INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps.

ProQuest Information and Learning
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
800-521-0600

UMI[®]

Estimating Function Variance Estimation in Complex Surveys and Combined Classification

by

Manuila Tausi, B.Sc., M.App.Sc.

A thesis submitted to
the Faculty of Graduate Studies and Research
in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy

School of Mathematics and Statistics

Carleton University

Ottawa, Ontario

March, 2002

© Copyright

2002, Manuila Tausi

**National Library
of Canada**

**Acquisitions and
Bibliographic Services**

**385 Wellington Street
Ottawa ON K1A 0N4
Canada**

**Bibliothèque nationale
du Canada**

**Acquisitions et
services bibliographiques**

**385, rue Wellington
Ottawa ON K1A 0N4
Canada**

Your file / Votre référence

Our file / Notre référence

The author has granted a non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of this thesis in microform, paper or electronic formats.

The author retains ownership of the copyright in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

L'auteur a accordé une licence non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de cette thèse sous la forme de microfiche/film, de reproduction sur papier ou sur format électronique.

L'auteur conserve la propriété du droit d'auteur qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

0-612-71950-2

Canada

The undersigned recommend to
the Faculty of Graduate Studies and Research
acceptance of the thesis

**“Estimating Function Variance Estimation in
Complex Surveys and Combined Classification”**

submitted by

Manuila Tausi, B.Sc., M.App.Sc.

in partial fulfillment of the requirements
for the degree of Doctor of Philosophy

Chair

School of Mathematics and Statistics

Thesis Supervisor

External Examiner

Carleton University

April, 2002

Abstract

There are three main areas addressed in this thesis. Part I discusses jackknife variance estimation of the calibration estimator of a total or a vector of finite population parameters. Part II compares coverage accuracy of confidence intervals for the mean using Edgeworth expansions. Part III looks at the idea of combining several data-based classifiers in a statistical classification problem.

In Part I of the thesis we focus our attention on calibration methods, such as poststratification and generalised regression, that use auxiliary information to improve the precision of the sample estimates or to benchmark to known population totals. The calibration methods make adjustments to the basic survey design weights and use these calibration adjusted weights in forming estimates of means and totals for the variables of interest that are collected in the survey.

For the full response case, we present asymptotic normality of the calibration estimator of a population mean and asymptotic consistency of the corresponding Taylor linearization variance estimator using the regularity conditions in Yung (1996) and Yung and Rao (2000). We apply the estimating function (EF) method of Hu and Kalbfleisch (2000) to the jackknife procedure and obtain EF jackknife variance estimators of the calibration estimators of a population total or a vector of finite population parameters.

We extend the EF jackknife method to cover Wald and Quasi-score tests where we are interested in testing hypotheses on a subset of the vector of finite population parameters. We then proceed to discuss EF variance estimation of the calibration

estimators of a population total or a vector of finite population parameters under unit nonresponse and item nonresponse.

We evaluate the loss in efficiency of the balanced repeated replication variance estimator of the population mean when two or more groups are employed to obtain the balanced replicates, thus extending the results of Krewski (1978).

In Part II of the thesis we use Edgeworth expansions to compare the coverage accuracies of confidence intervals for the studentized t -statistic and an alternative approximate pivotal statistic proposed by Godambe and Thompson (1999).

In Part III of the thesis we are interested in combining several data-based classifiers (or classification rules) in order to produce more effective classification rules in terms of smallest error rates. In particular, we extend kernel-based combined classification rules (Mojirsheibani, 2000) to accomodate varying smoothing parameters.

Acknowledgements

First and foremost, I would like to thank my supervisors, Professors J.N.K. Rao and M. Mojirsheibani, for their advice, encouragement and direction during the course of this work. I would also like to thank the Canadian Government, through the Commonwealth Scholarship and Fellowship Plan, for providing the financial support for this research programme. In addition, I would like to thank Professor J.N.K. Rao and the Government of Tuvalu for providing additional financial support to complete my programme.

I want to acknowledge the moral support of my Canadian family (Southminster United Church Congregation) while undertaking this programme.

Finally, special thanks go to my wife Naomi for her unwavering support throughout this research programme.

