the target population (Kalton and Anderson 1986).

Practical issues in the design of a dual frame or multiple frame survey are highly related to the characteristics of the target population and the availability of sampling frames. This chapter focuses on issues with estimation using data from dual or multiple frame surveys. Pseudo empirical likelihood methods and

C. Wu, M. E. Thompson, *Sampling Theory and Practice*, ICSA Book Series in Statistics, https://doi.org/10.1007/978-3-030-44246-0_16

the concept of *multiplicity* are shown to be very useful for point estimation and confidence intervals, especially for estimating population means of binary study variables such as the prevalence of a disease or the proportion of the population with a rare condition.

16.1 Estimation with Dual Frame Surveys

Estimation with dual frame surveys was first studied by Hartley (1962, 1974) based on post-stratified samples under the following classic setting. The same setting was also used by Fuller and Burmeister (1972), Skinner and Rao (1996), Lohr and Rao (2000, 2006), and Rao and Wu (2010b), among others.

Let A and B denote two sampling frames. Both frames can be incomplete but together they cover the entire finite population, $\mathbf{U}$. Let $\mathbf{A}$ be the set of population units covered by frame A and $\mathbf{B}$ be the set of population units in frame B. The population of interest, $\mathbf{U}$, may be divided into three mutually exclusive domains

$$a = \mathbf{A} \cap \mathbf{B}^c , \qquad b = \mathbf{A}^c \cap \mathbf{B} \qquad \text{and} \qquad ab = \mathbf{A} \cap \mathbf{B}$$

such that

$$\mathbf{U} = \mathbf{A} \cup \mathbf{B} = a \cup b \cup ab .$$

Note that $\mathbf{A}^c$ and $\mathbf{B}^c$ denote the complement sets of $\mathbf{A}$ and $\mathbf{B}$. The domain ab is the common part covered by both frames. Let N, N_A, N_B, N_a, N_b and N_{ab} be the number of population units in $\mathbf{U}$, $\mathbf{A}$, $\mathbf{B}$, a, b and ab, respectively. We have

$$N_A = N_a + N_{ab} , \qquad N_B = N_b + N_{ab}$$

and

$$N = N_a + N_b + N_{ab} = N_A + N_B - N_{ab} .$$

Let μ_y, μ_{ya}, μ_{yb} and μ_{yab} denote respectively the population or the domain mean of the study variable y for $\mathbf{U}$, a, b and ab. It follows that

$$\mu_y = W_a \mu_{ya} + W_b \mu_{yb} + W_{ab} \mu_{yab} , \tag{16.1}$$

where $W_a = N_a/N$, $W_b = N_b/N$ and $W_{ab} = N_{ab}/N$. The main objective is to estimate μ_y using dual frame survey samples as well as auxiliary information that is available. Let $\mathbf{S}_A$ be a sample of size n_A taken from frame A with first order inclusion probabilities $\pi_i^A = P(i \in \mathbf{S}_A)$; let $\mathbf{S}_B$ be a sample of size n_B selected from frame B with first order inclusion probabilities $\pi_i^B = P(i \in \mathbf{S}_B)$. The study variable y is observed for both samples. The two survey samples $\mathbf{S}_A$ and $\mathbf{S}_B$ are independent.

Estimation based on post-stratified samples requires a critical assumption: the frame membership for each sampled units can be correctly identified. Under this assumption, the frame A sample $\mathbf{S}_A$ can be post-stratified as $\mathbf{S}_A = \mathbf{S}_a \cup \mathbf{S}_{ab}$ over the two domains a and ab, where $\mathbf{S}_a = a \cap \mathbf{S}_A$ and $\mathbf{S}_{ab} = (ab) \cap \mathbf{S}_A$. Similarly, the frame B sample $\mathbf{S}_B$ can be post-stratified as $\mathbf{S}_B = \mathbf{S}_b \cup \mathbf{S}_{ba}$ over the two domains b and ab, where $\mathbf{S}_b = b \cap \mathbf{S}_B$ and $\mathbf{S}_{ba} = (ab) \cap \mathbf{S}_B$. Note that both $\mathbf{S}_{ab}$ and $\mathbf{S}_{ba}$ are from the common domain ab, but $\mathbf{S}_{ab}$ is part of the frame A sample and $\mathbf{S}_{ba}$ is part of the frame B sample. This notation differs from $\mathbf{S}'_{ab}$ and $\mathbf{S}''_{ab}$ used by Hartley (1962) and other authors. Also note that the notation $ba = \mathbf{B} \cap \mathbf{A}$ would be identical to $ab = \mathbf{A} \cap \mathbf{B}$.

A dual frame estimator of μ_y can be constructed based on (16.1) and the post-stratified samples. The population means μ_{ya} and μ_{yb} for domains a and b can be estimated by using the samples $\mathbf{S}_a$ and $\mathbf{S}_b$ and the Hájek estimators

$$\hat{\mu}_{ya\text{H}} = \sum_{i \in \mathbf{S}_a} \tilde{d}_i(\mathbf{S}_a) y_i \qquad \text{and} \qquad \hat{\mu}_{yb\text{H}} = \sum_{i \in \mathbf{S}_b} \tilde{d}_i(\mathbf{S}_b) y_i \,, \tag{16.2}$$

where $\tilde{d}_i(\mathbf{S}_a) = d_i^A / \hat{N}_a$, $\tilde{d}_i(\mathbf{S}_b) = d_i^B / \hat{N}_b$, $\hat{N}_a = \sum_{i \in \mathbf{S}_a} d_i^A$, $\hat{N}_b = \sum_{i \in \mathbf{S}_b} d_i^B$, with the design weights $d_i^A = 1/\pi_i^A$ and $d_i^B = 1/\pi_i^B$ for each of the two frames.

For the domain population mean μ_{yab}, there are two Hájek estimators available based on $\mathbf{S}_{ab}$ and $\mathbf{S}_{ba}$:

$$\hat{\mu}_{yab\text{H}} = \sum_{i \in \mathbf{S}_{ab}} \tilde{d}_i(\mathbf{S}_{ab}) y_i \qquad \text{and} \qquad \hat{\mu}_{yba\text{H}} = \sum_{i \in \mathbf{S}_{ba}} \tilde{d}_i(\mathbf{S}_{ba}) y_i \,, \tag{16.3}$$

where $\tilde{d}_i(\mathbf{S}_{ab}) = d_i^A / \hat{N}_{ab}$, $\tilde{d}_i(\mathbf{S}_{ba}) = d_i^B / \hat{N}_{ba}$, $\hat{N}_{ab} = \sum_{i \in \mathbf{S}_{ab}} d_i^A$, $\hat{N}_{ba} = \sum_{i \in \mathbf{S}_{ba}} d_i^B$. Note that $\hat{N}_{ab}$ and $\hat{N}_{ba}$ both are estimators for N_{ab} but use samples from different frames. A popular strategy for estimating μ_{yab} with data from the combined sample $\mathbf{S}_{ab} \cup \mathbf{S}_{ba}$ is to use (Fuller and Burmeister 1972; Hartley 1974)

$$\hat{\mu}_{yab}(\eta) = \eta \hat{\mu}_{yab\text{H}} + (1 - \eta) \hat{\mu}_{yba\text{H}} \,,$$

where $\eta \in (0, 1)$ is a constant. The optimal choice of η, which minimizes the variance of $\hat{\mu}_{yab}(\eta)$, is given by

$$\eta_o = \frac{V_B\big(\hat{\mu}_{yba\text{H}}\big)}{V_A\big(\hat{\mu}_{yab\text{H}}\big) + V_B\big(\hat{\mu}_{yba\text{H}}\big)} \,, \tag{16.4}$$

where $V_A(\cdot)$ and $V_B(\cdot)$ denote respectively the design-based variance under the survey designs for frames A and B. Let $\hat{\eta}_o$ be a consistent estimator of η_o. Replacing η_o by $\hat{\eta}_o$ does not change the asymptotic optimality of $\hat{\mu}_{yab}(\eta_o)$ since

$$\hat{\mu}_{yab}(\hat{\eta}_o) - \hat{\mu}_{yab}(\eta_o) = \big(\hat{\eta}_o - \eta_o\big)\big(\hat{\mu}_{yab\mathrm{H}} - \hat{\mu}_{yba\mathrm{H}}\big) = o_p\big(n_c^{-1/2}\big)\,,$$

where $n_c = \min\{n_{ab}, n_{ba}\}$ and (n_{ab}, n_{ba}) are the sample sizes of $(\mathbf{S}_{ab}, \mathbf{S}_{ba})$. The optimal choice of η depends on the variable y. A sub-optimal choice of η is to use $\gamma_o = V_B(\hat{N}_{ba})/\{V_A(\hat{N}_{ab}) + V_B(\hat{N}_{ba})\}$, which provides optimal estimation of N_{ab} through $\hat{N}_{ab}(\gamma) = \gamma\hat{N}_{ab} + (1-\gamma)\hat{N}_{ba}$ and is independent of any y variables (Skinner and Rao 1996).

If the frame sizes N_A and N_B and the common domain size N_{ab} are all known, the final dual frame estimator of μ_y is given by

$$\hat{\mu}_{y\mathrm{DF}} = W_a\hat{\mu}_{ya\mathrm{H}} + W_b\hat{\mu}_{yb\mathrm{H}} + W_{ab}\hat{\mu}_{yab}(\hat{\eta}_o)\,. \tag{16.5}$$

One practical scenario is that the two frame population sizes N_A and N_B are known but the domain size N_{ab} is unknown. Using the optimal estimator $\hat{N}_{ab}(\hat{\gamma}_o)$, the other two domain sizes can be estimated by $\hat{N}_a = N_A - \hat{N}_{ab}(\hat{\gamma}_o)$ and $\hat{N}_b = N_B - \hat{N}_{ab}(\hat{\gamma}_o)$.

Another practical scenario is that the frame A is complete but the frame B is not. In this case the domain $b = \mathbf{A}^c \cap \mathbf{B}$ becomes empty. We have $N_b = 0$ and $N_{ab} = N_B$. Estimation based on the post-stratified samples can be modified by removing the domain b and the sample $\mathbf{S}_b$ from the formulation.

Variance estimation for dual frame or multiple frame estimators based on post-stratified samples under general probability sampling designs is a very challenging topic. See, for instance, Skinner and Rao (1996) and Lohr and Rao (2000, 2006). The rest of the chapter focuses on point estimation and confidence intervals using the pseudo empirical likelihood methods. The requirement on frame membership identification can be relaxed through a multiplicity-based approach presented in Sect. 16.3. A bootstrap procedure is introduced in Sect. 16.4 to overcome the difficulties with variance estimation.

16.2 Pseudo Empirical Likelihood for Dual Frame Surveys

We show that the pseudo empirical likelihood formulation can be used for post-stratified dual frame samples. Although $\mathbf{S}_{ab}$ and $\mathbf{S}_{ba}$ should be viewed as two independent samples from the same domain ab, it is strategically and computationally more convenient to use two duplicated domains $ab = \mathbf{A} \cap \mathbf{B}$ and $ba = \mathbf{B} \cap \mathbf{A}$ and view $\mathbf{S}_{ab}$ as a sample from ab and $\mathbf{S}_{ba}$ a sample from ba.

The pseudo empirical likelihood approach to dual frame surveys possesses two major advantages: the flexibility for incorporating different types of population auxiliary information through the constrained maximization and the construction of pseudo empirical likelihood ratio confidence intervals.

16.2.1 Point Estimation

We first consider scenarios where the frame sizes N_A and N_B and the domain size N_{ab} are all known. The overall population mean μ_y given in (16.1) can be re-written as

$$\mu_y = W_a\mu_{ya} + W_{ab}(\eta)\mu_{yab} + W_{ba}(\eta)\mu_{yba} + W_b\mu_{yb}\,,$$

where $W_{ab}(\eta) = \eta N_{ab}/N$ and $W_{ba}(\eta) = (1-\eta)N_{ab}/N$ with $\eta \in (0, 1)$ being a fixed constant to be specified. Note that $\mu_{yab} = \mu_{yba}$ is the mean for the common domain $ab = ba$ and $W_{ab}(\eta)\mu_{yab} + W_{ba}(\eta)\mu_{yba} = W_{ab}\mu_{yab}$ for any η. The dual frame samples $\mathbf{S}_A$ and $\mathbf{S}_B$ can be simply combined into a single "post-stratified sample" $(\mathbf{S}_a, \mathbf{S}_{ab}, \mathbf{S}_{ba}, \mathbf{S}_b)$, with post-stratum sample sizes $(n_a, n_{ab}, n_{ba}, n_b)$. The two frame sample sizes are $n_A = n_a + n_{ab}$ and $n_B = n_b + n_{ba}$.

Following the stratified formulation of the pseudo empirical likelihood methods described in Sect. 8.3, we define the pseudo empirical likelihood function for the post-stratified dual frame samples as

$$\begin{aligned}
&\ell_{\mathrm{DF}}(\mathbf{p}_a, \mathbf{p}_{ab}, \mathbf{p}_{ba}, \mathbf{p}_b)\\
&= n\Big\{W_a \sum_{i\in\mathbf{S}_a} \tilde{d}_i(\mathbf{S}_a)\log(p_{ai}) + W_{ab}(\eta)\sum_{i\in\mathbf{S}_{ab}} \tilde{d}_i(\mathbf{S}_{ab})\log(p_{abi})\\
&\quad + W_{ba}(\eta)\sum_{i\in\mathbf{S}_{ba}} \tilde{d}_i(\mathbf{S}_{ba})\log(p_{bai}) + W_b \sum_{i\in\mathbf{S}_b} \tilde{d}_i(\mathbf{S}_b)\log(p_{bi})\Big\}\,,
\end{aligned} \tag{16.6}$$

where $n = n_A + n_B$, and the normalized stratum survey weights $\tilde{d}_i(\mathbf{S}_a)$, $\tilde{d}_i(\mathbf{S}_{ab})$, $\tilde{d}_i(\mathbf{S}_{ba})$ and $\tilde{d}_i(\mathbf{S}_b)$ are defined in (16.2) and (16.3). The four sets of discrete probability measures $\mathbf{p}_a = (p_{a1}, \cdots, p_{an_a})'$, $\mathbf{p}_{ab} = (p_{ab1}, \cdots, p_{abn_{ab}})'$, $\mathbf{p}_{ba} = (p_{ba1}, \cdots, p_{ban_{ba}})'$ and $\mathbf{p}_b = (p_{b1}, \cdots, p_{bn_b})'$ correspond to the post-stratified samples $\mathbf{S}_a$, $\mathbf{S}_{ab}$, $\mathbf{S}_{ba}$ and $\mathbf{S}_b$, respectively.

The set of normalization constraints for the post-stratified dual frame samples is specified as

$$\sum_{i\in\mathbf{S}_a} p_{ai} = 1\,, \quad \sum_{i\in\mathbf{S}_{ab}} p_{abi} = 1\,, \quad \sum_{i\in\mathbf{S}_{ba}} p_{bai} = 1 \quad \text{and} \quad \sum_{i\in\mathbf{S}_b} p_{bi} = 1\,. \tag{16.7}$$

The constraint induced by the common domain mean $\mu_{yab} = \mu_{yba}$ needs to be enforced and is given by

$$\sum_{i\in\mathbf{S}_{ab}} p_{abi}\,y_i = \sum_{i\in\mathbf{S}_{ba}} p_{bai}\,y_i\,. \tag{16.8}$$

Let $\hat{p}_{ai}$, $\hat{p}_{abi}$, $\hat{p}_{bai}$ and $\hat{p}_{bi}$ be the maximizer of $\ell_{\mathrm{DF}}(\mathbf{p}_a, \mathbf{p}_{ab}, \mathbf{p}_{ba}, \mathbf{p}_b)$ under the normalization constraints (16.7) and the common domain mean constraint (16.8).

The maximum pseudo empirical likelihood estimator of μ_y is computed as

$$\hat{\mu}_{y\mathrm{PEL}} = W_a\hat{\mu}_{ya} + W_{ab}(\eta)\hat{\mu}_{yab} + W_{ba}(\eta)\hat{\mu}_{yba} + W_b\hat{\mu}_{yb}\,, \tag{16.9}$$

where $\hat{\mu}_{ya} = \sum_{i\in\mathbf{S}_a}\hat{p}_{ai}y_i$, $\hat{\mu}_{yab} = \sum_{i\in\mathbf{S}_{ab}}\hat{p}_{abi}y_i$, $\hat{\mu}_{yba} = \sum_{i\in\mathbf{S}_{ba}}\hat{p}_{bai}y_i = \hat{\mu}_{yab}$ due to constraint (16.8), and $\hat{\mu}_{yb} = \sum_{i\in\mathbf{S}_b}\hat{p}_{bi}y_i$. It can be seen that under the current setting without any additional constraints on auxiliary information, we have $\hat{\mu}_{ya} = \hat{\mu}_{ya\mathrm{H}}$, $\hat{\mu}_{yb} = \hat{\mu}_{yb\mathrm{H}}$ and

$$\hat{\mu}_{y\mathrm{PEL}} = W_a\hat{\mu}_{ya} + W_{ab}\hat{\mu}_{yab} + W_b\hat{\mu}_{yb}\,.$$

The choice of η only affects the estimator $\hat{\mu}_{yab} = \hat{\mu}_{yba}$ for the common domain mean μ_{yab}. It has been shown by Rao and Wu (2010b) that under suitable regularity conditions,

$$\hat{\mu}_{yab} = \eta\hat{\mu}_{yab\mathrm{H}} + (1-\eta)\hat{\mu}_{yba\mathrm{H}} + o_p\big(n_c^{-1/2}\big)$$

where $n_c = \min\{n_{ab}, n_{ba}\}$. The optimal choice of η is η_o given in (16.4) and the resulting maximum pseudo empirical likelihood estimator $\hat{\mu}_{y\mathrm{PEL}}$ is asymptotically equivalent to the dual frame estimator $\hat{\mu}_{y\mathrm{DF}}$ given in (16.5). The estimator $\hat{\mu}_{y\mathrm{PEL}}$ remains unchanged asymptotically if we replace η_o by a consistent estimator $\hat{\eta}_o$.

16.2.2 Confidence Intervals

We follow the setting used for point estimation without any additional auxiliary information and treat $\hat{\eta}_o$ as the fixed η_o. Let $\hat{p}_{ai}$, $\hat{p}_{abi}$, $\hat{p}_{bai}$ and $\hat{p}_{bi}$ be the maximizer of $\ell_{\mathrm{DF}}(\mathbf{p}_a, \mathbf{p}_{ab}, \mathbf{p}_{ba}, \mathbf{p}_b)$ under the constraints (16.7) and (16.8). The pseudo empirical likelihood ratio statistic for the population mean μ_y is computed as

$$r_{\mathrm{DF}}(\theta) = -2\big\{\ell_{\mathrm{DF}}\big(\tilde{\mathbf{p}}_a(\theta), \tilde{\mathbf{p}}_{ab}(\theta), \tilde{\mathbf{p}}_{ba}(\theta), \tilde{\mathbf{p}}_b(\theta)\big) - \ell_{\mathrm{DF}}\big(\hat{\mathbf{p}}_a, \hat{\mathbf{p}}_{ab}, \hat{\mathbf{p}}_{ba}, \hat{\mathbf{p}}_b\big)\big\}\,, \tag{16.10}$$

where $\tilde{\mathbf{p}}_a(\theta)$, $\tilde{\mathbf{p}}_{ab}(\theta)$, $\tilde{\mathbf{p}}_{ba}(\theta)$ and $\tilde{\mathbf{p}}_b(\theta)$ maximize $\ell_{\mathrm{DF}}(\mathbf{p}_a, \mathbf{p}_{ab}, \mathbf{p}_{ba}, \mathbf{p}_b)$ subject to (16.7), (16.8) and the parameter constraint

$$W_a\sum_{i\in\mathbf{S}_a}p_{ai}y_i + W_{ab}(\eta_o)\sum_{i\in\mathbf{S}_{ab}}p_{abi}y_i + W_{ba}(\eta_o)\sum_{i\in\mathbf{S}_{ba}}p_{bai}y_i + W_b\sum_{i\in\mathbf{S}_b}p_{bi}y_i = \theta \tag{16.11}$$

for the given θ.

The asymptotic distribution of $r_{\mathrm{DF}}(\theta)$ involves a design effect, $\mathrm{deff}_{\mathrm{DF}}$, which requires variance estimation. It follows the same arguments described in Sect. 8.4.2 for the design effect $\mathrm{deff}_{\mathrm{ST}}$ under stratified sampling, except for one major difference: the post-stratified samples are not independent of each other. The variance part of the design effect needs to be handled by using the sampling designs for the two samples $\mathbf{S}_A$ and $\mathbf{S}_B$ which are independent. For instance, the variance formula for the dual frame estimator $\hat{\mu}_{y\mathrm{DF}}$ given in (16.5) can be derived by using

$$Var\big(\hat{\mu}_{y\mathrm{DF}}\big) = V_A\big\{W_a\hat{\mu}_{ya\mathrm{H}} + W_{ab}(\eta_o)\hat{\mu}_{yab\mathrm{H}}\big\} + V_B\big\{W_b\hat{\mu}_{yb\mathrm{H}} + W_{ba}(\eta_o)\hat{\mu}_{yba\mathrm{H}}\big\}\,.$$

Further details on the design effect $\mathrm{deff}_{\mathrm{DF}}$ can be found in Rao and Wu (2010b). Under suitable regularity conditions, the adjusted pseudo empirical likelihood ratio statistics $r_{\mathrm{DF}}^{[a]}(\theta) = r_{\mathrm{DF}}(\theta)/\mathrm{deff}_{\mathrm{DF}}$ converges in distribution to a χ^2 random variable with one degree of freedom when $\theta = \mu_y$. The $(1-\alpha)$-level pseudo empirical likelihood ratio confidence interval for μ_y can be constructed as $\{\theta \mid r_{\mathrm{DF}}^{[a]}(\theta) \le \chi_1^2(\alpha)\}$, where $\chi_1^2(\alpha)$ is the upper α-quantile of the χ^2 distribution with one degree of freedom.

16.2.3 Auxiliary Information and Computational Procedures

There are two types of auxiliary information for dual frame surveys: frame-specific information and overall population information. The pseudo empirical likelihood approach is very flexible in using either type of auxiliary information through suitable calibration constraints. This is also closely related to computational procedures for the constrained maximization problem. The normalization constraints (16.7) can be reformulated into (8.19) and (8.20) given in Sect. 8.4.2. Any other calibration constraints or parameter constraints need to follow the same structure as (8.20) so that the computational algorithms of Sect. 8.4.2 can be used.

Suppose that $\mathbf{x}_{Ai}$ is only measured for the frame A sample and the frame population mean $\mu_{\mathbf{x}A}$ is available. The frame-specific calibration constraints can be written as

$$\frac{N_a}{N_A}\sum_{i\in\mathbf{S}_a} p_{ai}\mathbf{x}_{Ai} + \frac{N_{ab}}{N_A}\sum_{i\in\mathbf{S}_{ab}} p_{abi}\mathbf{x}_{Ai} = \mu_{\mathbf{x}A}\,, \tag{16.12}$$

which can further be reformulated as

$$W_a\sum_{i\in\mathbf{S}_a} p_{ai}\mathbf{x}_{Ai} + W_{ab}(\eta_o)\sum_{i\in\mathbf{S}_{ab}} p_{abi}\frac{\mathbf{x}_{Ai}}{\eta_o} + W_{ba}(\eta_o)\sum_{i\in\mathbf{S}_{ba}} p_{bai}\cdot\mathbf{0} + W_b\sum_{i\in\mathbf{S}_b} p_{bi}\cdot\mathbf{0} = \frac{N_A}{N}\mu_{\mathbf{x}A}\,. \tag{16.13}$$

If the overall population mean $\mu_{\mathbf{x}}$ is available and $\mathbf{x}_i$ is observed for both samples $\mathbf{S}_A$ and $\mathbf{S}_B$, this information can be utilized through two sets of calibration constraints given by

$$W_a \sum_{i\in\mathbf{S}_a} p_{ai}\mathbf{x}_i + W_{ab}(\eta_o)\sum_{i\in\mathbf{S}_{ab}} p_{abi}\mathbf{x}_i + W_{ba}(\eta_o)\sum_{i\in\mathbf{S}_{ba}} p_{bai}\mathbf{x}_i + W_b\sum_{i\in\mathbf{S}_b} p_{bi}\mathbf{x}_i = \mu_{\mathbf{x}} \tag{16.14}$$

and

$$\sum_{i\in\mathbf{S}_{ab}} p_{abi}\mathbf{x}_i = \sum_{i\in\mathbf{S}_{ba}} p_{bai}\mathbf{x}_i\,. \tag{16.15}$$

The constraints (16.15) are enforced due to the common domain mean for $ab = ba$. The computational version of (16.15) is given by

$$W_a\sum_{i\in\mathbf{S}_a} p_{ai}\cdot\mathbf{0} + W_{ab}(\eta_o)\sum_{i\in\mathbf{S}_{ab}} p_{abi}\frac{\mathbf{x}_i}{\eta_o} + W_{ba}(\eta_o)\sum_{i\in\mathbf{S}_{ba}} p_{bai}\frac{-\mathbf{x}_i}{1-\eta_o} + W_b\sum_{i\in\mathbf{S}_b} p_{bi}\cdot\mathbf{0} = \mathbf{0}\,. \tag{16.16}$$

Point estimation in the presence of auxiliary information presents no additional difficulties since the computational procedures are straightforward. However, pseudo empirical likelihood ratio confidence intervals require a modified design effect based on fitted residuals as described in Sect. 8.4.2. The multiplicity-based approach presented in the next section and the bootstrap procedure described in Sect. 16.4 are attractive alternatives.

16.3 A Multiplicity-Based Approach for Multiple Frame Surveys

Inferences with multiple frame surveys based on post-stratified samples become cumbersome if the number of frames used is three or more. For a three-frame survey involving frames A, B and C, the overall population is partitioned into seven domains a, b, c, ab, ac, bc and abc, and samples from each frame need to have a corresponding partition. For instance, the frame A sample is partitioned as $\mathbf{S}_A = \mathbf{S}_a \cup \mathbf{S}_{ab} \cup \mathbf{S}_{ac} \cup \mathbf{S}_{abc}$. In addition to the notational complexities, the most crucial theoretical assumption that the frame membership for each unit can be accurately identified becomes a practical nightmare. In this section we present a single-frame multiplicity-based approach to inferences from multiple frame surveys. The method requires less information on frame membership details and is generally applicable to L-frame ($L \geq 2$) survey samples.

16.3.1 The Single-Frame Multiplicity-Based Estimator

Let $A_1, \cdots, A_L$ denote the L frames. Let $\mathbf{A}_l$ be the set of all units covered by frame A_l, $l = 1, \cdots, L$. The term *multiplicity* is defined as *the number of* *frames* a particular unit belongs to. It is a partial membership information that does not require detailed membership identifications for each frame. Information on multiplicity can often be obtained with some minor effort during the data collection process (Mecatti 2007).

Let $\mathbf{S}_1, \cdots, \mathbf{S}_L$ be the L independent survey samples drawn respectively from the L frames with sample sizes $n_1, \cdots, n_L$. Let $\{(y_{li}, \mathbf{x}_{li}), i \in \mathbf{S}_l\}, l = 1, \cdots, L$ be the L-frame samples, where y_{li} is the common study variable attached to unit i on frame A_l and $\mathbf{x}_{li}$ is a vector of auxiliary variables which are not necessarily common across different frames. Let $d_{li} = 1/\pi_{li}$ be the design weights associated with frame A_l, where $\pi_{li} = P(i \in \mathbf{S}_l)$ are the first order inclusion probabilities for the frame A_l sampling design. Note that any frame-specific unit (li) corresponds to a unique j in the overall population $\mathbf{U} = \{1, 2, \cdots, N\}$ but a unit $j \in \mathbf{U}$ may belong to more than one frame. Let m_{li} be the multiplicity of unit i in frame A_l, which is the number of frames unit (li) belongs to. For dual frame surveys, we have $m_{li} = 1$ if $i \in a$ or $i \in b$ and $m_{li} = 2$ if $i \in ab$.

The so-called single-frame multiplicity-based approach is to pool all the L samples from the L-frame surveys without specific frame membership details and construct an estimator using the information on multiplicity. It is straightforward to show that

$$\sum_{l=1}^{L} \sum_{i \in \mathbf{A}_l} \frac{y_{li}}{m_{li}} = \sum_{i=1}^{N} y_i = N\mu_y \,.$$

If the overall population size N is known, then a design-unbiased multiplicity-based estimator of the population mean μ_y is given by

$$\hat{\mu}_{yM} = \frac{1}{N} \sum_{l=1}^{L} \sum_{i \in \mathbf{S}_l} d_{li} \frac{y_{li}}{m_{li}} \,. \tag{16.17}$$

The estimator $\hat{\mu}_{yM}$ given in (16.17) keeps all the duplicated units from different frames but replaces y_{li} by y_{li}/m_{li}. This amounts to letting the value of the study variable, y_{li}, be shared by the m_{li} frames to which the unit (li) belongs. This is the idea of *variable sharing* which was first used by Rao (1968) to handle a single frame with an unknown amount of duplication. If the population size N is unknown, an unbiased estimator of N is given by

$$\hat{N}_M = \sum_{l=1}^{L} \sum_{i \in \mathbf{S}_l} \frac{d_{li}}{m_{li}} \,.$$

The single-frame multiplicity-based *variable sharing* estimator $\hat{\mu}_{yM}$ given in (16.17) can also be viewed from a different angle. If we re-write the estimator as

$$\hat{\mu}_{yM} = \frac{1}{N} \sum_{l=1}^{L} \sum_{i \in \mathbf{S}_l} \frac{d_{li}}{m_{li}} y_{li} \,,$$

then it is the so-called *weight share* estimator (Lavallée 2007). The basic design weights d_{li} attached to unit i in frame A_l is shared by the same unit on all m_{li} different frames.

16.3.2 Multiplicity-Based Pseudo Empirical Likelihood Approach

Let $\mathbf{p}_l = (p_{l1}, \cdots, p_{ln_l})'$ be the discrete probability measure over the sample $\mathbf{S}_l$, $l = 1, \cdots, L$. The single-frame multiplicity-based pseudo empirical likelihood function for the L-frame survey samples is defined as

$$\ell_M(\mathbf{p}_1, \cdots, \mathbf{p}_L) = \frac{n_M}{\hat{N}_M} \sum_{l=1}^{L} \sum_{i \in \mathbf{S}_l} \frac{d_{li}}{m_{li}} \log(p_{li}) \,, \tag{16.18}$$

where $n_M = \sum_{l=1}^{L} n_l$. Ignoring the constant $n_M/\hat{N}_M$, the pseudo empirical likelihood function $\ell_M(\mathbf{p}_1, \cdots, \mathbf{p}_L)$ is a design-unbiased estimator of the census empirical log-likelihood $\sum_{i=1}^{N} \log(p_i)$.

The single-frame multiplicity-based approach treats $(\mathbf{p}_1, \cdots, \mathbf{p}_L)$ as a single discrete probability measure and imposes the normalization constraint

$$\sum_{l=1}^{L} \sum_{i \in \mathbf{S}_l} p_{li} = 1 \,. \tag{16.19}$$

Maximizing $\ell_M(\mathbf{p}_1, \cdots, \mathbf{p}_L)$ subject to (16.19) leads to $\hat{p}_{li} = (d_{li}/m_{li})/\hat{N}_M$. The maximum pseudo empirical likelihood estimator of μ_y is computed as

$$\hat{\mu}_{y\text{MEL}} = \sum_{l=1}^{L} \sum_{i \in \mathbf{S}_l} \hat{p}_{li} y_{li} = \frac{1}{\hat{N}_M} \sum_{l=1}^{L} \sum_{i \in \mathbf{S}_l} \frac{d_{li}}{m_{li}} y_{li} \,,$$

which is the Hájek estimator corresponding to $\hat{\mu}_{yM}$. For dual frame surveys, the estimator $\hat{\mu}_{y\text{MEL}}$ is asymptotically equivalent to the post-stratified maximum pseudo empirical likelihood estimator $\hat{\mu}_{y\text{PEL}}$ given in (16.9) with unknown N_A, N_B, N_{ab} and the choice of $\eta = 1/2$. The multiplicity-based estimator is therefore not necessarily optimal under dual frame surveys.

Let $\tilde{p}_{li}(\theta)$ be the maximizer of $\ell_M(\mathbf{p}_1, \cdots, \mathbf{p}_L)$ under the constraint (16.19) and the parameter constraint

$$\sum_{l=1}^{L} \sum_{i \in \mathbf{S}_l} p_{li} y_{li} = \theta \tag{16.20}$$

for a given θ. The multiplicity-based pseudo empirical likelihood ratio function for the population mean μ_y is defined as

$$r_M(\theta) = -2\big\{\ell_M\big(\tilde{\mathbf{p}}_1(\theta), \cdots, \tilde{\mathbf{p}}_L(\theta)\big) - \ell_M\big(\hat{\mathbf{p}}_1, \cdots, \hat{\mathbf{p}}_L\big)\big\}\,. \tag{16.21}$$

Let deff_M be the design effect associated with the estimator $\hat{\mu}_{y\text{MEL}}$. Under suitable regularity conditions, the adjusted pseudo empirical likelihood ratio statistic $r_M^{[a]}(\theta) = r_M(\theta)/\text{deff}_M$ converges in distribution to a χ^2 random variable with one degree of freedom when $\theta = \mu_y$. The design effect is given by $\text{deff}_M = Var\big(\hat{\mu}_{y\text{MEL}}\big)/\big(\sigma_y^2/n_M\big)$, where σ_y^2 can be estimated by

$$\hat{\sigma}_y^2 = \frac{1}{\hat{N}_M} \sum_{l=1}^{L} \sum_{i \in \mathbf{S}_l} \frac{d_{li}}{m_{li}} \big(y_{li} - \hat{\mu}_{y\text{MEL}}\big)^2\,,$$

and $Var\big(\hat{\mu}_{y\text{MEL}}\big)$ can be estimated by

$$\frac{1}{\hat{N}_M^2} \sum_{l=1}^{L} v_l\Big(\sum_{i \in \mathbf{S}_l} d_{li} \tilde{y}_{li}\Big)\,,$$

where $\tilde{y}_{li} = (y_{li} - \hat{\mu}_{y\text{MEL}})/m_{li}$ and $v_l(\cdot)$ is the variance estimator under the sampling design for frame A_l.

If the overall population mean $\mu_{\mathbf{x}}$ is available and $\mathbf{x}_{li}$ are observed for all L samples, the information can be conveniently used through the calibration constraints

$$\sum_{l=1}^{L} \sum_{i \in \mathbf{S}_l} p_{li} \mathbf{x}_{li} = \mu_{\mathbf{x}}\,. \tag{16.22}$$

The design effect in this case is related to a regression-type estimator under the multiplicity-based approach. The design effect is required for constructing pseudo empirical likelihood ratio confidence intervals but can be bypassed through a bootstrap procedure.

16.4 A Bootstrap Procedure for Multiple Frame Surveys

Suppose that the L-frame survey samples $\mathbf{S}_1, \cdots, \mathbf{S}_L$ are all selected by single-stage unequal probability sampling designs with small sampling fractions. We treat d_{li} and m_{li} as part of the sample data so the complete sample observations from frame A_l have the form $\{(d_{li}, m_{li}, y_{li}, \mathbf{x}_{li}), i \in \mathbf{S}_l\}$. The following bootstrap procedure can be used to approximate the sampling distribution of the multiplicity-based pseudo empirical likelihood ratio statistic $r_M(\theta)$. We assume that the calibration constraints (16.22) are used.

1. Select a bootstrap sample $\mathbf{S}_l^*$ of size n_l from the original sample $\mathbf{S}_l$ using simple random sampling with replacement, and do this independently for $l = 1, \cdots, L$. Let $\{(d_{li}^*, m_{li}^*, y_{li}^*, \mathbf{x}_{li}^*), i \in \mathbf{S}_l^*\}$ be the bootstrap sample data, $l = 1, \cdots, L$.
2. Construct the bootstrap version of the pseudo empirical likelihood function as

$$\ell_M^*(\mathbf{p}_1, \cdots, \mathbf{p}_L) = \frac{n_M}{\hat{N}_M^*} \sum_{l=1}^{L} \sum_{i \in \mathbf{S}_l^*} \frac{d_{li}^*}{m_{li}^*} \log(p_{li}) \,,$$

   where $\hat{N}_M^* = \sum_{l=1}^{L} \sum_{i \in \mathbf{S}_l^*} d_{li}^*/m_{li}^*$; the bootstrap versions of the parameter constraint (16.20) and the calibration constraints (16.22) are

$$\sum_{l=1}^{L} \sum_{i \in \mathbf{S}_l^*} p_{li} y_{li}^* = \hat{\mu}_{y\text{MEL}} \qquad \text{and} \qquad \sum_{l=1}^{L} \sum_{i \in \mathbf{S}_l^*} p_{li} \mathbf{x}_{li}^* = \mu_{\mathbf{x}} \,,$$

   where $\hat{\mu}_{y\text{MEL}}$ is the multiplicity-based maximum pseudo empirical likelihood estimator of μ_y with the calibration constraints (16.22). The normalization constraint (16.19) remains unchanged.
3. Compute the bootstrap version of the unadjusted pseudo empirical likelihood ratio statistics $r_M^*(\theta)$ at $\theta = \hat{\mu}_{y\text{MEL}}$ using $\ell_M^*(\mathbf{p}_1, \cdots, \mathbf{p}_L)$ and the constraints described in Step 2 to obtain $r_M^*[1] = r_M^*(\hat{\mu}_{y\text{MEL}})$.
4. Repeat Steps 1–3 a large number B times, independently, to obtain a sequence $r_M^*[1], r_M^*[2], \cdots, r_M^*[B]$.

Let b_α^* be the $100(1-\alpha)$th sample quantile of the bootstrap sequence $(r_M^*[1], r_M^*[2], \cdots, r_M^*[B])$. The bootstrap calibrated pseudo empirical likelihood ratio confidence interval on μ_y is constructed as $\{\theta \mid r_M(\theta) \le b_\alpha^*\}$. If stratified unequal probability sampling or multi-stage sampling designs are used for one or more frames, the bootstrap procedure requires modifications. It may not work for certain complex survey designs.

16.5 Additional Remarks

Dual frame and multiple frame survey designs are often used in practice. Statistical analysis of multiple frame survey samples, however, faces several challenges. Obtaining accurate information on frame memberships is the first challenge in practice. Estimating the unknown domain population sizes under complex sampling designs is another problem. In addition, incorporating various types of auxiliary population information into inferential procedures is also difficult under the conventional approach with post-stratified samples. Variance estimation and confidence intervals are even harder to handle with multiple frame surveys under general unequal probability sampling designs.

The multiplicity-based pseudo empirical likelihood approach presented in Sect. 16.3.2 and the bootstrap procedure described in Sect. 16.4 have potential to be very useful in practice. They are very easy to implement for surveys involving two, three or more frames, do not require full frame membership information, and are flexible in using auxiliary population information that is available. Pseudo empirical likelihood ratio confidence intervals have a clear advantage over the customarily normal approximation-based intervals for population proportions, which are scenarios where dual frame or multiple frame surveys are often used. For example, a survey on a particular health condition among the homeless population in a large city may use a three-frame design. The first is a general area frame, the second is the list of homeless shelters, and the third is the list of soup kitchens and food banks in the city. The required information on multiplicity is also insensitive to domain misclassifications, as shown by Mecatti (2007).

One potential issue with the multiplicity-based approach is that if the samples are of very different sampling fractions for different frames. This might happen if one frame is more difficult to sample from than another. The survey weights could be more variable than we would want, resulting in less efficient estimators. Further investigations on estimation under such scenarios are required.

When all frames are complete, multiple frame sampling becomes the so-called *multiple surveys* scenario where several independent samples are taken from the same target population. Statistical analysis of multiple surveys involves strategies for combining samples that are very different from multiple frame surveys. Wu (2004b) contains discussions on combining two independent survey samples through the empirical likelihood method. Kim and Rao (2012) presents a model-assisted approach for combining data from two independent surveys. Chapter 17 discusses a unique topic on combining a reference probability survey sample with a non-probability survey sample to draw valid statistical inferences.

Chapter 17
Non-probability Survey Samples

We provide an overview of the emerging topic of non-probability survey samples which has drawn increased attention in the fields of survey methodology and official statistics. We highlight some of the issues in analyzing non-probability survey samples and present some of the methodological advances that have appeared in recent years. One important message from this chapter is that probability survey samples and design-based inference, which are the main focus of most other chapters of the book, play a fundamental role in the theoretical framework for non-probability survey samples.