Contents

Acceptance Sheet	i
Abstract	ii
Acknowledgements	iv
Contents	v
List of Tables	xi
List of Figures	xii
1 Introduction	1
1.1 Sampling Design and Asymptotic Setup	3
1.1.1 Sampling Design	3
1.1.2 Calibration	5
1.1.3 Asymptotic Setup	7
1.1.4 Results of Chapter 2	7
1.2 General Parameters θ	10
1.2.1 Parameter Estimation	11
1.2.2 Results of Chapter 3	12
1.3 Nonresponse	13

1.3.1	Unit Nonresponse	14
1.3.2	Results of Chapter 4	15
1.3.3	Item Nonresponse	17
1.3.4	Mean Imputation	17
1.3.5	Results of Chapter 5	20
1.4	Efficiency of the Balanced Repeated Replication Variance Estimator .	21
1.4.1	Results of Chapter 6	22
1.5	Coverage Accuracy of Confidence Intervals	23
1.6	Combining Classifiers	24
1.6.1	Results of Chapter 9	26
I	Multistage Sampling	28
2	Variance Estimation of a Total	29
2.1	Introduction	29
2.2	Stratified Multistage Sampling Design	30
2.2.1	Basic Estimator of a Total	31
2.2.2	Calibration Estimator of a Total	32
2.2.3	Taylor Linearization Variance Estimator	34
2.2.4	Large Sample Results	36
2.3	Resampling Variance Estimators	46
2.3.1	Linearized Jackknife	46
2.3.2	Bootstrap	48

2.3.3	Balanced Repeated Replication	52
2.3.4	Modified Balanced Repeated Replication	54
2.4	Estimating Function Variance Estimators	55
2.4.1	Estimating Function Jackknife	55
2.4.2	Estimating Function Bootstrap	58
2.4.3	Estimating Function Balanced Repeated Replication	60
2.4.4	Estimating Function Modified Balanced Repeated Replication	61
3	Variance Estimation of General Parameters θ	63
3.1	Introduction	63
3.2	Parameter Estimation	64
3.3	Variance Estimators	67
3.3.1	Taylor linearization	67
3.3.2	Linearized jackknife	75
3.3.3	Estimating function jackknife	77
3.4	Wald and Quasi-Score Tests	81
3.4.1	Wald Tests	82
3.4.2	Quasi-Score Tests	83
3.4.3	Taylor linearization variance estimator	85
3.4.4	Linearized jackknife variance estimator	86
3.4.5	Estimating function jackknife variance estimator	87
4	Variance Estimation Under Unit Nonresponse	90
4.1	Introduction	90

4.2	Calibration Estimator of a Total	91
4.2.1	Linearized Jackknife	93
4.2.2	Estimating Function Jackknife	94
4.3	Calibration Estimator of θ	97
4.3.1	Linearized Jackknife	98
4.3.2	Estimating Function Jackknife	99
5	Variance Estimation Under Mean Imputation	102
5.1	Introduction	102
5.2	Calibration Estimator of a Total	103
5.2.1	One Imputation Class	104
5.2.2	Multiple Imputation Classes	111
5.3	Calibration Estimator of θ	117
5.3.1	Linearized Jackknife	118
5.3.2	Estimating Function Jackknife	119
6	Efficiency of Balanced Repeated Replication Variance Estimator in the Linear Case	124
6.1	Introduction	124
6.2	Variance Estimation in the Linear Case	125
6.3	BRR for Two Groups	126
6.4	BRR and Orthogonal Arrays	128
6.5	BRR for Two or More Groups	131

II	Coverage Accuracy	140
7	Coverage Accuracy of Confidence Intervals for the Mean	141
7.1	Introduction	141
7.2	Studentized t -statistic	142
7.3	Alternative Approximate Pivotal Statistic	148
7.4	Comparison of Coverage Probabilities	152
III	Combined Classification	154
8	A Review of Classification	155
8.1	Introduction	155
8.2	Statistical Classification Setup	156
8.3	Classification Rules	158
8.3.1	Linear Discriminant Rules	159
8.3.2	Nearest Neighbour Rules	160
8.3.3	Kernel Rules	161
8.3.4	Classification Trees	163
8.4	Error Estimation	166
8.5	Empirical Risk Minimization	167
8.6	Vapnik - Chervonenkis Distribution - Free Performance Bound	170
8.7	Properties of Shatter Coefficient and VC Dimension	174
8.8	Partitioning and Splitting the Data	175

9	Combining Classifiers	181
9.1	Introduction	181
9.2	Linear Combined Classifiers	182
9.3	Combined Classifiers Via Discretization	185
9.3.1	Example	186
9.3.2	Large-sample Results	188
9.4	Kernel-based Combined Classifier	192
9.4.1	Fixed Smoothing Parameter	192
9.4.2	Variable Smoothing Parameters	195
9.4.3	Large Sample Results	196
10	Conclusions and Future Research	211
10.1	Conclusions	211
10.2	Future Research	213
	References	215

List of Tables

7.1	Coverage Probabilities (%) of the confidence intervals for the mean:	
	nominal level = 0.95	153
9.1	Predicted Class Memberships	187

List of Figures

8.1	Binary tree	164
-----	-----------------------	-----

Chapter 1

Introduction

The thesis contains three different parts. A large portion of the thesis, consisting of Part I, deals with variance estimation in stratified multistage sampling. In Part II of the thesis Edgeworth expansions are used to obtain coverage probabilities of pivotal statistics. Part III of the thesis looks at the idea of combining several data-based classifiers in a statistical classification problem, where we are interested in predicting the unknown nature of an observation, based on the available data.