17.1 Non-probability Samples

The use of non-probability survey samples has a very long history. Quota sampling, for instance, serves as a cost-effective alternative method to select a survey sample when one is limited by resources or restricted by the availability of sampling frames. While the design of quota samples attempts to achieve demographical representativeness of the sample data, the actual data collection process is left in the hands of field workers or contractors. Despite their use in various contexts, non-probability sampling methods have not gained much momentum in survey practice, especially in official statistics, partly due to the lack of theoretical foundation for statistical inference and partly due to the dominance and success of probability sampling methods and the widespread acceptance of the design-based framework.

The success of probability sampling has led to more frequent surveys and more ambitious research projects that involve long and sophisticated questionnaires and measurements. Response burden and privacy concerns, along with many other factors, have led to dramatic decreases in response rates for almost all surveys. The challenge of low participation rates and the ever-increasing costs for conducting

C. Wu, M. E. Thompson, *Sampling Theory and Practice*, ICSA Book
Series in Statistics, https://doi.org/10.1007/978-3-030-44246-0_17

surveys using probability sampling methods, coupled with technology advances, have resulted in a paradigm shift in recent years for government agencies, research institutions and industrial organizations, in the quest for other cheaper and quicker alternatives for data collection (Citro 2014). These alternative sources include administrative data, commercial transaction records, social media and streaming activities, and non-probability survey samples.

Non-probability survey samples have become popular in recent years due to their cost-and-time efficiency and the rise of the web based surveys. The most popular type of web survey employs so-called *opt-in panels*. These panels consist of volunteers who have agreed to participate and who typically receive a form of compensation for completing surveys. Opt-in panelists are recruited through various convenient but non-probability methods. Online research through opt-in panel surveys has become prevalent due to an efficient recruitment process, quick responses, and low maintenance expenses. Tourangeau et al. (2013) contains many examples for web based surveys.

Government organizations and statistical agencies are embracing the new opportunities and moving beyond a (probability) survey-first approach, adopting new methods and integrating data from a variety of existing sources. However, there are serious issues and major challenges for the use of non-probability survey samples and data from unconventional sources. The "*Summary Report of the AAPOR Task Force on Non-probability Sampling*" by Baker et al. (2013), which was commissioned by the American Association of Public Opinion Research (AAPOR) Executive Council, contains a well documented list of such issues and challenges. The task force's conclusions and recommendations include the following: (i) in contrast with the case of probability sampling, there is no single framework that adequately encompasses all of non-probability sampling; (ii) making inferences for any probability or non-probability survey requires some reliance on modeling assumptions; and (iii) if non-probability samples are to gain wider acceptance among survey researchers there must be a more coherent framework and accompanying set of measures for evaluating their quality.

The discussions in the rest of the chapter focus on establishing a general inferential framework for analysis of non-probability survey samples under an ideal setting where there exists a probability survey sample with auxiliary information about the target population. Issues and challenges not addressed by this framework are briefly discussed in the last section.

17.2 General Setting and Basic Assumptions

We follow the same notation used in Parts I and II of the book. Let $\mathscr{F}_N = \{(\mathbf{x}_i, y_i), \; i = 1, 2, \cdots, N\}$ represent the data structure for the finite population $\mathbf{U} = \{1, 2, \cdots, N\}$; let $\mu_y = N^{-1}\sum_{i=1}^{N} y_i$ be the parameter of interest. The general setting used in this chapter is as follows. Let $\mathbf{S}_A$ be the set of n_A units selected from the target population using a non-probability data collection method.

The study variable y and auxiliary variables $\mathbf{x}$ are measured for each unit in the sample. Let $\{(\mathbf{x}_i, y_i), i \in \mathbf{S}_A\}$ be the dataset for the non-probability survey sample. In addition, there exists a reference probability survey sample $\mathbf{S}_B$ of size n_B with relevant auxiliary information on $\mathbf{x}$. The probability survey dataset is denoted as $\{(\mathbf{x}_i, w_i), i \in \mathbf{S}_B\}$, where w_i are the survey weights. The probability sample may also contain information on other variables but they are ignored if the variables are irrelevant to the measurements in the non-probability sample. The two samples $\mathbf{S}_A$ and $\mathbf{S}_B$ are assumed to be independent.

The most crucial aspect of non-probability survey samples is the selection mechanism. It is unknown and requires a suitable model for the inclusion indicator variable given the characteristics of the unit. Let $R_i = 1$ if $i \in \mathbf{S}_A$ and $R_i = 0$ otherwise, $i = 1, 2, \cdots, N$. The propensity scores are defined as

$$\pi_i^A = E_q\big(R_i \mid \mathbf{x}_i, y_i\big) = P_q\big(R_i = 1 \mid \mathbf{x}_i, y_i\big), \quad i = 1, 2, \cdots, N,$$

where the subscript q refers to the model for the selection mechanism of the non-probability sample $\mathbf{S}_A$, i.e., the propensity score model. The subscript may be dropped if there is no confusion. Note that the existence of a selection mechanism which can be modeled by a conditional probability distribution is itself a strong assumption.

It is important to see the similarities and differences in the notation and terms used here and those used for probability samples and missing data. First, the π_i^A are similar to the inclusion probabilities for probability samples, with one crucial difference: the inclusion probabilities π_i for probability samples are known from the sampling design, while the propensity scores π_i^A for non-probability samples are unknown and require estimation using a model. The latter are also called the *quasi-randomization* inclusion probabilities (Kott 1994). Second, the term "propensity score" has been used in the context of missing data or causal inference. The scenario with non-probability samples is very different from regular missing data problems. Recall that we have $\delta_i = 1$ for $i \in \mathbf{S}_R$ and $\delta_i = 0$ for $i \in \mathbf{S}_M$ for samples with missing data. What we have here is that $R_i = 1$ for all the units in the sample $\mathbf{S}_A$ and $R_i = 0$ for all the units outside the sample; for these units no information is available from the observed sample data.

The selection mechanism is called *ignorable* if $\pi_i^A = P(R_i = 1 \mid \mathbf{x}_i, y_i) = P(R_i = 1 \mid \mathbf{x}_i)$ for all i. This corresponds to *missing at random* (MAR) as defined by Rubin (1976) and Little and Rubin (2002). More formally, we consider the following three basic assumptions or conditions for the selection mechanism.

A1 The conditional distribution of the selection indicator R_i given the set of covariates $\mathbf{x}_i$ and the response variable y_i is independent of y_i.

A2 All units have a non-zero propensity score, i.e., $\pi_i^A > 0$ for all i.

A3 The indicator variables R_i and R_j are independent given $\mathbf{x}_i$ and $\mathbf{x}_j$ for $i \neq j$.

Condition **A1** is the MAR assumption. Condition **A2** implies that the sampling frame is complete and every unit in the population have a non-zero probability to be included in the non-probability sample. Condition **A3** is usually satisfied in practice unless important determinants of sample inclusion are excluded from $\mathbf{x}$. The three conditions are the standard starting point for dealing with non-probability survey samples (Elliott and Valliant 2017; Chen et al. 2020).

There are three possible sources of randomization under the current setting for analyzing non-probability survey samples: the probability sampling design p for sample $\mathbf{S}_B$, the propensity score model q for sample $\mathbf{S}_A$, and the outcome regression model ξ for y given $\mathbf{x}$. The sampling design p is always involved for the approaches described in the next two sections. One or both of q and ξ will also be involved, depending on the method.

17.3 Sample Matching and Mass Imputation

The naive estimator $\hat{\mu}_y = n_A^{-1} \sum_{i \in \mathbf{S}_A} y_i$ is typically biased for μ_y unless the propensity scores π_i^A are a constant, i.e., the non-probability sample can be viewed as an independent sample. The concept of *sample matching* was first discussed by Rivers (2007). Note that the non-probability sample dataset $\{(\mathbf{x}_i, y_i), i \in \mathbf{S}_A\}$ does not have the "weights" (i.e., $1/\pi_i^A$) available while the reference probability sample dataset $\{(\mathbf{x}_i, w_i), i \in \mathbf{S}_B\}$ has no measurements on the study variable y. It is tempting to use the weights in $\mathbf{S}_B$ or the y's in $\mathbf{S}_A$ to create a complete dataset with the triplets $(\mathbf{x}_i, y_i, w_i)$.

Rivers (2007) proposed to use the nearest neighbor method to do the matching. There are two possible directions in terms of sample matching. The first is to use the y_i's in $\mathbf{S}_A$ to complete the dataset for $\mathbf{S}_B$. For each $i \in \mathbf{S}_B$, we let $y_i^* = y_j$, where the unit $j \in \mathbf{S}_A$ and satisfies $\|\mathbf{x}_j - \mathbf{x}_i\| = \min_{k \in \mathbf{S}_A} \|\mathbf{x}_k - \mathbf{x}_i\|$. If j is not unique, one unit is randomly selected from the pool of eligible candidates. This is the same idea as the nearest neighbor imputation. The sample matching estimator for μ_y is given by

$$\hat{\mu}_{yB} = \frac{1}{\hat{N}} \sum_{i \in \mathbf{S}_B} w_i y_i^* , \qquad (17.1)$$

where $\hat{N} = \sum_{i \in \mathbf{S}_B} w_i$. The estimator $\hat{\mu}_{yB}$ is essentially an imputation-based estimator using the probability sample $\mathbf{S}_B$ for which the study variable y is missing for all the units in the sample. In the recent literature it is termed the *mass imputation estimator*. The requirement that the auxiliary variables $\mathbf{x}$ are available in both samples $\mathbf{S}_A$ and $\mathbf{S}_B$ makes the sample matching possible.

The estimator $\hat{\mu}_{yB}$ given in (17.1) can be constructed using other suitable imputation methods. Kim et al. (2019) considered a semiparametric model ξ such that $E_\xi(y \mid \mathbf{x}) = m(\mathbf{x}; \beta)$. Under the assumption **A1**, we have $E_\xi(y \mid \mathbf{x}, R = 1) = E_\xi(y \mid \mathbf{x})$, and a consistent estimator $\hat{\beta}$ for the model parameters β can be obtained by using the non-probability sample data. For instance, under a linear regression model with $m(\mathbf{x}_i; \beta) = \mathbf{x}_i'\beta$, we can use the ordinary least squares estimator of β, which is given by

$$\hat{\beta} = \Big(\sum_{i\in\mathbf{S}_A} \mathbf{x}_i\mathbf{x}_i'\Big)^{-1}\Big(\sum_{i\in\mathbf{S}_A} \mathbf{x}_i y_i\Big) .$$

For each $i \in \mathbf{S}_B$, we impute y_i by $y_i^* = m(\mathbf{x}_i; \hat{\beta})$. The mass imputation estimator of μ_y is then constructed in the same way as $\hat{\mu}_{yB}$ given in (17.1). The estimator is consistent for μ_y under the joint randomization of p and ξ, since $E_p(\hat{N}) \doteq N$ and

$$E\Big(\sum_{i\in\mathbf{S}_B} w_i y_i^*\Big) = E_p\Big[\sum_{i\in\mathbf{S}_B} w_i E_\xi\big\{m(\mathbf{x}_i; \hat{\beta})\big\}\Big] \doteq \sum_{i=1}^{N} m(\mathbf{x}_i; \beta) = E\Big(\sum_{i=1}^{N} y_i\Big) .$$

The second direction for sample matching is to let $w_i^* = w_j$, where unit $i \in \mathbf{S}_A$ is assigned a matched weight w_j from unit $j \in \mathbf{S}_B$ which is selected based on the nearest neighbor method. The sample matching estimator for μ_y is computed as

$$\hat{\mu}_{yA} = \frac{1}{\hat{N}^*}\sum_{i\in\mathbf{S}_A} w_i^* y_i , \tag{17.2}$$

where $\hat{N}^* = \sum_{i\in\mathbf{S}_A} w_i^*$. The estimator $\hat{\mu}_{yA}$ given in (17.2), however, is conceptually difficult to interpret. The matched weight $w_i^* = w_j$ is from the sampling design for sample $\mathbf{S}_B$ and has no connection to the propensity score π_i^A. Theoretical properties of the sample matching estimator $\hat{\mu}_{yA}$ have not been investigated in the existing literature. A related approach is the inverse probability weighted estimator presented in Sect. 17.5 using the estimated propensity scores.

17.4 Estimation of Propensity Scores

The propensity scores π_i^A for the non-probability sample play the same role as inclusion probabilities do for the probability sample. Direct estimation of π_i^A, however, is impossible since $R_i = 1$ for all units in the sample and no information is available for units with $R_i = 0$. Under the current setting with auxiliary information from a reference probability sample $\mathbf{S}_B$, there have been efforts to use the combined sample $\mathbf{S}_A \cup \mathbf{S}_B$ to estimate the propensity scores. See, for instance, Lee (2006), Lee and Valliant (2009), and Valliant and Dever (2011). The targets of the estimation process are essentially $\tilde{\pi}_i^A = P(\tilde{R}_i = 1 \mid \mathbf{x}_i)$, where $\tilde{R}_i = 1$ if $i \in \mathbf{S}_A$ and $\tilde{R}_i = 0$

if $i \in \mathbf{S}_B$. The resulting estimated propensity scores do not provide consistent estimators for the finite population parameters.

Consider the hypothetical situation where $\mathbf{x}_i$ is observed for all units in the finite population $\mathbf{U}$ while y_i is only observed for the non-probability sample $\mathbf{S}_A$. Estimation of the propensity scores under this scenario becomes the standard missing data problem with observations $\{(R_i, R_i y_i, \mathbf{x}_i), i = 1, 2, \cdots, N\}$. Suppose that the propensity scores can be modeled parametrically as $\pi_i^A = P(R_i = 1 \mid \mathbf{x}_i) = \pi(\mathbf{x}_i, \phi_0)$, where ϕ_0 is the true value of the unknown model parameters. The maximum likelihood estimator of π_i^A is computed as $\hat{\pi}_i^A = \pi(\mathbf{x}_i, \hat{\phi})$, where $\hat{\phi}$ maximizes the log-likelihood function

$$\begin{aligned}\ell(\phi) &= \sum_{i=1}^{N} \left\{ R_i \log \pi_i^A + (1 - R_i) \log (1 - \pi_i^A) \right\} \\ &= \sum_{i \in \mathbf{S}_A} \log \left\{ \frac{\pi(\mathbf{x}_i, \phi)}{1 - \pi(\mathbf{x}_i, \phi)} \right\} + \sum_{i=1}^{N} \log \left\{ 1 - \pi(\mathbf{x}_i, \phi) \right\}. \qquad (17.3)\end{aligned}$$

The log-likelihood function specified in (17.3), however, cannot be used in practice since we do not observe $\mathbf{x}_i$ for all the units in the finite population. Instead of using $\ell(\phi)$, Chen et al. (2020) proposed to compute the estimator $\hat{\phi}$ by maximizing the following pseudo log-likelihood function

$$\ell^*(\phi) = \sum_{i \in \mathbf{S}_A} \log \left\{ \frac{\pi(\mathbf{x}_i, \phi)}{1 - \pi(\mathbf{x}_i, \phi)} \right\} + \sum_{i \in \mathbf{S}_B} w_i \log \left\{ 1 - \pi(\mathbf{x}_i, \phi) \right\}, \qquad (17.4)$$

where the population total $\sum_{i=1}^{N} \log \left\{ 1 - \pi(\mathbf{x}_i, \phi) \right\}$ is replaced by the survey weighted estimator $\sum_{i \in \mathbf{S}_B} w_i \log \left\{ 1 - \pi(\mathbf{x}_i, \phi) \right\}$ using the reference probability sample $\mathbf{S}_B$.

Under the commonly used logistic regression model for the propensity scores where $\pi_i^A = \pi(\mathbf{x}_i, \phi_0) = \exp(\mathbf{x}_i'\phi_0)/\{1 + \exp(\mathbf{x}_i'\phi_0)\}$, the pseudo log-likelihood function (17.4) becomes

$$\ell^*(\phi) = \sum_{i \in \mathbf{S}_A} \mathbf{x}_i'\phi - \sum_{i \in \mathbf{S}_B} w_i \log \left\{ 1 + \exp(\mathbf{x}_i'\phi) \right\}.$$

The maximum pseudo likelihood estimator $\hat{\phi}$ can be obtained by solving the score equations (Sect. 7.3.2; Problem 9.2)

$$\frac{\partial}{\partial \phi} \ell^*(\phi) = \sum_{i \in \mathbf{S}_A} \mathbf{x}_i - \sum_{i \in \mathbf{S}_B} w_i \, \pi(\mathbf{x}_i, \phi) \, \mathbf{x}_i = \mathbf{0}. \qquad (17.5)$$

The estimator $\hat{\phi}$ is consistent under the joint randomization of p and q. Chen et al. (2020) contains discussions on alternative approaches to estimation of ϕ based on more general estimating equations or calibration equations.

17.5 Estimation of the Population Mean

The estimated propensity scores $\hat{\pi}_i^A = \pi(\mathbf{x}_i, \hat{\phi})$ are the main tool for valid statistical inferences with non-probability samples. We focus on the inverse probability weighted estimator and the doubly robust estimator of the population mean μ_y but general parameters defined through census estimating equations can also be handled.

The term "Inverse Probability Weighted" (IPW) estimator is used in this section, since there are no other survey weighting schemes for sample $\mathbf{S}_A$. The method is the same as the propensity score adjusted (PSA) estimator introduced in Sect. 9.3.2. The IPW estimator of μ_y is defined as

$$\hat{\mu}_{y\text{IPW}} = \frac{1}{\hat{N}^A} \sum_{i \in \mathbf{S}_A} \frac{y_i}{\hat{\pi}_i^A}, \tag{17.6}$$

where $\hat{N}^A = \sum_{i \in \mathbf{S}_A} (\hat{\pi}_i^A)^{-1}$. Note that the Hájek-type estimator given by (17.6) is preferred even if the population size N is known. The estimator $\hat{\mu}_{y\text{IPW}}$ is consistent under the assumed propensity score model and the probability sampling design for $\mathbf{S}_B$, as can be shown through the unbiasedness of the joint estimating functions for defining $\theta = \mu_y$ and ϕ (Sect. 9.6.1).

Let $\hat{\beta}$ be a consistent estimator for the model parameters in the semiparametric outcome regression model ξ such that $E_\xi(y \mid \mathbf{x}) = m(\mathbf{x}, \beta)$. If the $\mathbf{x}_i$ are available for all the units in the population, the general form of the doubly robust (DR) estimator of μ_y is given by

$$\hat{\mu}_{y\text{DR0}} = \frac{1}{N}\left[\sum_{i=1}^{N} \frac{R_i\{y_i - m(\mathbf{x}_i, \hat{\beta})\}}{\pi(\mathbf{x}_i, \hat{\phi})} + \sum_{i=1}^{N} m(\mathbf{x}_i, \hat{\beta})\right]. \tag{17.7}$$

The form of the estimator $\hat{\mu}_{y\text{DR0}}$ given by (17.7) is identical to the model-assisted "generalized difference estimator" discussed in Wu and Sitter (2001a) under the scenarios where the $\pi(\mathbf{x}_i, \hat{\phi})$ are the inclusion probabilities for the probability sample and the complete auxiliary information $\{\mathbf{x}_1, \cdots, \mathbf{x}_N\}$ is available. The doubly robust estimator under the current setting proposed by Chen et al. (2020) is given by

$$\hat{\mu}_{y\text{DR}} = \frac{1}{\hat{N}^A} \sum_{i \in \mathbf{S}_A} d_i^A\{y_i - m(\mathbf{x}_i, \hat{\beta})\} + \frac{1}{\hat{N}^B} \sum_{i \in \mathbf{S}_B} w_i m(\mathbf{x}_i, \hat{\beta}), \tag{17.8}$$

where $\hat{N}^A = \sum_{i \in \mathbf{S}_A} d_i^A$ with $d_i^A = 1/\hat{\pi}_i^A$, and $\hat{N}^B = \sum_{i \in \mathbf{S}_B} w_i$. The estimator $\hat{\mu}_{y\mathrm{DR}}$ is doubly robust in the sense that it is consistent under correct specification of the probability sampling design p for the reference probability sample together with one of the propensity score model q and the regression model ξ; it is not necessary that both be specified correctly. See Problem 9.5 for the properties of the standard doubly robust estimator with missing data. For non-probability survey samples, the sampling design p for the reference probability sample $\mathbf{S}_B$ is always part of the inferential framework.

17.6 Variance Estimation

Variance estimation with non-probability survey samples could follow the two basic strategies that have been used in many other scenarios: (i) Formulate the estimator for the main parameter of interest and the estimators of the nuisance parameters as the solution to a system of estimating equations; (ii) Decompose the total variance into variance components based on the sources of randomization involved. The theoretical variance or asymptotic variance formula is developed first, and a variance estimator can be constructed accordingly.

We first discuss variance estimation for the IPW estimator $\hat{\mu}_{y\mathrm{IPW}}$ defined in (17.6). The estimating equation for defining $\hat{\theta} = \hat{\mu}_{y\mathrm{IPW}}$ is given by

$$\sum_{i=1}^{N} \frac{R_i}{\pi(\mathbf{x}_i, \phi)}(y_i - \theta) = 0\,.$$

The score equations (17.5) are used for defining $\hat{\phi}$. Following techniques similar to those described in Sect. 9.6.1 on variance estimation for the PSA estimator, it can be shown (Chen et al. 2020) that

$$Var\big(\hat{\mu}_{y\mathrm{IPW}}\big) = \frac{1}{N^2}\sum_{i=1}^{N}(1-\pi_i^A)\pi_i^A\left(\frac{y_i-\mu_y}{\pi_i^A} - \mathbf{b}'\mathbf{x}_i\right)^2 + \mathbf{b}'\mathbf{D}\,\mathbf{b} + o\big(n_A^{-1}\big)\,, \tag{17.9}$$

where

$$\mathbf{b} = \left\{\sum_{i=1}^{N}\pi_i^A(1-\pi_i^A)\mathbf{x}_i\mathbf{x}_i'\right\}^{-1}\left\{\sum_{i=1}^{N}(1-\pi_i^A)(y_i-\mu_y)\mathbf{x}_i\right\},$$

and $\mathbf{D} = N^{-2}\,V_p\big(\sum_{i \in \mathbf{S}_B} w_i \pi_i^A \mathbf{x}_i\big)$ is the design-based variance from the reference probability sample $\mathbf{S}_B$. A consistent variance estimator can be constructed by plugging in estimates for each required component.

Variance estimation for the doubly robust estimator $\hat{\mu}_{y\mathrm{DR}}$ turns out to be a difficult problem. This was discussed briefly in Sect. 9.3.4 under the context of the doubly robust estimator with missing data. The variance of $\hat{\mu}_{y\mathrm{DR}}$, however, can be derived by assuming that the propensity score model is correctly specified. One important observation is that, under the propensity score model, the estimation of the regression model parameters β has no impact on the asymptotic variance of $\hat{\mu}_{y\mathrm{DR}}$. This is due to the difference structure of $\hat{\mu}_{y\mathrm{DR}}$ given by

$$\hat{\mu}_{y\mathrm{DR}} = \frac{1}{\hat{N}^A}\sum_{i\in\mathbf{S}_A} d_i^A y_i + \frac{1}{\hat{N}^B}\sum_{i\in\mathbf{S}_B} w_i m(\mathbf{x}_i,\hat{\beta}) - \frac{1}{\hat{N}^A}\sum_{i\in\mathbf{S}_A} d_i^A m(\mathbf{x}_i,\hat{\beta}),$$

and the fact that the last two terms in $\hat{\mu}_{y\mathrm{DR}}$ estimate the same population quantity $N^{-1}\sum_{i=1}^N m(\mathbf{x}_i,\beta)$. Let $\hat{\beta} = \beta^* + O_p\big(n_A^{-1/2}\big)$, where β^* is not necessarily the same as β_0 since the regression model is not assumed to be true. We replace $\hat{\beta}$ by β^* in the expression for $\hat{\mu}_{y\mathrm{DR}}$ and the resulting difference is of the order $o_p\big(n_A^{-1/2}\big)$. The main estimating equation for defining $\hat{\theta} = \hat{\mu}_{y\mathrm{DR}}$ can be written as

$$\sum_{i=1}^N \frac{R_i}{\pi(\mathbf{x}_i,\phi)}\big\{y_i - m(\mathbf{x}_i,\beta^*) + \bar{m}_B - \theta\big\} = 0, \tag{17.10}$$

where $\bar{m}_B = \big(\hat{N}^B\big)^{-1}\sum_{i\in\mathbf{S}_B} w_i m(\mathbf{x}_i,\beta^*)$, which depends only on the reference probability sample $\mathbf{S}_B$.

If the logistic regression model is used for the propensity scores, the estimating equation (17.10) for θ can be combined with the score equations (17.5) on ϕ to derive the asymptotic variance formula for $\hat{\mu}_{y\mathrm{DR}}$. Chen et al. (2020) contains further details on variance estimation for the doubly robust estimator and discussions on using the method of Kim and Haziza (2014) to construct a doubly robust variance estimator.

17.7 An Example and Additional Remarks

The general setting used in this chapter on non-probability survey samples requires an existing reference probability sample with auxiliary information. The auxiliary variables need to be available from both samples, and the regularity conditions need to be satisfied, especially the condition **A1** on the selection mechanism for the non-probability sample. We present an example in this section and provide some additional remarks on related issues.

We discuss results from analyzing a dataset collected by the Pew Research Center in 2015. The results were originally presented in Chen et al. (2020). The dataset, denoted as PRC, contains 56 variables which aim to reveal the relations between people and their communities. There are a total of 9301 individuals in the PRC dataset, who have been provided by eight different vendors with unknown sampling and data collection strategies. We treat the PRC dataset as a non-probability survey sample with the sample size $n_A = 9301$. There are seven study variables included in the PRC dataset that are of particular interest, of which six are binary variables: Talked with neighbors frequently (y_1), Tended to trust neighbors (y_2), Expressed opinions at a government level (y_3), Voted in local elections (y_4), Participated in school groups (y_5), Participated in service organizations (y_6). The seventh variable is treated as a continuous variable: Days had at least one drink in previous month (y_7). Valid inferences on those study variables are not immediately available from the PRC sample given its non-probability based selection methods.

There are two potential reference probability survey samples which were taken in the same period of time as the PRC sample. The first sample is the volunteer supplement survey sample from the Current Population Survey (CPS), a survey which is rigorously conducted by the UW Census Bureau. The CPS dataset contains $80,075$ individuals with measurements on volunteerism, which is highly relevant to the study variables considered in the PRC dataset. The second probability sample consists of data from the Behavioral Risk Factor Surveillance System (BRFSS). The BRFSS is designed to measure behavioral risk factors for US residents and has a large sample size, 441,456 in the relevant period. Neither of the two probability samples contain measurements of the study variables, but both share a rich set of common survey items with the PRC dataset as shown in Table 17.1.

We first examine marginal distributions of the variables from the three datasets. Table 17.1 contains the estimated population means using each of the three datasets. For the PRC dataset, because the sampling strategy is unknown and no survey weights are available, estimates presented are unadjusted simple sample means. For the BRFSS and the CPS datasets where survey weights are available as part of the data files, survey weighted estimates are used. The entries with an "NA" in the table indicate that the variable is not available from the dataset. It can be seen that there are noticeable differences between the naive estimates from the PRC sample and the survey weighted estimates from the two probability samples for variables such as Origin (Hispanic/Latino), Education (High school or less), Household (with children), Health (Smoking) and Volunteer work. It is a strong evidence that the PRC dataset is not a representative simple random sample from the target population. The estimates from the two reference probability samples, on the other hand, agree to each other for almost all the variables.

There is a common set of auxiliary variables which are available in all three samples, as shown in the top part of Table 17.1. We use this common set of variables to compute estimates for the study variable y, with either CPS or BRFSS as the reference probability sample. For each y, we compute the IPW estimator $\hat{\mu}_{y\text{IPW}}$, the doubly robust estimator $\hat{\mu}_{y\text{DR}}$, and the mass imputation estimator $\hat{\mu}_{y\text{REG}}$ based on (17.1) and the linear regression or the logistic regression model. The naive simple sample mean based on the PRC non-probability sample is also included and denoted as $\hat{\mu}_{PRC}$. Estimates for binary study variables are reported in percentages (%). The results are presented in Table 17.2.

Table 17.1 Marginal distributions of common auxiliary variables in the three survey samples

		$\hat{\mu}_{PRC}$	$\hat{\mu}_{BRFSS}$	$\hat{\mu}_{CPS}$
Age category	<30	18.29	20.91	21.24
	>=30, <50	32.60	33.28	33.55
	>=50, <70	38.68	32.69	32.60
	>=70	10.43	13.12	12.62
Gender	Female	54.36	51.32	51.83
Race	White only	82.28	75.05	78.61
	Black only	8.83	12.59	12.46
Origin	Hispanic/Latino	9.27	16.52	15.60
Region	Northeast	19.96	17.70	18.00
	South	27.53	38.27	37.34
	West	29.87	23.18	23.47
Marital status	Married	50.35	50.82	52.82
Employment	Working	52.13	56.63	58.90
	Retired	24.34	17.93	14.34
Education	High school or less	21.63	42.66	40.65
	Bachelor's degree and above	41.57	26.32	30.90
	Bachelor's degree	22.07	NA	19.83
	Postgraduate	19.50	NA	11.07
Household	Presence of child in household	28.93	36.78	NA
	Home ownership	65.42	67.19	NA
Health	Smoke everyday	15.74	11.49	NA
	Smoke never	79.80	83.28	NA
Financial status	No money to see doctors	20.68	13.27	NA
	Having medical insurance	89.15	87.83	NA
	Household income <20K	16.14	NA	15.32
	Household income >100K	19.89	NA	23.32
Volunteer works	Volunteered	50.98	NA	24.83

There are two separate larger sets of auxiliary variables which are common between PRC and CPS and between PRC and BRFSS. For instance, the common set of auxiliary variables between PRC and CPS include all variables which are available in all three datasets, plus five more variables which are available from CPS but not BRFSS, as shown in Table 17.1. Estimates for the population means using the larger set of common auxiliary variables for each reference probability sample are presented in Table 17.3.

There are four major observations from the results presented in Tables 17.2 and 17.3, which shed light on practical applications of the estimation methods introduced in this chapter for non-probability survey samples. First, two different reference probability survey samples lead to very similar results if the set of auxiliary variables used in estimation is the same for both reference samples. Second, two different sets of auxiliary variables could produce quite different

Table 17.2 Estimated population mean (%) using a single set of common auxiliary variables

Study variable y		$\hat{\mu}_{PRC}$	$\hat{\mu}_{y\text{IPW}}$	$\hat{\mu}_{y\text{REG}}$	$\hat{\mu}_{y\text{DR}}$
y_1: Talked with neighbours frequently	BRFSS	46.13	45.53	45.72	45.72
	CPS		45.41	45.76	45.79
y_2: Tended to trust neighbours	BRFSS	58.97	55.69	55.30	55.37
	CPS		55.53	55.66	55.71
y_3: Expressed opinions at a government level	BRFSS	26.54	24.23	23.98	24.17
	CPS		24.40	24.31	24.52
y_4: Voted in local elections	BRFSS	74.97	71.28	70.70	70.87
	CPS		71.38	71.59	71.69
y_5: Participated in school groups	BRFSS	20.97	20.18	20.02	20.16
	CPS		20.73	20.55	20.71
y_6: Participated in service organizations	BRFSS	14.11	13.26	13.28	13.29
	CPS		13.51	13.50	13.56
y_7: Days had at least one drink last month	BRFSS	5.30	4.95	4.93	4.97
	CPS		4.95	4.99	5.01

Table 17.3 Estimated population mean (%) using separate sets of common auxiliary variables

Study variable y		$\hat{\mu}_{PRC}$	$\hat{\mu}_{y\text{IPW}}$	$\hat{\mu}_{y\text{REG}}$	$\hat{\mu}_{y\text{DR}}$
y_1: Talked with neighbours frequently	BRFSS	46.13	44.23	44.64	44.63
	CPS		40.32	40.43	40.14
y_2: Tended to trust neighbours	BRFSS	58.97	56.07	56.06	56.04
	CPS		53.92	52.97	53.12
y_3: Expressed opinions at a government level	BRFSS	26.54	22.38	22.33	22.37
	CPS		20.17	19.91	19.91
y_4: Voted in local elections	BRFSS	74.97	71.84	71.51	71.77
	CPS		69.43	68.10	68.26
y_5: Participated in school groups	BRFSS	20.97	19.68	19.79	19.75
	CPS		13.74	13.32	13.12
y_6: Participated in service organizations	BRFSS	14.11	12.04	12.09	11.98
	CPS		8.98	8.68	8.54
y_7: Days had at least one drink last month	BRFSS	5.30	4.78	4.81	4.82
	CPS		5.07	5.06	5.09

estimation results, as shown in Table 17.3. Third, when the propensity score model and the outcome regression model both seem to be reasonable, the results from mass imputation, inverse probability weighting, and doubly robust estimation are all similar to each other. And lastly, the naive simple sample means from the non-probability survey sample behave very differently and are most likely unreliable.

There are unsettled issues with analysis of non-probability survey samples. The assumption **A1** is often difficult to verify in practice with the given set of common auxiliary variables available in the reference probability sample. A subjective judgement is required on whether or not essential auxiliary variables are included.

A sensitivity analysis might also be feasible by following a similar approach as the one proposed in Zhao et al. (2019). The assumption **A2** is another fundamental pillar for the methodologies described in this chapter. It amounts to stating that the sampling frame for the non-probability sample is complete and there are no coverage issues. In practice, this assumption is often untrue with non-probability samples, and accounting for departures from it requires further methodological development.

As non-probability survey samples and data from other unconventional sources have been increasingly used by statistical agencies and government organizations, there has been a lingering question of whether probability sampling methods and probability survey samples are still needed in the future. The theory and methods presented in this chapter indicate that a few large scale, high quality probability surveys with rich information on auxiliary variables can play indispensable roles for valid statistical inferences with non-probability survey samples. Probability sampling methods will remain as a major data collection tool for many fields of scientific study.

Appendix A
R Code for PPS Sampling, Empirical Likelihood and Regression Analysis

We include some practically useful R codes for unequal probability sampling procedures, pseudo empirical likelihood methods and regression analysis. These codes might be modified to fit into specific needs for survey design and analysis or for simulation studies.

A.1 Randomized Systematic PPS Sampling

The randomized systematic PPS sampling method is described in Sect. 4.4.2. The inputs for the R function are the size variable $x = (x_1, \cdots, x_N)'$ and the sample size n. The output of the R function is the set of n sampled units satisfying $\pi_i = nx_i/T_x$.

```
syspps=function(x,n){
##
##Order of population units is randomized first!
##
N=length(x)
U=sample(N,N)
xx=x[U]
z=rep(0,N)
for(i in 1:N) z[i]=n*sum(xx[1:i])/sum(x)
r=runif(1)
s=numeric()
for(i in 1:N){
if(z[i]>=r){
  s=c(s,U[i])
  r=r+1
         }    }
return(s[order(s)])
       }
```

C. Wu, M. E. Thompson, *Sampling Theory and Practice*, ICSA Book Series in Statistics, https://doi.org/10.1007/978-3-030-44246-0

A.2 Rao-Sampford PPS Sampling

The Rao-Sampford PPS sampling method is described in Sect. 4.4.3. The procedure selects a PPS sample with pre-specified first order inclusion probabilities π_i and has closed form expressions for the second order inclusion probabilities π_{ij}.

A.2.1 Select a PPS Sample

The inputs of the R function are the size variable $p = (p_1, \cdots, p_N)'$ and the sample size n. Assuming $p_i > 0$ and $\sum_{i=1}^N p_i = 1$, the first unit is selected with probability p_i and all $n - 1$ subsequent units are selected with probability $q_i = cp_i/(1 - np_i)$ from the population with replacement, where the constant c satisfies $\sum_{i=1}^N q_i = 1$. The *cumulative sum method* described at the beginning of Sect. 4.4 and a bisection search algorithm are used for selecting each individual unit. Samples with duplicated units are abandoned and the selection process starts anew. The numbers 1963 and 1989 appeared in the R function are arbitrary thresholds. The R function returns the set of n selected units for the final sample with $\pi_i = np_i$.

```
RSsample=function(p,n){
N=length(p)
p=p/sum(p)
lam=rep(0,N+1)
for(i in 1:N) lam[i+1]=lam[i]+p[i]
q=rep(0,N)
for(i in 1:N) q[i]=p[i]/(1-n*p[i])
q=q/sum(q)
lam2=rep(0,N+1)
for(i in 1:N) lam2[i+1]=lam2[i]+q[i]
ntot=1
while(ntot<1963){
sam=NULL
rand=runif(1)
dif=1
L=1
U=N+1
while(dif>0){
M=floor((U-L)/2)
if(lam[L+M]>rand) U=U-M
if(lam[L+M]<=rand) L=L+M
if(lam[U]>=rand & lam[U-1]<rand){
si=U-1
dif=0
      }
if(lam[L]<rand & lam[L+1]>=rand){
si=L
dif=0
      }
     }
sam=c(sam,si)
nn=0
while(nn<n-1){
dif=1
rand=runif(1)
L=1
U=N+1
while(dif>0){
M=floor((U-L)/2)
if(lam2[L+M]>rand) U=U-M
if(lam2[L+M]<=rand) L=L+M
if(lam2[U]>=rand & lam2[U-1]<rand){
si=U-1
dif=0
      }
if(lam2[L]<rand & lam2[L+1]>=rand){
si=L
dif=0
      }
     }
if(min(abs(si-sam))==0) nn=n+1
if(min(abs(si-sam))>0){
sam=c(sam,si)
nn=nn+1
          }
         }
if(nn==n+1) ntot=1
if(nn==n-1) ntot=1989
             }
return(sam)
           }
```

A.2.2 Compute Second Order Inclusion Probabilities

It is a relatively slow process to compute the second order inclusion probabilities for the Rao-Sampford sampling method with not-so-powerful computing machines. For simulation studies, it is recommended to calculate the $N \times N$ matrix for π_{ij} prior to repeated simulation runs with the given sample size n. The vector $p =$

$(p_1, \cdots, p_N)'$ is the normalized size variable. The final $N \times N$ matrix (π_{ij}) should be checked using $\sum_{i=1}^{N} \pi_{ii} = n$ and $\sum_{i=1}^{N} \sum_{j=1}^{N} \pi_{ij} = n^2$. The code works well when $N \leq 1000$, and can be modified to compute the $n \times n$ matrix for (ij) in the sample only. The line on *Lm3* is to make the next line fit into the space.

```
piij=matrix(0,N,N)
Lm=rep(0,n)
q=rep(0,N)
for(i in 1:N) q[i]=p[i]/(1-n*p[i])
Lm[1]=1
for(i in 2:n){
for(r in 1:(i-1)){
Lm[i]=Lm[i]+((-1)^(r-1))*sum(q^r)*Lm[i-r]
                              }
Lm[i]=Lm[i]/(i-1)
               }
t1=(n+1)-(1:n)
Kn=1/sum(t1*Lm/n^t1)
Lm2=rep(0,n-1)
t2=(1:(n-1))
t3=n-t2
for(i in 2:N){
for(j in 1:(i-1)){
Lm2[1]=1
Lm2[2]=Lm[2]-(q[i]+q[j])
for(m in 3:(n-1)){
Lm2[m]=Lm[m]-(q[i]+q[j])*Lm2[m-1]-q[i]*q[j]*Lm2[m-2]
                  }
Lm3=Lm2[t3]/n^(t2-1))
piij[i,j]=Kn*q[i]*q[j]*sum((t2+1-n*(p[i]+p[j]))*Lm3
piij[j,i]=piij[i,j]
                  }
piij[i,i]=n*p[i]
                }
piij[1,1]=n*p[1]
```

A.3 Empirical Likelihood Methods

We provide R computing programs for the algorithms described in Sect. 8.4 on pseudo empirical likelihood. The Newton-Raphson procedure implemented in Appendix A.3.2 can be modified for the calibration weighting methods discussed in Sect. 6.2 with suitably re-defined $\Delta_1(\lambda)$ and $\Delta_2(\lambda)$.