In Part I of the thesis, we consider stratified multistage sampling designs where calibration methods, such as poststratification and generalised regression, are commonly employed for two reasons: to improve the precision of the sample estimates or to benchmark to known population totals. The calibration methods make adjustments to the basic survey design weights and use these calibration adjusted weights in forming estimates of means and totals for the variables of interest that are collected in the survey.

Our particular interest in Part I of the thesis is in the estimation of the variance

of estimators of a population total or a vector of finite population parameters, using the Estimating Function (EF) method of Hu and Kalbfleisch (2000). We apply the EF method to three resampling techniques: jackknife, balance repeated replication (BRR) and bootstrap. An advantage of these resampling techniques is that they employ a single variance formula for all parameter estimates, unlike the Taylor linearization method where a separate variance formula is derived for each parameter estimate. However, the Taylor linearization method is applicable to general sampling designs. The jackknife and the BRR methods are applicable only to those stratified multistage designs in which clusters within strata are sampled with replacement or the first-stage sampling fraction is negligible. The bootstrap method is applicable to more general designs (Rao, Wu and Yue, 1992) but its properties have not been fully studied.

In Part I of the thesis, we will study the calibration estimator and the corresponding jackknife variance estimators under the full response case and the nonresponse case. The sampling design and asymptotic setup are described in Section 1.1. Results of Chapter 2 are summarized in Subsection 1.1.4. In Section 1.2 we consider the estimation of a vector of finite population parameters for the sample using the estimating equations approach, and assuming auxiliary information is available at the estimation stage. Results of Chapter 3 are given in Subsection 1.2.2. Section 1.3 considers the case of nonresponse; both unit nonresponse and item nonresponse. Subsection 1.3.2 and Subsection 1.3.5 give results for Chapter 4 and Chapter 5 respectively. In Section 1.4, we discuss the loss in efficiency of the BRR variance estimator, with the results of Chapter 6 being given in Subsection 1.4.1.

Part II of the thesis, consisting of Chapter 7, is summarized in Section 1.5. In Chapter 7 we compare the coverage accuracy of confidence intervals obtained using the studentized t -statistic and an alternative pivotal statistic proposed by Godambe and Thompson (1999).

Part III of the thesis, consisting of Chapters 8 and 9, is summarized in Section 1.6. Here we are interested in combining several individual classifiers; for example, Fisher's linear discriminant functions, nearest-neighbour rules, kernel rules, partitioning rules or tree classifiers, in order to produce more effective classification rules, in terms of smallest error rates. Chapter 8 gives a review of classification. The results of Chapter 9 are outlined in Subsection 1.6.1.

1.1 Sampling Design and Asymptotic Setup

1.1.1 Sampling Design

We now describe the stratified multistage sampling design which will be considered in Part I of the thesis. Suppose the population under consideration is stratified into L nonoverlapping design strata with N_h first-stage clusters or primary sampling units (PSU's) in the h^{th} stratum. From each stratum h , $n_h \geq 2$ first-stage clusters are selected independently across the strata. These first-stage clusters are usually selected without replacement to avoid the selection of the same cluster more than once. But at the stage of variance estimation, the first-stage clusters are treated as if they were selected with replacement. This approximation leads to overestimation of

the true variance, but the relative bias is small if the first-stage sampling fractions, n_h/N_h , are small.

Within the $(hi)^{\text{th}}$ sampled first-stage cluster, m_{hi} ultimate units (elements) are selected according to some probability sampling method, $i = 1, \dots, n_h$, $h = 1, \dots, L$. Note that we do not specify the number of stages and the sampling methods used after the first-stage sampling, but we will simply assume that subsampling within the sampled clusters is performed to ensure unbiased estimation of cluster totals, Y_{hi} , $i = 1, \dots, n_h$, $h = 1, \dots, L$.

Associated with the k^{th} sampled element in the i^{th} sampled cluster of the h^{th} stratum, denoted by element (hik) , is the variable or characteristic of interest, y_{hik} , and the basic design weight, w_{hik} . Using these basic weights, an unbiased estimator of the population total, Y , is of the form

$$\hat{Y} = \sum_s w_{hik} y_{hik}$$

where $\sum_s$ denotes summation over $(hik) \in s$. Under the assumption of with replacement sampling of the clusters, $\hat{Y}$ can be written as

$$\hat{Y} = \sum_h n_h^{-1} \sum_i y_{hi}$$

where the cluster totals, $y_{hi} = \sum_k (n_h w_{hik}) y_{hik}$, are independent and identically distributed random variables within each stratum, but independent across strata. It follows that an unbiased estimator of the variance of $\hat{Y}$ is simply given by

$$v(\hat{Y}) = \sum_h n_h^{-1} (n_h - 1)^{-1} \sum_i (y_{hi} - \bar{y}_h)^2 = v(y_{hi})$$

where $\bar{y}_h = n_h^{-1} \sum_i y_{hi}$. The operator notation $v(y_{hi})$ indicates that $v(\hat{Y})$ depends

only on the weighted cluster totals y_{hi} . Under the assumption of without replacement sampling of the clusters, $v(\hat{Y})$ will overestimate the true variance of $\hat{Y}$.

1.1.2 Calibration

If auxiliary information is available at the estimation stage, calibration methods are commonly used to improve the precision of sample estimates or to ensure consistency with known population totals $\mathbf{X}$. We assume that $\mathbf{x}_{hik}$ is some auxiliary data vector known for all population units, or for all sampled units and the population total $\sum_U \mathbf{x}_{hik} = \mathbf{X}$ is known. The basic design weights, w_{hik} , are then adjusted by the calibration procedure to obtain calibration weights, $\tilde{w}_{hik} = w_{hik} g_{hik}$, where g_{hik} is an adjustment factor, by minimizing a distance measure from w_{hik} subject to the restriction that $\sum_s \tilde{w}_{hik} \mathbf{x}_{hik} = \mathbf{X}$. The calibration estimator of the population total, Y , is then given by

$$\hat{Y}_G = \sum_s \tilde{w}_{hik} y_{hik}.$$

Two commonly used calibration procedures are the poststratified estimator and the Generalised Regression (GREG) estimator. In Part I of the thesis, we consider the GREG estimator where the calibration adjustment factor is

$$g_{hik} = 1 + (\mathbf{X} - \hat{\mathbf{X}})^T \hat{\mathbf{A}}^{-1} \mathbf{x}_{hik}$$

with $\hat{\mathbf{X}} = \sum_s w_{hik} \mathbf{x}_{hik}$ and $\hat{\mathbf{A}} = \sum_s w_{hik} \mathbf{x}_{hik} \mathbf{x}_{hik}^T$. Furthermore, we focus on the special case of $\mathbf{x}_{hik}$ being a vector of indicator auxiliary variables corresponding to the categories of the poststratifiers with known marginal population totals $\mathbf{X}$, where

the calibration adjustment factor reduces to

$$g_{hik} = \mathbf{X}^T \widehat{\mathbf{A}}^{-1} \mathbf{x}_{hik}.$$

A standard Taylor linearization variance estimator of $\hat{Y}_G$ is given by

$$v_{L1}(\hat{Y}_G) = \sum_h n_h^{-1} (n_h - 1)^{-1} \sum_i (e_{hi} - \bar{e}_h)^2 = v(e_{hi}), \quad (1.1.1)$$

where $e_{hi} = \sum_k (n_h w_{hik}) e_{hik}$, $e_{hik} = y_{hik} - \mathbf{x}_{hik}^T \hat{\mathbf{B}}$ denotes the residuals obtained by regressing y_{hik} on the auxiliary variables $\mathbf{x}_{hik}$ and $\bar{e}_h = n_h^{-1} \sum_i e_{hi}$. An alternative Taylor linearization variance estimator of $\hat{Y}_G$, which uses the calibration weights $\tilde{w}_{hik}$, is given by

$$v_{L2}(\hat{Y}_G) = \sum_h n_h^{-1} (n_h - 1)^{-1} \sum_i (\tilde{e}_{hi} - \bar{\tilde{e}}_h)^2 = v(\tilde{e}_{hi}), \quad (1.1.2)$$

where $\tilde{e}_{hi} = \sum_k (n_h \tilde{w}_{hik}) e_{hik}$ and $\bar{\tilde{e}}_h = n_h^{-1} \sum_i \tilde{e}_{hi}$. Valliant (1993) has shown that $v_{L2}(\hat{Y}_G)$ is design-consistent as well as model-consistent (under a model) for the special case of the poststratified estimator of a total, defined by