A.3.1 Lagrange Multiplier: Univariate Case

The R function for finding the vector-valued Lagrange multiplier as the solution to (8.18) is given in Appendix A.3.2, which also works for the univariate case described here. However, if the constraint only involves a single variable, i.e., finding $\hat{p}_i(\theta)$ under $\sum_{i\in\mathbf{S}} p_i(y_i-\theta)=0$ without any other calibration constraints or finding $\hat{p}_i$ under $\sum_{i\in\mathbf{S}} p_i(x_i-\mu_x)=0$ for a single x, the following R function based on a bi-section search method is extremely efficient. The inputs are design weights $ds=(d_1,\cdots,d_n)'$, $u=(u_1,\cdots,u_n)'$ and mu corresponding to the constraint $\sum_{i\in\mathbf{S}} p_i(u_i-mu)=0$. The output is the Lagrange multiplier λ.

```
Lag1=function(u,ds,mu){
dif=1
tol=1e-8
if(min(u-mu)>=0 | max(u-mu)<=0){
dif=0
M=0
    }
L=-1/max(u-mu)
R=-1/min(u-mu)
while(dif>tol){
M=(L+R)/2
glam=sum((ds*(u-mu))/(1+M*(u-mu)))
if(glam>0) L=M
if(glam<0) R=M
dif=abs(glam)
              }
return(M)
          }
```

A.3.2 Lagrange Multiplier: Vector Case

The R function *Lag2* below solves (8.18) for the vector-valued λ. The inputs are the data matrix $u=(\mathbf{x}_1,\cdots,\mathbf{x}_n)'$, design weights $ds=(d_1,\cdots,d_n)'$, and the vector mu in the constraints $\sum_{i\in\mathbf{S}} p_i(\mathbf{x}_i-mu)=\mathbf{0}$. The output is the Lagrange multiplier λ. The procedure is capped by 50 iterations and returns $\lambda=\mathbf{0}$ otherwise (a sign of non-convergence).

```
Lag2=function(u,ds,mu){
  n=length(ds)
  u=u-rep(1,n)%*%t(mu)
  M=0*mu
  dif=1
  tol=1e-8
  k=0
  while(dif>tol & k<=50){
  D1=t(u)%*%((ds/(1+u%*%M))*rep(1,n))
  DD=-t(u)%*%(c((ds/(1+u%*%M)^2))*u)
  D2=solve(DD,D1,tol=1e-40)
  dif=max(abs(D2))
  rule=1
  while(rule>0){
    rule=0
    if(min(1+t(M-D2)%*%t(u))<=0) rule=rule+1
    if(rule>0) D2=D2/2
               }
    M=M-D2
    k=k+1
                   }
  if(k>=50) M=0*mu
  return(as.vector(M))
          }
```

A.3.3 Lagrange Multiplier: Stratified Sampling

Finding the Lagrange multiplier as the solution to (8.21) for stratified samples involves preparation of the data into the right format. The required data are: (i) the strata sample sizes $n = (n_1, \cdots, n_{\mathrm{H}})'$; (ii) the data matrix $x = \{\mathbf{x}_{hi}, i = 1, \cdots, n_h, \;\; h = 1, \cdots, H\}$; (iii) the population means $X = \mu_{\mathbf{x}}$; (iv) the vector of design weights normalized within stratum: $\{\tilde{d}_{hi}(\mathbf{S}_h), i = 1, \cdots, n_h, \;\; h = 1, \cdots, H\}$; and (v) the vector of stratum weights $W = (W_1, \cdots, W_{\mathrm{H}})'$. The code below shows how to compute $\hat{p}_{hi}$.

```
nst=sum(n)
k=length(n)-1
ntot=rep(0,k)
   ntot[1]=n[1]
   for(j in 2:k) ntot[j]=ntot[j-1]+n[j]
ist=matrix(0,nst,k)
   ist[1:n[1],1]=1
   for(j in 2:k) ist[(ntot[j-1]+1):ntot[j],j]=1
uhi=cbind(ist,x)
mu=c(W[1:k],X)
whi=rep(W[1],n[1])
   for(j in 2:(k+1)) whi=c(whi,rep(W[j],n[j]))
dhi=whi*ds
```

```
M=Lag2(uhi,dhi,mu)
phi=as.vector(ds/(1+(uhi-rep(1,nst)%*%t(mu))%*%M))
```

A simple step to check the computation: all components of the final *phi* should be positive, with the first n_1 components summing up to 1, the next n_2 components summing up to 1, etc.

A.3.4 Empirical Likelihood Ratio Confidence Intervals

The following R code demonstrate how to find the 95% pseudo EL ratio confidence interval for μ_y in the absence of additional calibration constraints as described in Sect. 8.4.3 (Fig. 8.1). The *YEL* denotes the point estimator $\hat{\theta}_{\text{PEL}}$; the *ys* represents the observed data $(y_1, \cdots, y_n)$, the *ds* is the vector of design weights, and *nss* equals to $n/\hat{c}$, which is the effective sample size. The final interval $(\hat{\theta}_L, \hat{\theta}_U)$ is given by *(LB, UB)*.

```
tol=1e-08
cut=qchisq(0.95,1)
#------------
t1=YEL
t2=max(ys)
dif=t2-t1
while(dif>tol){
      tau=(t1+t2)/2
      M=Lag1(ys,ds,tau)
      elratio=2*nss*sum(ds*log(1+M*(ys-tau)))
      if(elratio>cut) t2=tau
      if(elratio<=cut) t1=tau
      dif=t2-t1
            }
UB=(t1+t2)/2
#------------
t1=YEL
t2=min(ys)
dif=t1-t2
while(dif>tol){
      tau=(t1+t2)/2
      M=Lag1(ys,ds,tau)
      elratio=2*nss*sum(ds*log(1+M*(ys-tau)))
      if(elratio>cut) t2=tau
      if(elratio<=cut) t1=tau
      dif=t1-t2
            }
LB=(t1+t2)/2
```

A.4 Survey Weighted Regression Analysis

The survey weighted regression analysis is discussed in Sect. 7.3. The R code is for single-stage sampling design with a small sampling fraction using the variance formula given in Part (c) of Problem 7.1.

A.4.1 Linear Regression Analysis

The first column of the design matrix X is a vector of 1's corresponding to the intercept. The ys is the vector of the response variable. The ds is the vector of design weights. The R function below returns point estimates and standard errors for each regression coefficient β_j, along with t-values and p-values for testing H_0: $\beta_j = 0$ versus H_1: $\beta_j \neq 0$.

```
SurveyRegression=function(ys,X,ds){
n=length(ys)
betahat=solve(t(X)%*%(c(ds)*X),t(X)%*%(ds*ys))
z=(1/ds)/n
VU=var(c(1/z)*(c(ys-X%*%betahat)*X))/n
Vbetahat=     ### The whole line does not fit here!
solve(t(X)%*%(c(ds)*X),VU)%*%solve(t(X)%*%(c(ds)*X))
SEbetahat=sqrt(diag(Vbetahat))
tvalue=betahat/SEbetahat
pvalue=2*(1-pnorm(abs(tvalue)))
out=list(betahat,SEbetahat,tvalue,pvalue)
names(out)=c("beta","SE","tvalue","pvalue")
return(out)
                       }
```

A.4.2 Logistic Regression Analysis

The setting is the same as in Appendix A.4.1 for linear regression, except that the ys is now a vector of 0's and 1's. The Newton-Raphson iterative procedure for solving (7.19) requires an initial value for β. One possible choice is to use the unweighted least square estimator from a linear regression model; the other is to simply set $\beta = \mathbf{0}$. The following R function returns the solution $\hat{\beta}$ to (7.19). Standard errors and test results can be developed in R using similar techniques in Appendix A.4.1.

```
LogReg=function(ys,X,ds){
b0=solve(t(X)%*%X,t(X)%*%ys)
##b0=rep(0,dim(X)[2])
ds=ds/sum(ds)
tol=0.00000001
dif=1
while(dif>tol){
mu=exp(X%*%b0)/(1+exp(X%*%b0))
b1=b0+     ## The whole line does not fit here!
solve(t(X)%*%(c(ds*mu*(1-mu))*X),t(X)%*%(ds*(ys-mu)))
dif=max(abs(b1-b0))
b0=b1
              }
b0=as.vector(b0)
return(b0)
             }
```

References

Aitkin, M. (2008). Applications of the Bayesian bootstrap in finite population inference. *Journal of Official Statistics, 24*, 21–51.

Alexander, C. H. (1987). A class of methods for using person controls in household weighting. *Survey Methodology, 13*, 183–198.

Asok, C., & Sukhatme, B. V. (1976). On Sampford's procedure of unequal probability sampling without replacement. *Journal of the American Statistical Association, 71*, 912–918.

Bahadur, R. R. (1966). A note on quantiles in large samples. *Annals of Mathematical Statistics, 37*, 577–580.

Baker, R., Brick, J. M., Bates, N. A., Battaglia, M., Couper, M. P., Dever, J. A., et al. (2013). Report of the AAPOR task force on non-probability sampling. *Journal of Survey Statistics and Methodology, 1*, 90–143.

Basu, D. (1971). An essay on the logical foundations of survey sampling. Part One. In V. P. Godambe & D. A. Sprott (Eds.), *Foundations of statistical inference* (pp. 203–242). Toronto: Holt, Rinehart and Winston.

Béland, Y., & St-Pierre, M. (2008). Mode effects in the Canadian Community Health Survey: A comparison of CAPI and CATI. In J. M. Lepkowski, N. C. Tucker, J. M. Brick, E. de Leeuw, L. Japec, P. J. Lavrakas, et al. (Eds.) *Advances in telephone survey methodology* (pp. 297–314). Hoboken: Wiley.

Bellhouse, D. R. (1988). Systematic sampling. In P. R. Krishnaiah & C. R. Rao (Eds.), *Handbook of statistics, Vol. 6: Sampling* (pp. 125–145). Amsterdam: North-Holland.

Benedetti, R., Bee, M., Espa, G., & Piersimoni, F. (2010). *Agricultural survey methods*. Chichester: Wiley Ltd.

Benedetti, R., Piersimoni, F., & Postiglione, P. (2015). *Sampling spatial units for agricultural surveys*. Heidelberg: Springer.

Berger, Y. G. (1998). Rate of convergence to normal distribution for the Horvitz-Thompson estimator. *Journal of Statistical Planning and Inference, 67*, 209–226.

Berger, Y. G., & De La Riva Torres, O. (2016). Empirical likelihood confidence intervals for complex sampling designs. *Journal of the Royal Statistical Society, Series B, 78*, 319–341.

Berger, Y. G., Tirari, E. H. M., & Tillé, Y. (2003). Towards optimal regression estimation in sample surveys. *Australian & New Zealand Journal of Statistics, 45*, 319–329.

Bickel, P. J., & Freedman, D. A. (1984). Asymptotic normality and the bootstrap in stratified sampling. *Annals of Statistics, 12*, 470–482.

Binder, D. A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review, 51*, 279–292.

C. Wu, M. E. Thompson, *Sampling Theory and Practice*, ICSA Book
Series in Statistics, https://doi.org/10.1007/978-3-030-44246-0

Binder, D. A., & Patak, Z. (1994). Use of estimating functions for estimation from complex surveys. *Journal of the American Statistical Association, 89*, 1035–1043.

Binder, D. A., & Roberts, G. (2009). Design- and model-based inference for model parameters. In D. Pfeffermann & C. R. Rao (Eds.), *Handbook of statistics, Volume 29B: Sample surveys: Inference and analysis* (pp. 33–54). Amsterdam: Elsevier.

Binson, D., Canchola, J. A., & Catania, J. A. (2000). Random selection in a national telephone survey: A comparison of the Kish, next-birthday, and last-birthday methods. *Journal of Official Statistics, 16*, 53–60.

Booth, J. G., Butler, R. W., & Hall, P. (1994). Bootstrap methods for finite populations. *Journal of the American Statistical Association, 89*, 1282–1289.

Brackstone, G. J., & Rao, J. N. K. (1979). An investigation of raking ratio estimators. *Sankhyā, Series C, 41*, 97–114.

Breidt, F. J., Claeskens, G., Opsomer, J. D. (2005). Model-assisted estimation for complex surveys using penalised splines. *Biometrika, 92*, 831–846.

Breidt, F. J., & Opsomer, J. D. (2000). Local polynomial regression estimators in survey sampling. *The Annals of Statistics, 28*, 1026–1053.

Brewer, K. R. W. (1963a). A model of systematic sampling with unequal probabilities. *Australian Journal of Statistics, 5*, 5–13.

Brewer, K. R. W. (1963b). Ratio estimation and finite populations: Some results deducible from the assumption of an underlying stochastic process. *Australian Journal of Statistics, 5*, 93–105.

Brewer, K. R. W. (1999). Cosmetic calibration with unequal probability sampling. *Survey Methodology, 25*, 205–212.

Brewer, K. R. W., & Hanif, M. (1983). *Sampling with unequal probabilities. Lecture notes in statistics* (Vol. 15). New York: Springer.

Brick, J. M., & Kalton, G. (1996). Handling missing data in survey research. *Statistical Methods in Medical Research, 5*, 215–238.

Buckland, S. T. (2004). *Advanced distance sampling*. Oxford: Oxford University Press.

Cantrell, J., Hair, E. C., Smith, A., Bennett, M., Miller Rath, J., Thomas, R. K., et al. (2018). Recruiting and retaining youth and young adults: Challenges and opportunities in survey research for tobacco control. *Tobacco Control, 27*, 147–154.

Carrillo, I. A., Chen, J., & Wu, C. (2010). The pseudo-GEE approach to the analysis of longitudinal surveys. *The Canadian Journal of Statistics, 38*, 540–554.

Carrillo, I. A., Chen, J., & Wu, C. (2011). A pseudo-GEE approach to analyzing longitudinal surveys under imputation for missing responses. *Journal of Official Statistics, 27*, 255–277.

Cassel, C. M., Särndal, C. E., & Wretman, J. H. (1976). Some results on generalized difference estimation and generalized regression estimation for finite populations. *Biometrika, 63*, 615–620.

Chang, T., & Kott, P. S. (2008). Using calibration weighting to adjust for nonresponse under a plausible model. *Biometrika, 95*, 555–571.

Chauvet, G. (2007). *Méthodes de bootstrap en population finie*. Ph.D. thesis, Université de Rennes 2.

Chen, J., Chen, S. Y., & Rao, J. N. K. (2003). Empirical likelihood confidence intervals for the mean of a population containing many zero values. *The Canadian Journal of Statistics, 31*, 53–68.

Chen, S., & Haziza, D. (2019). Recent development in dealing with item non-response in surveys: A critical review. *International Statistical Review, 87*, S192–218.

Chen, S., Haziza, D., Léger, C., & Mashreghi, Z. (2019). Pseudo-population bootstrap methods for imputed survey data. *Biometrika, 106*, 369–384.

Chen, H. Y., & Little, R. J. A. (1999). A test of missing completely at random for generalized estimating equations with missing data. *Biometrika, 86*, 1–13.

Chen, J., & Qin, J. (1993). Empirical likelihood estimation for finite populations and the effective usage of auxiliary information. *Biometrika, 80*, 107–116.

Chen, J., & Rao, J. N. K. (2007). Asymptotic normality under two-phase sampling designs. *Statistica Sinica, 17*, 1047–1064.

Chen, J., & Shao, J. (2000). Nearest neighbor imputation for survey data. *Journal of Official Statistics, 16*, 113–131.

Chen, J., & Shao, J. (2001). Jackknife variance estimation for nearest-neighbor imputation. *Journal of the American Statistical Association, 96*, 260–269.

Chen, J., & Sitter, R. R. (1999). A pseudo empirical likelihood approach to the effective use of auxiliary information in complex surveys. *Statistica Sinica, 9*, 385–406.

Chen, J., Sitter, R. R., & Wu, C. (2002). Using empirical likelihood methods to obtain range restricted weights in regression estimators for surveys. *Biometrika, 89*, 230–237.

Chen, J., Thompson, M. E., & Wu, C. (2004). Estimation of fish abundance indices based on scientific research trawl surveys. *Biometrics, 60*, 116–123.

Chen, J., & Wu, C. (2002). Estimation of distribution function and quantiles using the model-calibrated pseudo empirical likelihood method. *Statistica Sinica, 12*, 1223–1239.

Chen, M., Thompson, M. E., & Wu, C. (2018). Empirical likelihood methods for complex surveys with data missing-by-design. *Statistica Sinica, 28*, 2027–2048.

Chen, M., Wu, C., & Thompson, M. E. (2015). An Imputation-based empirical likelihood approach to pretest-posttest studies. *The Canadian Journal of Statistics, 43*, 378–402.

Chen, M., Wu, C., & Thompson, M. E. (2016). Mann-Whitney test with empirical likelihood methods for pretest-posttest studies. *Journal of Nonparametric Statistics, 28*, 360–374.

Chen, S., & Kim, J. K. (2014). Population empirical likelihood for nonparametric inference in survey sampling. *Statistica Sinica, 24*, 335–355.

Chen, Y., Li, P., & Wu, C. (2020). Doubly robust inference with non-probability survey samples. *Journal of the American Statistical Association* (to appear)

Chowdhury, S., Khare, M., & Wolter, K. (2007). Weight trimming in the national immunization survey. *Proceedings of the Section on Survey Research Methods, 2007* (pp. 2651–2658). Alexandria, VA: American Statistical Association.

Citro, C. F. (2014). From multiple modes for surveys to multiple data sources for estimates. *Survey Methodology, 40*, 137–161.

Cochran, W. G. (1939). The use of analysis of variance in enumeration by sampling. *Journal of the American Statistical Association, 34*, 492–510.

Cochran, W. G. (1953). *Sampling techniques* (1st ed.). New York: Wiley.

Cochran, W. G. (1977). *Sampling techniques* (3rd ed.). New York: Wiley.

Copeland, K. R., Peitzmeier, F. K., & Hoy, C. E. (1987). An alternative method of controlling current population survey estimates to population counts. *Survey Methodology, 13*, 173–181.

Cox, B. G., Binder, D., Chinnappa, B. N., Christianson, A., Colledge, M.J., & Kott, P. S. (1995). *Business survey methods*. New York: Wiley.

David, H. A. (1968). Gini's mean difference rediscovered. *Biometrika, 55*, 573–575.

Deming, W. E., & Stephan, F. F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *The Annals of Mathematical Statistics, 11*, 427–444.

Deville, J. C., & Särndal, C. E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association, 87*, 376–382.

Deville, J. C., Särndal, C. E., & Sautory, O. (1993). Generalized raking procedures in survey sampling. *Journal of the American Statistical Association, 88*, 1013–1020.

Dillman, D. A. (2017). The promise and challenge of pushing respondents to the Web in mixed-mode surveys. *Survey Methodology, 43*, 3–30.

Dillman, D., Phelps, G., Tortora, R., Swift, K., Kohrell, J., Berck, J., et al. (2009). Response rates and measurement differences in mixed-mode surveys: Using mail, telephone, interactive voice response and the internet. *Social Science Research, 38*, 1–18.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics, 7*, 1–26.

Efron, B. (1982). *The jackknife, the bootstrap and other resampling plans*. Philadelphia: Society for Industrial and Applied Mathematics.

Efron, B. (1994). Missing data, imputation, and the bootstrap. *Journal of the American Statistical Association, 89*, 463–479.

Elliott, M. R., & Valliant, R. (2017). Inference for nonprobability samples. *Statistical Science* **32**, 249–264.

Erdös, P., & Rényi, A. (1959). On the central limit theorem for samples from a finite population. *Publication of the Mathematical Institute of the Hungarian Academy of Sciences, 4*, 49–61.

Ericson, W. A. (1969). Subjective Bayesian models in sampling finite populations, I. *Journal of the Royal Statistical Society, Series B, 31*, 195–234.

Fan, J., & Gijbels, I. (1996). *Local polynomial modelling and its applications*. New York: Chapman & Hall/CRC.

Fan, J., & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association, 96*, 1348–60.

Fay, R. E. (1984). Some properties of estimators of variance based on replication methods. *Proceedings of the Survey Research Methods Section* (pp. 495–500). Alexandria, VA: American Statistical Association.

Fay, R. E. (1992). When are inferences from multiple imputation valid? *Proceedings of the Survey Research Methods Section* (pp. 227–232). American Statistical Association, Alexandria, VA.

Fay, R. E. (1996). Alternative paradigms for the analysis of imputed survey data. *Journal of the American Statistical Association, 91*, 490–498.

Fay, R. E., & Dippo, C. S. (1989). Theory and application of replicate weighting for variance calculations. *Proceedings of the Survey Research Methods Section* (pp. 212–217). Alexandria, VA: American Statistical Association.

Fitzmaurice, G. M., Laird, N. M., & Ware, J. H. (2004). *Applied longitudinal analysis.* New York: Wiley.

Folsom, R. E. (1991). Exponential and logistic weight adjustment for sampling and nonresponse error reduction. *Proceedings of the Section on Social Statistics* (pp. 197–202). Alexandria, VA: American Statistical Association.

Francisco, C. A., & Fuller, W. A. (1991). Quantile estimation with a complex survey design. *Annals of Statistics, 19*, 454–469.

Frank, O., & Snijders, T. (1994). Estimating the size of hidden populations using snowball sampling. *Journal of Official Statistics, 10*, 53–67.