$$\begin{aligned} \hat{Y}_{PS} &= \sum_{c=1}^C {}_cM \frac{{}_c\hat{Y}}{{}_c\hat{M}} \\ &= \sum_{c=1}^C \sum_{(hik) \in {}_cs} {}_cw_{hik} y_{hik}, \end{aligned}$$

where ${}_cs$ denotes the sample in the c^{th} poststratum, ${}_cM$ are the known poststrata counts, ${}_c\hat{Y} = \sum_{(hik) \in {}_cs} w_{hik} y_{hik}$ and ${}_c\hat{M} = \sum_{(hik) \in {}_cs} w_{hik}$ are the estimates of the poststrata totals and counts respectively, and ${}_cw_{hik} = w_{hik}({}_cM / {}_c\hat{M})$ are the poststratified adjusted weights ($c = 1, \dots, C$).

1.1.3 Asymptotic Setup

We follow the asymptotic setup in Krewski and Rao (1981), who considered the behaviour of estimators as the number of strata, L , grows large. This setup includes the important special case where $n_h = 2$ clusters are selected within each stratum which allows maximum degree of stratification while still permitting valid variance estimation. We assume that there is a sequence of populations $\{p_\nu, \nu = 1, 2, \dots\}$, where population p_ν contains L_ν strata and finitely many elements, and the population under consideration is a member of this sequence of populations. For simplicity, the subscript ν is suppressed and all limiting processes are understood to be as $\nu \rightarrow \infty$. We also assume that n_h is bounded and that $L \rightarrow \infty$ as $\nu \rightarrow \infty$.

1.1.4 Results of Chapter 2

Asymptotic consistency and normality of the calibration (GREG) estimator of the population mean $\bar{Y} = (Y/M)$, where M is the total number of elements in the population, are given in Chapter 2. The calibration estimator of $\bar{Y}$ is given by

$$\bar{y}_G = \sum_s \tilde{w}_{hik}^\mu y_{hik},$$

where $\tilde{w}_{hik}^\mu = w_{hik}^\mu g_{hik}$ and $w_{hik}^\mu = (w_{hik}/M)$ are the normalized weights. Consistency of the Taylor linearization variance estimators, $v_{L1}(\bar{y}_G)$ and $v_{L2}(\bar{y}_G)$, is also presented in Chapter 2. Note that $v_{L1}(\bar{y}_G)$ and $v_{L2}(\bar{y}_G)$ are defined in equations (1.1.1) and (1.1.2) respectively, with the basic design weights, w_{hik} , replaced by the normalized weights, w_{hik}^μ .

Resampling Variance Estimators

Resampling techniques (also known as replication methods), as described by Rust and Rao (1996):

use computational intensity to overcome difficulties and inconveniences in utilizing an analytic solution to the problem at hand. Briefly, the replication approach consists of estimating the variance of a population parameter of interest by using a large number of somewhat different subsamples (or somewhat different sampling weights) to calculate the parameter of interest. The variability among the resulting estimates is used to estimate the true sampling error of the initial, or full-sample, estimate.

Three resampling techniques commonly used in complex survey sampling are known as the jackknife, balance repeated replication (BRR) and the bootstrap. An advantage of these resampling techniques is that they employ a single variance formula for all parameter estimates, unlike the Taylor linearization method where a separate variance formula is derived for each parameter estimate. To implement these resampling techniques, we need to store on the record of the sample element (hik) the following weights:

- the basic design weight,
- the replicate weight,
- the calibration weight,
- the replicate calibration weight or

- the estimating function replicate calibration weight.

Note that the estimating function replicate calibration weight, proposed in this thesis, is computationally simpler to obtain and it may retain confidentiality in the data file. These points will be discussed in more detail in Chapter 2. The remainder of Chapter 2 presents resampling variance estimators for these resampling techniques and compare them with an Estimating Function (EF) resampling approach, proposed by Hu and Kalbfleisch (2000) for the bootstrap. We will only illustrate the jackknife method for the estimation of a total Y here as it is applied in the rest of Part I of the thesis.

To apply the jackknife method in stratified multistage sampling, we delete one first-stage cluster at a time. The jackknife basic weights when the $(gj)^{\text{th}}$ cluster is deleted are given by $w_{hik(gj)} = w_{hik}b_{gj}$, where b_{gj} is 0 if $(hi) = (gj)$; is $n_g/(n_g - 1)$ if $h = g$ and $i \neq j$; is 1 if $h \neq g$. Then the jackknife calibration weights are given by

$$\tilde{w}_{hik(gj)} = w_{hik(gj)}g_{hik(gj)} \quad (1.1.3)$$

where the jackknife adjustment factor

$$g_{hik(gj)} = \mathbf{X}^T \widehat{\mathbf{A}}_{(gj)}^{-1} \mathbf{x}_{hik}$$

needs to be recalculated for every cluster (gj) that is deleted. We note that the jackknife calibration weights become zero whenever cluster (gj) is deleted. This may not retain confidentiality in the data file as analysts may be able to identify clusters by just looking at the column of jackknife calibration weights. It may also be possible that $\widehat{\mathbf{A}}_{(gj)}^{-1}$ may not exist for some (gj) .