Frieze, A., & Karoński, M. (2016). *Introduction to random graphs*. Cambridge: Cambridge University Press.

Fu, Y., Wang, X., & Wu, C. (2009). Weighted empirical likelihood inference for multiple samples. *Journal of Statistical Planning and Inference, 139*, 1462–1473.

Fuller, W. A. (2002). Regression estimation for survey samples. *Survey Methodology, 28*, 5–23.

Fuller, W. A. (2009). *Sampling statistics*. Hoboken, NJ: Wiley.

Fuller, W. A., & Burmeister, L. F. (1972). Estimators for samples selected from two overlapping frames. *Proceedings of the Social Statistics Section* (pp. 245–249). Alexandria, VA: American Statistical Association.

Gastwirth, J. L. (1971). A general definition of Lorenz curve. *Econometrica, 39*, 1037–1039.

Gelman, A. (2007). Struggles with survey weighting and regression modeling. *Statistical Science, 22*, 153–164.

Gile, K. J. (2011). Improved inference for respondent-driven sampling data with application to HIV prevalence estimation. *Journal of the American Statistical Association, 106*, 135–146.

Godambe, V. P. (1955). A unified theory of sampling from finite populations. *Journal of the Royal Statistical Society, Series B, 17*, 269–278.

Godambe, V. P. (1960). An optimum property of regular maximum likelihood estimation. *Annals of Mathematical Statistics, 31*, 1208–1212.

Godambe, V. P. (1966). A new approach to sampling from finite populations. *Journal of the Royal Statistical Society, Series B, 28*, 310–328.

Godambe, V. P. (1968). Bayesian sufficiency in survey-sampling. *Annals of Mathematical Statistics, 20*, 363–373.

Godambe, V. P. (Ed.) (1991a). *Estimating functions*. New York: Oxford University Press.

Godambe, V. P. (1991b). Orthogonality of estimating functions and nuisance parameters. *Biometrika, 78*, 143–151.

Godambe, V. P., & Thompson, M. E. (1971). Bayes, fiducial and frequency aspects of statistical inference in regression analysis in survey-sampling. *Journal of the Royal Statistical Society, Series B, 33*, 361–390.

Godambe, V. P., & Thompson, M. E. (1973). Estimation in sampling theory with exchangeable prior distributions. *Annals of Statistics, 1*, 1212–1221.

Godambe, V. P., & Thompson, M. E. (1986). Parameters of superpopulation and survey population: Their relationships and estimation. *International Statistical Review, 54*, 127–138.

Godambe, V. P., & Thompson, M. E. (1989). An extension of quasi-likelihood estimation. *Journal of Statistical Planning and Inference, 22*, 137–152.

Godambe, V. P., & Thompson, M. E. (1999). A new look at confidence intervals in survey sampling. *Survey Methodology, 25*, 161–173.

Godambe, V. P., & Thompson, M. E. (2009). Estimating functions and survey sampling. In D. Pfeffermann & C. R. Rao (Eds.), *Handbook of statistics, Volume 29B: Sample surveys: Inference and analysis* (pp. 83–101). Amsterdam: Elsevier.

Goodman, R., & Kish, L. (1950). Controlled selection - A technique in probability sampling. *Journal of the American Statistical Association, 45*, 350–372.

Govindarajulu, Z. (1999). *Elements of sampling theory and methods*. Upper Saddle River, NJ: Prentice Hall.

Goyder, J. (1987). *The silent minority: Nonrespondents on sample surveys*. Boulder, CO: Westview Press.

Gregoire, T. G., & Valentine, H. T. (2007). *Sampling Strategies for Natural Resources and the Environment*. Boca Raton: Chapman & Hall/CRC.

Gross, S. (1980). Median estimation in sample surveys. *Proceedings of the Survey Research Methods Section* (pp. 181–184). Alexandria, VA: American Statistical Association.

Gunderson, D. R. (1993). *Surveys of fisheries resources*. New York: Wiley.

Hájek, J. (1960). Limiting distributions in simple random sampling from a finite population. *Publication of the Mathematical Institute of the Hungarian Academy of Sciences, 5*, 361–374.

Hájek, J. (1964). Asymptotic theory of rejective sampling with varying probabilities from a finite population. *Annals of Mathematical Statistics, 35*, 1491–1523.

Hájek, J. (1971). Discussion of "An essay on the logical foundations of survey sampling, Part One" by D. Basu. In V. P. Godambe & D. A. Sprott (Eds.) *Foundations of statistical inference* (Vol. 236). Toronto: Holt, Rinehart and Winston.

Hájek, J. (1981). *Sampling from a finite population*. New York: Marcel Dekker.

Haldane, J. B. S. (1945). On a method of estimating frequencies. *Biometrika, 33*, 222–225.

Han, P. (2014). Multiply robust estimation in regression analysis with missing data. *Journal of the American Statistical Association, 109*, 1159–1173.

Han, P., & Wang, L. (2013). Estimation with missing data: Beyond double robustness. *Biometrika, 100*, 417–430.

Handcock, M. S., & Gile, K. J. (2010). Modeling social networks from sampled data. *Annals of Applied Statistics, 4*, 5–25.

Hansen, M. H., & Hurwitz, W. N. (1943). On the theory of sampling from finite populations. *Annals of Mathematical Statistics, 14*, 333–362.

Hansen, M. H., Hurwitz, W. N., & Madow, W. G. (1953). *Sample survey methods and theory* (Vols. I and II). New York: Wiley.

Hansen, M. H., Madow, W. G., & Tepping, B. J. (1983). An evaluation of model-dependent and probability-sampling inferences in sample surveys. *Journal of the American Statistical Association, 78*, 776–793.

Hartley, H. O. (1962). Multiple frame surveys. *Proceedings of the Social Statistics Section* (pp. 203–206). Alexandria, VA: American Statistical Association.

Hartley, H. O. (1974). Multiple frame methodology and selected applications. *Sankhyā, Series C, 36*, 99–118.

Hartley, H. O., & Rao, J. N. K. (1962). Sampling with unequal probabilities and without replacement. *Annals of Mathematical Statistics, 33*, 350–374.

Hartley, H. O., & Rao, J. N. K. (1968). A new estimation theory for sample surveys. *Biometrika, 55*, 547–557.

Hartley, H. O., & Rao, J. N. K. (1969). A new estimation theory for sample surveys, II. In N. L. Johnson & H. Smith (Eds.), *New developments in survey sampling* (pp. 147–169). New York: Wiley.

Haziza, D. (2009). Imputation and inference in the presence of missing data. In D. Pfeffermann & C. R. Rao (Eds.) *Handbook of Statistics, Vol. 29A: Sample Surveys: Design, Methods and Applications* (pp. 215–246). Amsterdam: Elsevier.

Haziza, D., & Beaumont, J.-F. (2007). On the construction of imputation classes in surveys. *International Statistical Review, 75*, 25–43.

Haziza, D., & Beaumont, J.-F. (2017). Construction of weights in surveys: A review. *Statistical Science, 32*, 206–226.

Haziza, D., & Lesage, E. (2016). A discussion of weighting procedures for unit nonresponse. *Journal of Official Statistics, 32*, 129–145.

Haziza, D., Mecatti, F., & Rao, J. N. K. (2008). Evaluation of some approximate variance estimators under the Rao-Sampford unequal probability sampling design. *Metron, 66*, 91–108.

Heckathorn, D. D. (1997). Respondent-driven sampling: A new approach to the study of hidden populations. *Social Problems, 44*, 174–199.

Heeringa, S. G., West, B. T., & Berglund, P. A. (2017). *Applied survey data analysis* (2nd ed.). New York: Chapman & Hall/CRC.

Hoadley, B. (1969). The compound multinomial distribution and Bayesian analysis of categorical data from finite populations. *Journal of the American Statistical Association, 64*, 216–229.

Holmberg, A. (1998). A bootstrap approach to probability proportional to size sampling. *Proceedings of the Survey Research Methods Section* (pp. 378–383). Alexandria, VA: American Statistical Association.

Horvitz, D. G., & Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association, 47*, 663–685.

Hu, F., & Kalbfleisch, J. D. (2000). The estimating function bootstrap. *The Canadian Journal of Statistics, 28*, 449–499.

Huang, E. T., & Fuller, W. A. (1978). Nonnegative regression estimation for sample survey data. *Proceedings of the Section on Social Statistics* (pp. 300–305). Alexandria, VA: American Statistical Association.

Huang, Y. C., Thompson, M. E., Boudreau, C., & Fong, G. T. (2010). Accounting for the effects of data collection modes in population surveys. In *Proceedings of Statistics Canada's XXVI International Methodology Symposium*.

Ireland, C. T., & Kullback, S. (1968). Contingency tables with given marginals. *Biometrika, 55*, 179–188.

Isaki, C. T., & Fuller, W. A. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association, 77*, 89–96.

Kalton, G., & Anderson, D. W. (1986). Sampling rare populations. *Journal of the Royal Statistical Society, Series A, 149*, 65–82.

Kalton, G., & Kish, L. (1984). Some efficient random imputation methods. *Communications in Statistics, A, 13*, 1919–1939.

Kim, J. K. (2010). Calibration estimation using exponential tilting in sample surveys. *Survey Methodology, 36*, 145–155.

Kim, J. K., & Fuller, W. A. (2004). Fractional hot deck imputation. *Biometrika, 91*, 559–578.

Kim, J. K., & Haziza, D. (2014). Doubly robust inference with missing data in survey sampling. *Statistica Sinica, 24*, 375–394.

Kim, J. K., & Kim, J. J. (2007). Nonresponse weighting adjustment using estimated response probability. *The Canadian Journal of Statistics, 35*, 501–514.

Kim, J. K., & Park, M. (2010). Calibration estimation in survey sampling. *International Statistical Review, 78*, 21–39.

Kim, J. K., Park, S., Chen, Y., & Wu, C. (2019). Combining non-probability and probability survey samples through mass imputation. Available at https://arxiv.org/abs/1812.10694

Kim, J. K., & Rao, J. N. K. (2012). Combining data from two independent surveys: A model-assisted approach. *Biometrika, 99*, 85–100.

Kim, J. K., & Shao, J. (2013). *Statistical methods for handling incomplete data*. Boca Raton: CRC Press, Taylor & Francis Group.

Kim, J. K., & Wu, C. (2013). Sparse and efficient replication variance estimation for complex surveys. *Survey Methodology* **39**, 91–120.

Koenker, R., & Bassett, G. (1978). Regression quantiles. *Econometrica, 46*, 33–50.

Kott, P. S. (1994). A note on handling nonresponse in sample surveys.*Journal of the American Statistical Association, 89*, 693–696.

Kott, P. S. (2003). A practical use for instrumental-variable calibration. *Journal of Official Statistics, 19*, 265–272.

Kott, P. S., & Chang, T. (2010). Using calibration weighting to adjust for nonignorable unit nonresponse. *Journal of the American Statistical Association, 105*, 1265–1275.

Kreuter, F., Presser, S., & Tourangeau, R. (2008). Social desirability bias in CATI, IVR and web surveys: The effects of mode and question sensitivity. *Public Opinion Quarterly, 72*, 847–865.

Krewski, D., & Rao, J. N. K. (1981). Inference from stratified samples: Properties of linearization, jackknife and balanced repeated replication methods. *Annals of Statistics, 9*, 1010–1019.

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics, 22*, 79–86.

Lavallée, P. (1995). Cross-sectional weighting of longitudinal surveys of individuals and households using the weight share method. *Survey Methodology, 21*, 25–32.

Lavallée, P. (2007). *Indirect sampling*. New York: Springer.

Lazar, N. A. (2003). Bayesian empirical likelihood. *Biometrika, 90*, 319–326.

Lee, S. (2006). Propensity score adjustment as a weighting scheme for volunteer panel web surveys. *Journal of Official Statistics* **22**, 329–349.

Lee, S., & Valliant, R. (2009). Estimation for volunteer panel web surveys using propensity score adjustment and calibration adjustment. *Sociological Methods* & *Research, 37*, 319–343.

Lemaitre, G., & Dufour, J. (1987). An integrated method for weighting persons and families. *Survey Methodology, 13*, 199–207.

Leng, C., & Tang, C. Y. (2012). Penalized empirical likelihood and growing dimensional general estimating equations. *Biometrika, 99*, 703–716.

Liang, K.-Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika, 73*, 13–22.

Link, M. W., Battaglia, M. P., Frankel, M. R., Osborn, L., & Mokdad, A.H. (2008). A comparison of address-based sampling (ABS) versus random-digit dialing (RDD) for general population surveys. *Public Opinion Quarterly, 72*, 6–27.

Little, R. J. A. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association, 83*, 1198–1202.

Little, R. J. A. (2011). Calibrated Bayes, for statistics in general, and missing data in particular. *Statistical Science, 26*, 162–174.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). New York: Wiley

Liu, T. P., & Thompson, M. E. (1983). Properties of estimators of quadratic finite population functions: The batch approach. *The Annals of Statistics, 11*, 275–285.

Lohr, S. L. (2010). *Sampling: Design and analysis* (2nd ed.). San Francisco, CA: Brooks/Cole, Cengage Learning.

Lohr, S. L., & Rao, J. N. K. (2000). Inference from dual frame surveys. *Journal of the American Statistical Association, 95*, 271–280.

Lohr, S. L., & Rao, J. N. K. (2006). Estimation in multiple-frame surveys. *Journal of the American Statistical Association, 101*, 1019–1030.

Lorenz, M. O. (1905). Methods of measuring concentration of wealth. *Journal of the American Statistical Association, 9*, 209–219.

Lu, W. W., & Sitter, R. R. (2008). Disclosure risk and replication-based variance estimation. *Statistica Sinica, 18*, 1669–1687.

Luery, D. (1980). An alternative to principal person weighting. *Internal Memorandum*. Suitland: Bureau of the Census.

Luery, D. (1986). Weighting survey data under linear constraints on the weights. *Proceedings of the Section on Survey Research Methods* (pp. 325–330). Alexandria, VA: American Statistical Association.

Lumley, T., Shaw, P. A., & Dai, J. Y. (2011). Connections between survey calibration estimators and semiparametric models for incomplete data. *International Statistical Review, 79*, 200–220.

Lundström, S., & Särndal, C. E. (1999). Calibration as a standard method for the treatment of nonresponse. *Journal of Official Statistics, 15*, 305–327.

Madow, W. G. (1948). On the limiting distributions of estimates based on samples from finite universe. *Annals of Mathematical Statistics, 19*, 535–545.

Mashreghi, Z., Haziza, D., & Léger, C. (2016). A survey of bootstrap methods in finite population sampling. *Statistics Surveys, 10*, 1–52.

McCarthy, P. J. (1966). Replication: An approach to the analysis of data from complex surveys. In *Vital and health statistics* (Ser. 2, No. 14). Washington, DC: U.S. Government Printing Office.

McCarthy, P. J. (1969). Pseudo-replication: Half samples. *Review of the International Statistical Institute, 37*, 239–264.

McCarthy, P. J., & Snowden, C. B. (1985). The bootstrap and finite population sampling. In *Vital and health statistics* (Ser. 2, No. 95). Public health service publication (pp. 85–1369). Washington, DC: U.S. Government Printing Office.

McCullagh, P., & Nelder, J. (1983). *Generalized linear models*. London: Chapman & Hall.

Mecatti, F. (2007). A single frame multiplicity estimator for multiple frame survey. *Survey Methodology, 33*, 151–157.

Meeden, G., & Vardeman, S. (1991) A non-informative Bayesian approach to interval estimation in finite population sampling. *Journal of the American Statistical Association, 86*, 972–980.

Meng, X.-L. (1994). Multiple-imputation inferences with uncongenial sources of input. *Statistical Science, 9*, 538–573.

Midzuno, H. (1952). On the sampling system with probability proportional to sum of sizes. *Annals of the Institute of Statistical Mathematics, 3*, 99–107.

Molina, C. E. A., & Skinner, C. J. (1992). Pseudo-likelihood and quasi-likelihood estimation for complex sampling schemes. *Computational Statistics & Data Analysis, 13*, 395–405.

Molloy, M., & Reed, B. (1995). A critical point for random graphs with a given degree sequence. *Random Structures and Algorithms, 6*, 161–180.

Montanari, G. E., & Ranalli, M. G. (2005). Nonparametric model calibration estimation in survey sampling. *Journal of the American Statistical Association, 100*, 1429–1442.

Murthy, M. N. (1967). *Sampling theory and methods*. Calcutta: Statistical Publishing Society.

Narain, R. D. (1951). On sampling without replacement with varying probabilities. *Journal of the Indian Society of Agricultural Statistics, 3*, 169–174.

Narisetty, N. N., & He, X. (2014). Bayesian variable selection with shrinking and diffusing priors. *Annals of Statistics, 42*, 789–817.

Newey, W. K., & McFadden, D. (1994). Large sample estimation and hypothesis testing. *Handbook of Econometrics, 4*, 2111–2245.

Neyman, J. (1934). On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society, 97*, 558–606.

Norris, D. A., & Paton, D. G. (1991). Canada's General Social Survey: Five years of experience. *Survey Methodology, 17*, 227–240.

Oguz-Alper, M., & Berger, Y. G. (2016). Modelling complex survey data with population level Information: An empirical likelihood approach. *Biometrika, 103*, 447–459.

Owen, A. B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika, 75*, 237–249.

Owen, A. B. (2001). *Empirical likelihood*. New York: Chapman & Hall/CRC.

Park, M., & Fuller, W. A. (2005). Towards nonnegative regression weights for survey samples. *Survey Methodology, 31*, 85–93.

Park, S., & Kim, J. K. (2014). Instrumental-variable calibration estimation in survey sampling. *Statistica Sinica, 24*, 1001–1015.

Potter, F. (1990). A study of procedures to identify and trim extreme sampling weights. *Proceedings of the Section on Survey Research Methods* (pp. 225–230). Alexandria, VA: American Statistical Association.

Pfeffermann, D. (1993). The role of sampling weights when modeling survey data. *International Statistical Review, **61***, 317–337.

Qin, J., & Lawless, J. F. (1994). Empirical likelihood and general estimating equations. *Annals of Statistics, 22*, 300–325.

Quenouille, M. (1956). Notes on bias in estimation. *Biometrika, 43*, 353–360.

Raghunathan, T. E., Lepkowski, J. M., van Hoewyk, J., & Solenberger, P. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology, 27*, 85–96.

Rao, J. N. K. (1965). On two simple schemes of unequal probability sampling without replacement. *Journal of the Indian Statistical Association, 3*, 173–180.

Rao, J. N. K. (1966a). Alternative estimators in PPS sampling for multiple characteristics. *Sankhyā, Series A, 28*, 47–60.

Rao, J. N. K. (1966b). On the relative efficiency of some estimators in PPS sampling for multiple characteristics. *Sankhyā, Series A, 28*, 61–70.

Rao, J. N. K. (1968). Some nonresponse sampling theory when the frame contains an unknown amount of duplication. *Journal of the American Statistical Association, 63*, 87–90.

Rao, J. N. K. (1998). Marginal models for repeated observations: Inference with survey data. *Proceedings of the Section on Survey Research Methods* (pp. 76–82). Alexandria, VA: American Statistical Association.

Rao, J. N. K. (2005). Interplay between sample survey theory and practice: An appraisal. *Survey Methodology, 31*, 117–138.

Rao, J. N. K. (2011). Impact of frequentist and Bayesian methods on survey sampling practice: A selective appraisal. *Statistical Science, 26*, 240–256.

Rao, J. N. K., & Ghangurde, P. D. (1972). Bayesian optimization in sampling finite populations. *Journal of the American Statistical Association, 67*, 439–443.

Rao, J. N. K., Hartley, H. O., & Cochran, W. G. (1962). On a simple procedure of unequal probability sampling without replacement. *Journal of the Royal Statistical Society, Series B, 24*, 482–491.

Rao, J. N. K., & Molina, I. (2015). *Small area estimation* (2nd ed.). Hoboken, NJ: Wiley.

Rao, J. N. K., & Shao, J. (1992). Jackknife variance estimation with survey data under hot deck imputation. *Biometrika, 79*, 811–822.

Rao, J. N. K., & Shao, J. (1996). On balanced half-sample variance estimation in stratified random sampling. *Journal of the American Statistical Association, 91*, 343–348.

Rao, J. N. K., & Shao, J. (1999). Modified balanced repeated replication for complex survey data. *Biometrika, 86*, 403–415.

Rao, J. N. K., & Singh, A. C. (1997). A ridge-shrinkage method for range-restricted weight calibration in survey sampling. *Proceedings of the Section on Survey Research Methods* (pp. 57–65). Alexandria, VA: American Statistical Association.