Yung and Rao (1996) obtained a linearized jackknife variance estimator of $\hat{Y}_G$ as

$$v_{JL}(\hat{Y}_G) = v(\tilde{e}_{gj}),$$

which is identical to the Taylor linearization variance estimator $v_{L2}(\hat{Y}_G)$ given by (1.1.2).

Estimating Function Variance Estimators

The EF jackknife method, following the idea in Hu and Kalbfleisch (2000), involve EF jackknife calibration weights defined by

$$\begin{aligned} \check{w}_{hik(gj)} &= \tilde{w}_{hik} - w_{hik(gj)} g_{hik} \\ &+ w_{hik} \left(\sum_s w_{hik(gj)} g_{hik} \mathbf{x}_{hik}^T \right) \widehat{\mathbf{A}}^{-1} \mathbf{x}_{hik}. \end{aligned} \quad (1.1.4)$$

These EF jackknife calibration weights are not zero when cluster (gj) is deleted, unlike the jackknife calibration weights in (1.1.3). Furthermore, the EF jackknife calibration weights are computationally simpler to implement because we only have to invert the full matrix $\widehat{\mathbf{A}}$ thereby avoiding the repeated inversions of possibly illconditioned matrices $\widehat{\mathbf{A}}_{(gj)}$ that may be obtained from the ordinary jackknife method. The latter problem is more severe for the bootstrap method. An EF jackknife variance estimator of $\hat{Y}_G$ is shown to be identical to $v_{L2}(\hat{Y}_G)$ and $v_{JL}(\hat{Y}_G)$.

1.2 General Parameters θ

Finite population parameters of interest such as means, ratios, linear and logistic regression coefficients, can be expressed as solutions to suitable “census” estimating

equations (EE). Parameter estimates are obtained by solving sample EE which involve the basic design weights as well as calibration adjustment factors based on poststratification information. Binder (1983), Godambe and Thompson (1986), Kovacevic and Binder (1997) and others studied the EE approach for making inference from complex survey data, using the Taylor linearization method for variance estimation. Rao, Yung and Hidioglou (2002) discussed the implication of calibration adjustment on variance estimation, as well as presenting analogues of C.R. Rao's score tests that take account of the survey design features and the calibration procedure.

1.2.1 Parameter Estimation

A finite parameter vector θ_N can be regarded as the solution to "census" EE

$$\mathbf{S}(\theta) = \sum_U \mathbf{u}_{hik}(\theta) = 0$$

where $\sum_U$ denotes the summation over the finite population U and the estimating functions $\mathbf{u}_{hik}(\theta)$ are suitably chosen. For example, $\mathbf{u}_{hik}(\theta) = \mathbf{y}_{hik} - \theta$ gives the finite population mean $\theta_N = \bar{\mathbf{Y}}$.

In reality, we do not know the values for the whole finite population U but only for those elements in a sample drawn from the population, according to some well-defined sampling scheme. We assume that the sample design provides consistent, asymptotically normal estimators of population totals and associated standard errors. An estimator, $\hat{\theta}$, of the finite population parameter, θ_N , is then obtained as the solution to the sample EE

$$\hat{\mathbf{S}}(\theta) = \sum_s w_{hik} \mathbf{u}_{hik}(\theta) = 0,$$

where $\hat{\mathbf{S}}(\boldsymbol{\theta})$ is the basic estimator of $\mathbf{S}(\boldsymbol{\theta})$.

If we have auxiliary information available at the estimation stage, then we can use calibration methods to improve the precision of sample estimates or to benchmark to known population totals. A calibration estimator, $\hat{\boldsymbol{\theta}}_G$, of the finite population parameter, $\boldsymbol{\theta}_N$, is defined as the solution to the sample EE given by

$$\hat{\mathbf{S}}_G(\boldsymbol{\theta}) = \sum_s \tilde{w}_{hik} \mathbf{u}_{hik}(\boldsymbol{\theta}) = \mathbf{0}, \quad (1.2.1)$$

where $\hat{\mathbf{S}}_G(\boldsymbol{\theta})$ is the calibration estimator of $\mathbf{S}(\boldsymbol{\theta})$. The solution to (1.2.1) has a closed form in simple cases (e.g. mean, least squares regression). For logistic regression and other complex cases, it is necessary to solve (1.2.1) iteratively (e.g. Newton-Raphson algorithm) to obtain the solution $\hat{\boldsymbol{\theta}}_G$.