Rao, J. N. K., & 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., & Wu, C. F. J. (1988). Resampling inference with complex survey data. *Journal of the American Statistical Association, 83*, 231–241.

Rao, J. N. K., & Wu, C. (2009). Empirical likelihood methods. In D. Pfeffermann & C. R. Rao (Eds.), *Handbook of Statistics, Vol. 29B: Sample Surveys: Inference and Analysis* (pp. 189–207). Amsterdam: Elsevier.

Rao, J. N. K., & Wu, C. (2010a). Bayesian pseudo empirical likelihood intervals for complex surveys. *Journal of the Royal Statistical Society, Series B, 72*, 533–544.

Rao, J. N. K., & Wu, C. (2010b). Pseudo empirical likelihood inference for multiple frame surveys. *Journal of the American Statistical Association, 105*, 1494–1503.

Rao, J. N. K., Wu, C. F. J., & Yue, K. (1992). Some recent work on resampling methods for complex surveys. *Survey Methodology, 18*, 209–217.

Rao, P. S. R. S., & Rao, J. N. K. (1971). Small sample results for ratio estimation. *Biometrika, 58*, 625–630.

Rivers, D. (2007). Sampling for web surveys. *Proceedings of the Survey Research Methods Section* (pp. 1–26). Alexandria, VA: American Statistical Association.

Robins, J. M., Rotnitzky, A., & Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association, 89*, 846–866.

Robins, J. M., & Wang, N. (2000). Inference for imputation estimators. *Biometrika, 87*, 113–124.

Roman, A. (1982). Proposal for weighting research in the consumer expenditure surveys. *Internal Memorandum*. Suitland: Bureau of the Census.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*, 41–55.

Royall, R. M. (1970). On finite population sampling theory under certain linear regression models. *Biometrika, 57*, 377–387.

Royall, R. M., & Cumberland, W. G. (1978). An empirical study of prediction theory in finite population sampling: simple random sampling and the ratio estimator. In N. K. Namboodiri (Ed.), *Survey sampling and measurements* (pp. 293–309). New York: Academic.

Royall, R. M., & Pfeffermann, D. (1982). Balanced samples and robust Bayesian inference in finite population sampling. *Biometrika, 69*, 401–409.

Rubin, D. B. (1976). Inference and missing data. *Biometrika, 63*, 581–590.

Rubin, D. B. (1978). Multiple imputations in sample surveys - A phenomenological Bayesian approach to nonresponse. In *Proceedings of the Survey Research Methods Section* (pp. 20–28). Alexandria, VA: American Statistical Association.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: Wiley.

Saigo, H. (2010). Comparing four bootstrap methods for stratified three-stage sampling. *Journal of Official Statistics, 26*, 193–207.

Saigo, H., Shao, J., & Sitter, R. R. (2001). A repeated half-sample bootstrap and balanced repeated replications for randomly imputed data. *Survey Methodology, 27*, 189–196.

Sampford, M. R. (1967). On sampling without replacement with unequal probabilities of selection. *Biometrika, 54*, 499–513.

Särndal, C.-E. (1980). On π-inverse weighting versus best linear unbiased weighting in probability sampling. *Biometrika, 67*, 639–650.

Särndal, C.-E. (2007). The calibration approach in survey theory and practice. *Survey Methodology, 33*, 99–119.

Särndal, C.-E., & Lundström, S. (2005). *Estimation in surveys with nonresponse*. Chichester: Wiley Ltd.

Särndal, C.-E., Swensson, B., & Wretman, J. H. (1992). *Model assisted survey sampling*. New York: Springer.

Scharfstein, D. O., Rotnitzky, A., & Robins, J. M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association, 94*, 1096–1120.

Schnute, J. T. (1994). A general framework for developing sequential fisheries models. *Canadian Journal of Fisheries and Aquatic Science, 51*, 1676–1688.

Scott, A., & Smith, T. M. F. (1969). A note on estimating secondary characteristics in multivariate surveys. *Sankhyā, Series A, 31*, 497–498.

Seber, G. A. F. (1982). *The estimation of animal abundance and related parameters*. Caldwel, NJ: Blackburn Press.

Sen, A. R. (1953). On the estimate of the variance in sampling with varying probabilities. *Journal of the Indian Society of Agricultural Statistics, 5*, 119–127.

Sen, P. K. (1988). Asymptotics in finite population sampling. In P. R. Krishnaiah & C. R. Rao (Eds.), *Handbook of statistics, Vol. 6: Sampling* (pp. 291–331). Amsterdam: North-Holland.

Shao, J. (2009). Nonparametric variance estimation for nearest neighbor imputation. *Journal of Official Statistics, 25*, 55–62.

Shao, J., & Sitter, R. R. (1996). Bootstrap for imputed survey data. *Journal of the American Statistical Association, 91*, 1278–1288.

Shao, J., & Tu, D. (1995). *The jackknife and bootstrap*. Springer series in statistics. New York: Springer.

She, X., & Wu, C. (2019). Fully efficient joint fractional imputation for incomplete bivariate ordinal responses. *Statistica Sinica, 29*, 409–430.

She, X., & Wu, C. (2020). Validity and efficiency in analyzing ordinal responses with missing observations. *The Canadian Journal of Statistics* (in press).

Shook-Sa, B. E., Currivan, D., Roe, D., & Klein Warren, L. (2016). Random digit dialing versus address-based sampling using telephone data collection. *Survey Practice, 9*. https://doi.org/10.29115/SP-2016-0015.

Sitter, R. R. (1992a). A resampling procedure for complex survey data. *Journal of the American Statistical Association, 87*, 755–765.

Sitter, R. R. (1992b). Comparing three bootstrap methods for survey data. *The Canadian Journal of Statistics, 20*, 135–154.

Sitter, R. R., & Wu, C. (2001). A note on Woodruff confidence intervals for quantiles. *Statistics & Probability Letters, 52*, 353–358.

Sitter, R. R., & Wu, C. (2002). Efficient estimation of quadratic finite population functions in the presence of auxiliary information. *Journal of the American Statistical Association, 97*, 535–543.

Skinner, C. J., & Rao, J. N. K. (1996). Estimation in dual frame surveys with complex designs. *Journal of the American Statistical Association, 91*, 349–356.

Snijders, T. A. B., Pattison, P., Robins, G. L., & Handcock, M. S. (2006). New specifications for Exponential Random Graph Models. *Sociological Methodology, 36*, 99–153.

Snijkers, G., Haraldsen, G., Jones, J., & Willimack, D. (2013). *Designing and conducting business surveys*. New York: Wiley.

Stephan, F. F. (1942). An iterative method of adjusting sample frequency tables when expected marginal totals are known. *Annals of Mathematical Statistics, 13*, 166–178.

Sukhatme, P. V. (1954). *Sampling theory of surveys, with applications*. Ames, IA: Iowa State College Press.

Tan, Z. (2013). Simple design-efficient calibration estimators for rejective and high-entropy sampling. *Biometrika, 100*, 399–415.

Tan, Z., & Wu, C. (2015). Generalized pseudo empirical likelihood inferences for complex surveys. *The Canadian Journal of Statistics, 43*, 1–17.

Tang, C. Y., & Leng, C. (2010). Penalized high dimensional empirical likelihood. *Biometrika, 97*, 905–920.

Thompson, M. E. (1984). Model and design correspondence in finite population sampling. *Journal of Statistical Planning and Inference, 10*, 323–334.

Thompson, M. E. (1997). *Theory of sample surveys*. London: Chapman & Hall.

Thompson, M. E. (2008). International surveys: Motives and methodologies. *Survey Methodology, 34*, 131–141.

Thompson, M. E., Fong, G. T., Hammond, D., Boudreau, C., Driezen, P., Hyland, A., et al. (2006). Methods of the International Tobacco Control (ITC) four country survey. *Tobacco Control, 15*(Suppl. 3), i12–18.

Thompson, M. E., Ramirez Ramirez, L. L., Lyubchich, V., & Gel, Y. R. (2016). Using the bootstrap for statistical inference on random graphs. *The Canadian Journal of Statistics* **44**, 3–24.

Thompson, M. E., & Wu, C. (2008). Simulation-based randomized systematic PPS sampling under substitution of units. *Survey Methodology, 34*, 3–10.

Thompson, S. K. (1990). Adaptive cluster sampling. *Journal of the American Statistical Association, 85*, 1050–1059.

Thompson, S. K. (1991). Stratified adaptive cluster sampling. *Biometrika, 78*, 389–397.

Thompson, S. K., & Seber, G. (1996). *Adaptive sampling*. Hoboken: Wiley.

Tibshirani, R. J. (1996). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society, Series B, 58*, 267–288.

Tillé, Y. (2006). *Sampling algorithms*. New York: Springer.

Tourangeau, R., Conrad, F. G., & Couper, M. P. (2013). *The science of web surveys* (1st ed.). Oxford: Oxford University Press.

Tourangeau, R., Rips, L. J., & Rasinksi, K. (Eds.). (2000). *The psychology of survey response*. Cambridge: Cambridge University Press.

Tsiatis, A. A. (2006). *Semiparametric theory and missing data*. New York: Springer.

Valliant, R., & Dever, J. A. (2011). Estimating propensity adjustments for volunteer web surveys. *Sociological Methods* & *Research, 40*, 105–137.

Valliant, R., Dever, J. A., & Kreuter, F. (2013). *Practical tools for designing and weighting survey samples*. New York: Springer.

Valliant, R., Dorfman, A. H., & Royall, R. M. (2000). *Finite population sampling and inference: A prediction approach*. New York: Wiley.

van der Vaart, A. W. (2000). *Asymptotic statistics* (Vol. 3). Cambridge: Cambridge University Press.

Veitch, V., & Roy, D. M. (2019). Sampling and estimation for (sparse) exchangeable graphs. *Annals of Statistics, 47*, 3274–3299.

Volz, E., & Heckathorn, D. D. (2008). Probability based estimation theory for Respondent Driven Sampling. *Journal of Official Statistics, 24*, 79–97.

Wang, J., & Opsomer, J. D. (2011). On asymptotic normality and variance estimation for nondifferentiable survey estimators. *Biometrika, 98*, 91–106.

Wang, N., & Robins, J. M. (1998). Large-sample theory for parametric multiple imputation procedures. *Biometrika, 85*, 935–948.

Wang, Z., & Thompson, M. E. (2012). A resampling approach to estimate variance components of multilevel models. *The Canadian Journal of Statistics, 40*, 150–171.

Wasserman, L. (2000). Asymptotic inference for mixture models using data-dependent priors. *Journal of the Royal Statistical Society, Series B, 62*, 159–180.

Wei, G. C., & Tanner, M. A. (1990). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Association, 85*, 699–704.

West, B. T. (2011). Paradata in survey research. *Survey Practice, 4*. https://doi.org/10.29115/SP-2011-0018. How to refer to problems.

Wilson, E. B. (1927). Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association, 22*, 209–212.

Woodruff, R. S. (1952). Confidence intervals for medians and other position measures. *Journal of the American Statistical Association, 47*, 635–646.

Wu, C. (2003). Optimal calibration estimators in survey sampling. *Biometrika, 90*, 937–951.

Wu, C. (2004a). Some algorithmic aspects of the empirical likelihood method in survey sampling. *Statistica Sinica, 14*, 1057–1067.

Wu, C. (2004b). Combining information from multiple surveys through empirical likelihood method. *The Canadian Journal of Statistics, 32*, 15–26.

Wu, C. (2005). Algorithms and R codes for the pseudo empirical likelihood method in survey sampling. *Survey Methodology, 31*, 239–243.

Wu, C., & Lu, W. W. (2016). Calibration weighting methods for complex surveys. *International Statistical Review, 84*, 79–98.

Wu, C., & Luan, Y. (2003). Optimal calibration estimators under two-phase sampling. *Journal of Official Statistics, 19*, 119–131.

Wu, C., & Rao, J. N. K. (2006). Pseudo empirical likelihood ratio confidence intervals for complex surveys. *The Canadian Journal of Statistics, 34*, 359–375.

Wu, C., & Rao, J. N. K. (2010). Bootstrap procedures for the pseudo empirical likelihood method in sample surveys. *Statistics and Probability Letters, 80*, 1472–1478.

Wu, C., & Sitter, R. R. (2001a). A model-calibration approach to using complete auxiliary information from survey data. *Journal of the American Statistical Association, 96*, 185–193.

Wu, C., & Sitter, R. R. (2001b). Variance estimation for the finite population distribution function with complete auxiliary information. *The Canadian Journal of Statistics, 29*, 289–307.

Wu, C., Thompson, M. E., Fong, G. T., Jiang, Y., Yang, Y., Feng, G., et al. (2015). Methods of the International Tobacco Control (ITC) China survey: waves 1, 2 and 3. *Tobacco Control, 24*(Suppl. 4), iv1–5.

Wu, C., Thompson, M. E., Fong, G. T., Li, Q., Jiang, Y., Yang, Y., et al. (2010). Methods of the International Tobacco Control (ITC) China survey. *Tobacco Control, 19*(Suppl. 2), i1–5.

Wu, C., & Yan, Y. (2012). Weighted empirical likelihood inference for two-sample problems. *Statistics and Its Interface, 5*, 345–354.

Wu, C., & Zhang, S. (2019). Comments on: Deville and Särndal's calibration: revisiting a 25 years old successful optimization problem. *Test, 28*, 1082–1086.

Wu, C. F. (1982). Estimation of variance of the ratio estimator. *Biometrika, 69*, 183–189.

Xie, X., & Meng, X.-L. (2017). Dissecting multiple imputation from a multi-phase inference perspective: What happens when god's, imputer's and analyst's models are uncongenial (with discussion)? *Statistica Sinica, 27*, 1485–1594.

Yang, Y., & He, X. (2012). Bayesian empirical likelihood for quantile regression. *The Annals of Statistics, 40*, 1102–1131.

Yates, F., & Grundy, P. M. (1953). Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society, Series B, 15*, 253–261.

Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with *g* prior distributions. In P. Goel & A. Zellner (Eds.), *Bayesian Inference and Decision Techniques* (pp. 233–243). New York: Elsevier.

Zellner, A., & Siow, A. (1980). Posterior odds ratios for selected regression hypotheses. In J. M. Bernardo, M. H. DeGroot, D. V. Lindley, & A. F. M. Smith (Eds.), *Bayesian Statistics: Proceedings of the First International Meeting Held in Valencia* (pp. 585–603). Valencia: University of Valencia Press.

Zhang, S., Han, P., & Wu, C. (2019a). A unified empirical likelihood approach to testing MCAR and subsequent estimation. *Scandinavian Journal of Statistics, 46*, 272–288.

Zhang, S., Han, P., & Wu, C. (2019b). Empirical likelihood inference for non-randomized pretest-posttest studies with missing data. *Electronic Journal of Statistics, 13*, 2012–2042.

Zhao, P., Ghosh, M., Rao, J. N. K., & Wu, C. (2020a). Bayesian empirical likelihood inference with complex survey data. *Journal of the Royal Statistical Society, Series B, 82*, 155–174.

Zhao, P., Haziza, D., & Wu, C. (2020b). Survey weighted estimating equation inference with nuisance functionals. *Journal of Econometrics, 216*, 516–536.

Zhao, P., Haziza, D., & Wu, C. (2020c). Sample empirical likelihood and the design-based oracle variable selection theory. *Statistica Sinica* (revised).

Zhao, P., & Wu, C. (2019). Some theoretical and practical aspects of empirical likelihood methods for complex surveys. *International Statistical Review, 87*, S239–256.

Zhao, Q., Small, D. S., & Bhattacharya, B. B. (2019). Sensitivity analysis for inverse probability weighting estimators via the percentile bootstrap. *Journal of the Royal Statistical Society, Series B, 81*, 735–761.

Zhong, B., & Rao, J. N. K. (2000). Empirical likelihood inference under stratified random sampling using auxiliary population information. *Biometrika, 87*, 929–938.

Zieschang, K. D. (1986). A generalized least squares weighting system for the consumer expenditure survey. *Proceedings of the Section on Survey Research Methods* (pp. 64–71). Alexandria, VA: American Statistical Association.

Zieschang, K. D. (1990). Sample weighting methods and estimation of totals in the consumer expenditure survey. *Journal of the American Statistical Association, 85*, 986–1001.

Author Index

A
Aitkin, M., 248
Alexander, C.H., 115
Anderson, D.W., 305

B
Bahadur, R.R., 103
Baker, R., 320
Bassett, G., 141
Basu, D., 69
Beaumont, J.-F., 198, 204
Béland, Y., 273
Bellhouse, D.R., 24
Benedetti, R., 282, 283
Berger, Y.G., 120, 179
Bickel, P.J., 25, 37, 226
Binder, D.A., 148, 150
Binson, D., 260
Booth, J.G., 228
Brackstone, G.J., 134
Brewer, K.R.W., 72, 73, 86, 94, 120
Brick, J.M., 207
Buckland, S.T., 282
Burmeister, L.F., 306, 307

C
Cantrell, J., 275
Carrillo, I.A., 155, 156, 218
Cassel, C.M., 68, 104, 107
Chang, T., 135
Chauvet, G., 228, 229
Chen, J., 25, 103, 123, 126, 130, 147, 162, 205, 257, 288, 289
Chen, M., 218
Chen, S., 322, 326, 327
Chen, S.Y., 170, 173
Chen, Y., 216
Chowdhury, S., 135
Citro, C.F., 320
Cochran, W.G., 24, 68, 90
Copeland, K.R., 115
Cox, B.G., 282
Cumberland, W.G., 97

D
David, H.A., 231
Deming, W.E., 115, 132, 134
Dever, J. A., 323
Deville, J.C., 115, 118, 120, 121, 124, 126, 134
Dillman, D., 273
Dillman, D.A., 272
Dippo, C.S., 237
Dufour, J., 115

E
Efron, B., 226, 227, 239, 241
Elliott, M.R., 322
Erdös, P., 24
Ericson, W.A., 246

F
Fan, J., 151, 188, 288
Fay, R.E., 217, 237
Folsom, R.E., 122
Francisco, C.A., 147

C. Wu, M. E. Thompson, *Sampling Theory and Practice*, ICSA Book Series in Statistics, https://doi.org/10.1007/978-3-030-44246-0

Frank, O., 299
Freedman, D.A., 25, 37, 226
Frieze, A., 298
Fu, Y., 173
Fuller, W.A., 24, 25, 68, 101, 104, 115, 125, 147, 209, 306, 307

G
Gastwirth, J.L., 231
Gelman, A., 151
Ghangurde, P.D., 253
Gijbels, I., 151, 288
Gile, K.J., 298, 300
Godambe, V.P., 60, 68, 69, 141–144, 147, 150, 161, 209, 246, 254
Goodman, R., 74
Govindarajulu, Z., 31
Goyder, J., 273
Gregoire, T.G., 282
Gross, S., 228, 242
Grundy, P.M., 62, 74
Gunderson, D.R., 285

H
Hájek, J., 24, 25, 63–65, 71, 74, 84, 179, 248
Han, P., 135
Handcock, M.S., 298
Hanif, M., 73
Hansen, M.H., 78, 94, 245, 300
Hartley, H.O., 74, 162, 246, 306, 307
Haziza, D., 64, 198, 202, 204, 216, 327
He, X., 253
Heckathorn, D.D., 300
Hoadley, B., 246
Holmberg, A., 228
Horvitz, D.G., 59–67, 199
Hu, F., 230
Huang, E.T., 115, 125
Huang, Y.C., 275
Hurwitz, W.N., 78, 300

I
Ireland, C.T., 133
Isaki, C.T., 68

K
Kalbfleisch, J.D., 230
Kalton, G., 207, 305
Karoñski, M., 298
Kim, J.K., 120, 122, 126, 135, 198, 202, 206, 209, 217, 218, 237, 238, 317, 322, 327
Kish, L., 74, 75, 207
Koenker, R., 141
Kott, P.S., 135, 196, 321
Kreuter, F., 273
Krewski, D., 25, 37, 241
Kullback, S., 121, 133

L
Lavallée, P., 295, 300–302, 314
Lazar, N.A., 250
Lee, S., 323
Leibler, R.A., 121
Lemaitre, G., 115
Leng, C., 188
Lesage, E., 198
Li, R., 188
Liang, K.-Y., 155, 156
Link, M.W., 272
Little, R.J.A., 194, 206, 217, 254, 321
Liu, T.P., 131
Lohr, S.L., 103, 306, 308
Lorenz, M.O., 231
Lu, W.W., 121, 122, 124, 126, 127, 135, 237
Luan, Y., 135
Luery, D., 115
Lumley, T., 135
Lundström, S., 135