1.2.2 Results of Chapter 3

Asymptotic design-based consistency and normality of $\hat{\boldsymbol{\theta}}$ are given in Chapter 3. Variance estimators of the calibration estimator $\hat{\boldsymbol{\theta}}_G$ are also presented. One method uses the Taylor linearization approach while the other two are based on the jackknife procedure; namely the linearized jackknife and the EF jackknife. Confidence intervals on the components of $\boldsymbol{\theta}$, for example, the first component θ_1 , are also given in Chapter 3.

The remainder of Chapter 3 gives a discussion of design-model inferences on a model parameter vector $\boldsymbol{\theta}$ assuming a super-population model. In particular, we are interested in testing a null hypothesis of the form

$$H_0 : \boldsymbol{\theta}_2 = \boldsymbol{\theta}_{20}$$

using the sample data, where θ is partitioned as $\theta = (\theta_1^T, \theta_2^T)^T$. Wald tests and Quasi-score tests are presented, as well as jackknife variance estimators.

1.3 Nonresponse

Surveys typically collect responses to a large number of items for each sampled element. The results in Chapters 2 and 3 assumed that all sampled elements responded and fully completed the survey questionnaires. This situation is commonly referred to as the full response case. In practice though, the full response case is rarely achieved as most surveys suffer from nonresponse. Kalton and Kasprzyk (1986) described the two major types of nonresponse:

The problem of missing data occurs when some or all of the responses are not collected for a sampled element or when some responses are deleted because they fail to satisfy edit constraint. It is common practice to distinguish between total (unit) nonresponse, when none of the survey responses are available for a sampled element, and item nonresponse, when some but not all of the responses are available. Total nonresponse arises because of refusals, inability to participate, not-at-homes, and untraced elements. Item nonresponse arises because of item refusals, “don’t knows”, omissions and answers deleted in editing.

The distinction between unit and item nonresponse is useful since different methods are used to compensate for the missing data. In general, compensation for unit

nonresponse is handled through the use of weighting adjustments while imputation methods are used to compensate for item nonresponse.

In the presence of nonresponse, we divide the sample from a stratified multistage sampling design into respondents, s_r , and nonrespondents, s_m . In Part I of the thesis, we will assume that the sample can be divided into two or more weighting or imputation classes, according to the values of an auxiliary variable observed for all units in the sample. We then assume a uniform response mechanism within each weighting or imputation class.

1.3.1 Unit Nonresponse

Weighting adjustments are primarily used to compensate for unit nonresponse. In essence, these weighting adjustment procedures increase the weights of the respondents so that they represent the nonrespondents. In practice, the sample is divided into classes, called weighting classes, based on characteristics known for all sample elements, and weighting adjustments are then performed within each class.

We will illustrate the weighting adjustment method when we have multiple weighting classes. Let a_{hik} be the response indicator variable, where a_{hik} is 1 if element (hik) responds and 0 otherwise. We assume that we can form $M \geq 2$ weighting classes with uniform probability of response within each weighting class but different across the weighting classes. Then the nonresponse adjustment factor in the m^{th} weighting class is given by

$$d_m = \frac{\sum_s w_{hik} \eta_{hik}^m}{\sum_s w_{hik} a_{hik} \eta_{hik}^m} = \frac{\widehat{W}_m}{\widehat{R}_m}$$

where w_{hik} is the basic design weight for $(hik) \in s$ and η_{hik}^m is the weighting class indicator variable. Using this nonresponse adjustment, the basic estimator of the population total, Y , is

$$\hat{Y}_M = \sum_{m=1}^M \sum_s d_m w_{hik} a_{hik} \eta_{hik}^m y_{hik}$$

where the basic design weight is increased by the nonresponse adjustment factor to compensate for the nonresponding units. For the calibration estimator of the population total, Y , we use the calibration adjustment factor

$$g_{hik} = \mathbf{X}^T \bar{\mathbf{A}}_M^{-1} \mathbf{x}_{hik}$$

where $\mathbf{x}_{hik}$ is a P -dimensional vector of indicator auxiliary variables, $\mathbf{X}$ is the vector of known population totals and $\bar{\mathbf{A}}_M = \sum_{m=1}^M \sum_s d_m w_{hik} a_{hik} \eta_{hik}^m \mathbf{x}_{hik} \mathbf{x}_{hik}^T$. Then the calibration estimator is given by