M
Madow, W.G., 24
Mashreghi, Z., 241
McCarthy, P.J., 224, 241, 242
McCullagh, P., 211
McFadden, D., 142
Mecatti, F., 313, 317
Meeden, G., 246
Meng, X.-L., 217
Midzuno, H., 86
Molina, C.E.A., 144, 246
Molloy, M., 298, 300
Montanari, G.E., 135
Murthy, M.N., 24

N
Narain, R.D., 60
Narisetty, N.N., 253
Nelder, J., 211
Newey, W.K., 142

Neyman, J., 11, 40, 52, 53, 90, 95, 246
Norris, D.A., 272

O
Opsomer, J.D., 135, 147
Owen, A.B., 162, 163, 165, 189, 246

P
Park, M., 125, 135
Park, S., 120
Patak, Z., 150
Paton, D.G., 272
Pfeffermann, D., 151, 246
Potter, F., 135

Q
Qin, J., 162, 185
Quenouille, M., 241

R
Raghunathan, T.E., 218
Ranalli, M.G., 135
Rao, J.N.K., 19, 25, 37, 60, 68, 69, 73–75, 78, 97, 124, 126, 134, 155, 156, 161, 162, 164, 169–171, 176, 177, 226, 227, 237, 241, 245, 246, 249, 253, 255, 306, 308, 310, 311, 313, 317
Rao, P.S.R.S., 97
Reed, B., 298, 300
Rényi, A., 24, 25
Rivers, D., 195, 322
Roberts, G., 148
Robins, J.M., 207, 212, 213
Roman, A., 115
Rosenbaum, P.R., 195, 199
Roy, D.M., 298
Royall, R.M., 94, 97, 246
Rubin, D.B., 195, 199, 206, 217, 321

S
Saigo, H., 227, 229
Sampford, M.R., 75, 124
Särndal, C.E., 104, 115, 118, 120, 121, 124, 126, 135
Scharfstein, D.O., 201
Schnute, J.T., 285
Scott, A., 69
Seber, G., 295, 302
Seber, G.A.F., 282
Sen, A.R., 62
Sen, P.K., 25
Shao, J., 205, 206, 217, 218, 239, 241
She, X., 208, 209, 211
Shook-Sa, B.E., 272
Singh, A.C., 126
Siow, A., 252
Sitter, R.R., 103, 107, 128–131, 227, 229, 237, 239, 240, 325
Skinner, C.J., 144, 306, 308
Smith, T.M.F., 69
Snijders, T.A,B., 298, 299
Snijkers, G., 282
Snowden, C.B., 224, 242
Stephan, F.F., 115, 132–134
St-Pierre, M., 273
Sukhatme, P.V., 75

T
Tan, Z., 120, 124, 178, 179, 237
Tang, C.Y., 188
Tanner, M.A., 206
Thompson, D.J., 59–67, 199
Thompson, M.E., 19, 68, 74, 131, 139, 141, 143, 144, 146, 147, 150, 228, 229, 254, 261, 267, 270, 299
Thompson, S.K., 295, 302, 303
Tibshirani, R.J., 188
Tillé, Y., 85
Tourangeau, R., 273, 320
Tsiatis, A.A., 142, 209
Tu, D., 241

V
van der Vaart, A.W., 142
Valentine, H.T., 282
Valliant, R., 94, 260, 322, 323
Vardeman, S., 246
Veitch, V., 298
Volz, E., 300

W
Wang, J., 147
Wang, L., 135
Wang, N., 207, 213
Wang, Z., 228, 229
Wasserman, L., 253
Wei, G.C., 206
West, B.T., 276
Wilson, E.B., 159
Woodruff, R.S., 103

Wu, C., 4, 68, 74, 97, 103, 107, 120–122, 124, 127–131, 134, 135, 147, 164, 169–171, 173–179, 187, 189, 191, 208, 209, 211, 226, 227, 237, 238, 240, 241, 245, 249, 259, 267, 306, 310, 311, 317, 325

X

Xie, X., 217

Y

Yan, Y., 191
Yang, Y., 253
Yates, F., 62, 74

Z

Zeger, S.L., 155, 156
Zellner, A., 252
Zhang, S., 135, 218
Zhao, P., 181, 184, 187, 189, 230, 232, 331
Zhong, B., 171
Zieschang, K.D., 115

Subject Index

A
Absolute tolerable error, 26–28
Adaptive and network surveys, 295–303
Adaptive cluster sampling, 298, 302, 303
Agricultural surveys, 4, 23, 67, 282, 283
Analytic use of survey data, 90, 180
Anticipated variance, 68, 69, 107
Auxiliary information, 89–91, 95, 103, 107, 109, 115, 116, 127, 128, 132, 135, 163, 164, 167, 168, 171, 172, 177, 190, 198, 201–206, 234, 250, 275, 283, 291, 292, 297, 306, 308, 310–312, 318, 320, 321, 323, 325, 327
Auxiliary variables, 14, 28, 39, 53, 67, 95, 98, 108, 110, 111, 115–120, 124, 125, 128, 130, 134, 167, 168, 178, 179, 190, 191, 194, 204, 205, 233, 234, 286, 297, 313, 321, 322, 328–331

B
Bahadur representation, 103, 147
Balanced repeated replications (BRR), 223, 241
Balanced sampling, 109
Basu's elephant, 70, 71
Bayesian credible intervals, 248, 250, 251, 256
Bayesian empirical likelihood, 245–256
Bayesian inference with survey data, 245–247
Benchmark constraints, 117, 127, 128, 136
Bernoulli sampling, 84
Bootstrap, 135, 167, 223, 248, 308
Brewer's PPS sampling method, 86
Business surveys, 4, 67, 282, 283

C
Calibration equations, 117, 123, 127, 129, 162, 325
Calibration estimator, 115–137, 243, 292
Calibration weighting, 108, 115–137, 173, 177, 198, 200, 234, 336
Catch per unit effort (CPUE), 284–293
Census parameters, 105, 106, 128, 180
Central limit theorem, 24, 25
Chi-square distance, 117–121, 129, 131, 136, 137
Clustering, 7–9, 25, 33, 115, 233, 281, 298
Cluster sampling, 8, 9, 33–56, 64, 67, 73, 74, 80, 82, 85, 87, 88, 97, 98, 113, 226–229, 231, 262, 267, 298, 302, 303
Combined ratio estimator, 110
Complete auxiliary information, 89, 109, 116, 127, 128, 135, 325
Complete-case analysis, 196, 198, 210, 212, 213, 218, 219
Computer assisted telephone interviewing (CATI), 260
Confidence interval, 13, 24, 25, 37, 95, 102, 103, 135, 149, 150, 159, 162, 164, 169–171, 173, 176–179, 236, 237, 241, 244, 284, 292, 293, 306, 308, 310–312, 315–317, 339, 340
Constrained maximization, 162, 172–175, 191, 308, 311
Cooperation rate, 266, 270
Coverage error, 5, 14, 275, 305
Cross-sectional survey, 155, 158, 267, 269
Cross-sectional weights, 267–268

C. Wu, M. E. Thompson, *Sampling Theory and Practice*, ICSA Book Series in Statistics, https://doi.org/10.1007/978-3-030-44246-0

D

Design-based inference, ix, 11–13, 24, 89, 90, 94, 95, 116, 117, 127, 129, 134, 135, 143, 151, 153, 161, 165, 177, 245, 254, 287, 319
Design-based variable selection, 180, 187–189
Design-consistency, 95, 120, 135, 144, 146, 158, 181, 182, 192, 246
Design-unbiased, 19, 29, 52, 54, 59, 60, 62, 64, 66, 70, 76, 83, 88, 95, 96, 98, 99, 104, 106–109, 119, 129, 132, 152, 225, 226, 237, 238, 242, 313, 314
Deterministic imputation, 203
Dirichlet prior, 247, 248
Doubly robust estimation, 202, 330
Dual frame sampling, 308, 309

E

Eligibility criteria, 4, 5
Email frame, 17
Empirical likelihood, 134, 162, 218, 236, 245, 284, 305, 308, 333
Establishment surveys, 271, 281, 282
Estimating equation, 105, 119, 139–160, 180–183, 185–187, 202, 209, 211, 212, 214–216, 229, 230, 232, 233, 235, 251, 325–327
Estimating function, 140–149, 159, 210, 213, 220, 230, 232
Estimating function bootstrap, 230
Exponential tilting, 121–124, 133

F

Face-to-face interviews, 259–262, 273
First order inclusion probabilities, 57, 66, 67, 72–75, 82, 84, 86, 101, 106, 136, 163, 224, 291, 306, 313, 334
Fish abundance index, 284, 285, 287, 289, 291
Fish abundance surveys, 283–285, 291
Fractional imputation, 207–209, 211, 213, 214, 218
Frame population, 5, 10, 14, 308

G

Generalized difference estimator, 107, 108, 289, 290, 292, 325
Generalized estimating equations (GEE), 155–158, 160
Generalized pseudo empirical likelihood, 123–125, 177–181
Generalized ratio estimator, 64, 66, 100–102, 136
Generalized regression estimator, 103–107, 119, 130, 292
Gini coefficient, 231, 232, 243
Godambe class of linear estimators, 60, 85

H

Hájek estimator, 60, 85
Hájek variance estimator, 63
Horvitz–Thompson estimator, 59–61, 63–66, 71, 74, 75, 78, 84, 85, 88, 102, 116, 117, 120, 144, 145, 147, 163, 165, 177, 181, 185, 199, 223–225, 227, 230, 231, 239, 244, 289, 297
Household surveys, 23, 115, 259–270, 278, 302
Hundreds banks, 272

I

Imputation class, 204, 205
Imputation for missing values, 200–209, 233, 240, 243
Internal consistency, 116, 135
Intra-cluster correlation coefficient, 46, 54
Inverse sampling, 30, 31
Item nonresponse, 193, 194, 198–202, 206, 233, 235

J

Jackknife, 223, 241
Judgement or purposive sampling, 6, 70

K

Kullback–Leibler distance, 121, 123, 136

L

Lagrange multiplier, 40, 117, 123, 124, 166–168, 173, 174, 181, 184, 185, 188, 189, 251, 337–339
Linearization, 97–101, 110, 219, 223, 232, 241, 243
Linear regression analysis, 151–153, 340
Link tracing design, 295–298, 302
List frame, 17, 260
Logistic regression analysis, 153, 154, 213, 340, 341
Longitudinal surveys, 155–158, 160, 218, 267, 269
Longitudinal weights, 269
Lorenz curve, 231, 243

M
Mailed questionnaires, 260
Mass imputation, 322, 323, 328, 330, 348
Mean imputation, 204
Measurement error, 193, 273, 275
Midzuno's method, 86, 87
Missing at random (MAR), 194, 197, 199, 200, 202, 206, 209, 210, 212–214, 218, 220, 244, 321
Missing completely at random (MCAR), 194, 196–198, 204, 210, 218, 219, 269
Missing mechanism, 195, 197, 204, 209, 210, 213–215, 218–220, 240
Missing not at random (MNAR), 194, 220
Mixed mode surveys, 275
Model-assisted, 89–112, 151, 325, 344
Model-based prediction, 89–112, 287, 289
Model-calibration estimators, 128–131, 135, 137
Model-unbiased prediction estimator, 91–93, 95, 101, 104, 106, 108
Multiple frame sampling, 305, 317
Multiple imputation, 206–207, 211, 213, 214, 217, 218
Multiplicity, 306, 308, 312–317
Multiplier bootstrap, 229–233, 243
Multi-stage cluster sampling, 64, 67, 82, 87, 193, 226, 227, 229, 267

N
Natural resource inventory surveys, 89, 115, 281–293
Nearest neighbor imputation (NNI), 132, 205, 322
Neyman allocation, 39–42, 52, 53
Non-human population surveys, 281–283
Non-probability survey samples, 6, 319–331

O
Observational units, 9, 10, 281
Optimal allocation, 39, 41, 42, 52, 53
Optimal regression estimator, 109, 120
Optimal strategy, 67–70
Opt-in panel, 6, 274, 320
Outcome regression model, 195, 202, 220, 325, 330
Over-coverage, 251, 273, 275

P
Paradata, 270, 276
Parametric imputation, 205, 212
Penalized pseudo empirical likelihood, 188–189
Penalized sample empirical likelihood, 189
Poisson sampling, 64, 83–85, 87, 120, 136, 179, 229
Population distribution function, 65–66, 102, 130, 146, 169, 232
Population mean, 10, 19, 34, 64, 89, 119, 141, 162, 198, 227, 245, 273, 291, 295, 306, 325
Population parameters, 10–11, 13, 26, 34, 39, 90, 99, 105, 125, 127, 130, 131, 142–143, 146, 147, 149, 152, 156, 162, 167, 169, 171, 180, 188, 189, 195, 210, 224, 228, 229, 232, 245, 246, 251, 282, 284, 324
Population quantiles, 65, 66, 103, 146, 147, 149, 167, 169, 185, 231, 232, 241, 251, 255
Population regression coefficient, 99, 104, 105, 143, 152, 154, 175, 187, 292
Population surveys, 4, 13, 281–283, 328
Population variance, 10, 19, 28, 34, 35, 40, 53, 55, 56, 94, 96, 106, 131, 135, 137, 141, 165, 167, 175
Posterior predictive distribution, 206
Post-stratification, 38, 54
PPS sampling, 66–83, 86–89, 140, 150, 158, 161, 169, 186, 187, 224–226, 229, 230, 239, 242–244, 255, 256, 263, 267, 268, 333–343
PPS sampling with replacement, 78–83, 87, 158, 161, 226, 230
Probability sampling designs, 12, 13, 15, 19, 20, 57–59, 66, 69, 90, 91, 95, 98, 100, 106, 116, 127, 134, 144, 147, 153, 154, 160, 161, 179, 195, 196, 201, 224, 228, 230, 245–247, 284, 295, 297, 300, 308, 316, 317, 322, 325, 326
Propensity score, 195–200, 202, 204, 209, 210, 212–215, 217, 219, 220, 321–327, 330
Propensity score adjusted analysis, 199–200
Propensity score model, 195, 196, 198, 202, 210, 214, 215, 217, 321, 322, 325–327, 330
Proportional allocation, 38–40, 42, 52
Pseudo EL ratio statistic, 168, 183, 184, 190, 192
Pseudo empirical likelihood (PEL), 123–125, 134, 162–173, 176–185, 188–189, 191, 236, 245–250, 255, 291–293, 305, 308–312, 314–317, 333, 336
Pseudo-GEE, 155–158, 160

Pseudo population bootstrap, 227–229, 244
Public-use survey data, 129, 233–237, 240, 241

Q
Quality control, 14, 265, 266
Questionnaire development, 264
Quota sampling, 6, 319

R
Raking ratio estimation, 115, 116, 132–134
Random digit dialing (RDD), 260, 261, 271, 272, 274, 276–279
Random hot deck imputation, 204, 205
Random imputation, 203, 316
Range-restricted calibration, 121, 125–127, 135–137
Rao–Hartley–Cochran method, 64, 75–78, 87
Rao–Sampford method, 74–75
Ratio adjustment to survey weights, 197
Regression analysis, 139–160, 211–214, 219, 220, 251, 333–341
Regression imputation, 205, 217, 220, 239, 240, 244
Rejective sampling, 25, 73–75, 84, 85, 179
Relative tolerable error, 27–28
Replication methods, 223–244
Replication weights, 233–238, 240, 241, 243
Resampling methods, 223, 237, 241
Respondent driven sampling (RDS), 298, 300
Response rates, 260, 266, 270, 319
Restricted sampling, 6
Retention rate, 266, 270

S
Sample data, 7–9, 13, 35, 43, 45, 47, 48, 59, 76, 91, 92, 94, 95, 100, 104, 108–112, 116, 127, 144, 161, 169, 189, 196, 199, 204, 205, 210, 223–228, 234, 239, 241, 244–247, 249, 250, 252, 256, 287, 305, 316, 319, 321–323
Sampled population, 5, 10, 14, 59
Sample empirical likelihood (SEL), 162–164, 175, 181, 184–189, 191, 245, 251, 253
Sample matching, 322, 323
Sample of convenience, 6
Sample of volunteers, 6
Sample size allocation, 38–42, 53, 171, 285, 286
Sample size calculation, 25–28, 31
Sampling design, 8, 17, 33, 57, 89, 116, 140, 161, 195, 224, 245, 259, 271, 285, 295, 305, 321
Sampling frames, 7–10, 12–14, 17, 18, 33, 38, 47, 57, 193, 259, 260, 263, 271, 281–283, 305, 306
Sampling procedure, 18, 21, 22, 30, 43, 47, 50, 63, 67, 71–77, 81–85, 170, 193, 227, 242, 263, 295
Sampling scheme, 18, 57
Sampling units, 9, 10, 17, 82, 241, 282, 283, 285, 286
Scale-load approach, 161, 162, 246
Second order inclusion probabilities, 57–59, 62, 63, 72–75, 78, 85, 116, 164, 224, 237, 238, 334, 335
Separate ratio estimator, 110
Separate regression estimator, 111
Simple random sampling without replacement (SRSWOR), 17–22, 25–31, 33, 36, 37, 39, 43–48, 50, 51, 53–56, 58, 72, 81, 85–88, 93, 95–100, 104, 105, 108–112, 120, 132, 137, 159, 161, 165, 219, 225, 228, 242, 243, 250, 256, 263
Simple random sampling with replacement (SRSWR), 21, 22, 26, 29, 30, 81, 82, 85, 165, 167, 169, 224, 225, 227, 234, 239, 302
Simple ratio estimator, 95–98, 100
Simple regression estimator, 98–100
Simple systematic sampling (SSS), 22–24, 30, 54, 73
Single imputation, 203–205, 207, 216, 217
Single-stage cluster sampling, 42–46, 50, 54, 55, 80, 85, 97, 98
Snowball sampling, 299, 302, 303
Stratification, 7, 8, 25, 33, 37–39, 53, 54, 115, 197, 233, 281, 285
Stratified sampling, 8, 33–56, 89, 171–173, 175, 191, 225, 226, 229, 230, 239, 241, 261, 287, 291, 310, 338
Stratified simple random sampling, 25, 33–40, 52, 53, 110, 111, 262, 292
Study population, 5
Study variable, 10, 34, 39, 53, 65, 67, 70, 78, 116, 125, 140, 141, 143, 165, 178, 193, 194, 198, 200, 203, 205, 206, 208, 211, 218, 282, 306, 313, 321, 322, 328, 330
Superpopulation model, 68, 70, 90, 91, 94, 103, 107, 110, 111, 128, 140, 142–144, 147, 151, 152, 155, 195, 200, 209

Survey population, 3, 15, 18, 24, 30, 33, 42, 50, 54, 60, 65, 68, 89–92, 94, 95, 99, 101, 102, 115, 127, 139–144, 149, 163, 165, 228, 229, 236, 240, 260, 278, 295
Survey samples, x, 5–7, 12, 13, 18, 30, 104, 139, 270, 305–307, 312–314, 316, 319–331
Surveys of animal populations, 282
Survey weighted estimating equations, 119, 140, 144–150, 180, 182, 183, 185, 186, 229, 235
Survey weighted estimator, 106, 129, 137, 153, 209, 214, 220, 232, 243

T
Target population, 5, 9, 10, 260–262, 275, 300, 305, 317, 320, 328
Telephone frame, 17
Telephone interviews, 14, 259–261, 273
Two-phase sampling, 25, 111, 112
Two-stage cluster sampling, 9, 33, 43, 46–52, 56, 85, 88, 229

U
Under-coverage, 171, 193, 273
Unequal probability sampling, 20, 25, 57–88, 98, 100, 106, 134, 153, 154, 160, 170, 174, 224, 230, 247, 316
Unit nonresponse, 193, 196–198, 200, 233, 235, 267

V
Variance estimation, 13, 24, 51, 62, 74, 78, 81, 82, 93, 97, 100, 102, 103, 109, 112, 147, 151, 158, 164, 178, 180, 183, 185, 190, 201, 202, 207, 214–217, 220, 223, 224, 226, 227, 229, 231, 233, 236–238, 240, 241, 243, 287, 290, 308, 310, 317, 326, 327

W
Web surveys, 271–280
Weight calculation, 267–269
Weighting class, 197
With-replacement bootstrap, 223–227, 229, 233, 234, 236, 239, 242–244

Y
Yates-Grundy-Sen variance formula, 62, 69, 86, 87