$$\hat{Y}_{MG} = \mathbf{X}^T \tilde{\mathbf{B}}_M = \sum_{m=1}^M \sum_s d_m \tilde{w}_{hik} a_{hik} \eta_{hik}^m y_{hik}$$

where $\tilde{w}_{hik} = w_{hik} g_{hik}$, $\tilde{\mathbf{B}}_M = \bar{\mathbf{A}}_M^{-1} \tilde{\mathbf{U}}_M$ and $\tilde{\mathbf{U}}_M = \sum_{m=1}^M \sum_s d_m w_{hik} a_{hik} \eta_{hik}^m \mathbf{x}_{hik} y_{hik}$.

We assume here that the nonresponse adjustment is performed first followed by the calibration adjustment (Flyer, Rust and Morganstein, 1989). This is necessary because when the calibration adjustment is performed first, the resulting estimator will not benchmark to the known population totals $\mathbf{X}$.

1.3.2 Results of Chapter 4

In Chapter 4 we present jackknife variance estimators for the calibration estimator of a population total Y and the calibration estimator of a finite population vector

θ_N , under unit nonresponse with M weighting classes. In Section 4.2 the linearized jackknife variance estimator for the calibration estimator of a population total Y (Yung and Rao, 2000) is

$$v_{JL}(\hat{Y}_{MG}) = v(\tilde{r}_{gj})$$

defined in equation (1.1.2), where

$$\begin{aligned} \tilde{r}_{gj} = & \sum_{m=1}^M d_m \sum_k (n_g \tilde{w}_{gjk}) a_{gjk} \eta_{gjk}^m e_{gjk} \\ & + \sum_{m=1}^M \hat{\tau}_m \hat{R}_m^{-1} \sum_k (1 - d_m a_{gjk}) (n_g w_{gjk}) \eta_{gjk}^m, \end{aligned}$$

$\hat{\tau}_m = \sum_s \tilde{w}_{gjk} a_{gjk} \eta_{gjk}^m e_{gjk}$, $e_{gjk} = y_{gjk} - \mathbf{x}_{gjk}^T \bar{\mathbf{B}}_M$ and $\bar{r}_g = n_g^{-1} \sum_j \tilde{r}_{gj}$. We show that the EF jackknife variance estimator, $v_{EFJ}(\hat{Y}_{MG})$, is approximately equal to $v_{JL}(\hat{Y}_{MG})$.

Section 4.3 extends the results of estimating totals to situations where we want to estimate finite population parameters that are defined as solutions of “census” estimating equations (EE). In the presence of unit nonresponse and multiple weighting classes, we define the calibration estimator of the census EE, $\mathbf{S}(\theta) = \sum_U \mathbf{u}_{hik}(\theta)$, by

$$\hat{\mathbf{S}}_{MG}(\theta) = \sum_{m=1}^M \sum_s d_m \tilde{w}_{hik} a_{hik} \mathbf{u}_{hik}(\theta)$$

and denote the solution of $\hat{\mathbf{S}}_{MG}(\theta) = \mathbf{0}$ by $\hat{\theta}_{MG}$, the calibration estimator of θ_N . Section 4.3 presents the linearized jackknife variance estimator of $\hat{\theta}_{MG}$ (Hidioglou, Rao and Yung, 2000), $\hat{\mathbf{V}}_{JL}(\hat{\theta}_{MG})$, and we show that the EF jackknife variance estimator, $\hat{\mathbf{V}}_{EFJ}(\hat{\theta}_{MG})$, is approximately equal to $\hat{\mathbf{V}}_{JL}(\hat{\theta}_{MG})$.

1.3.3 Item Nonresponse

Item nonresponse is usually handled by some form of imputation, where values are inserted for the missing items. These imputed values are obtained from the respondents under a given imputation rule. Imputation procedures are used to create a complete data set, from which survey estimates can be computed. An advantage of imputation is that it permits the use of the same survey weight for all items, unlike weighting adjustment which is typically used for unit nonresponse. However, imputation in general, does not increase the statistical precision of survey estimates.

Imputation procedures include deterministic imputation; such as mean imputation, ratio imputation and regression imputation, and stochastic imputation; such as random imputation and hot deck imputation. For deterministic imputation, the imputed values are deterministic given the sample of respondents and any auxiliary information on the nonrespondents. Stochastic imputation substitutes a missing item by one of the respondents under a given imputation rule.

We will focus on mean imputation, where the missing items are substituted by the mean of the respondents, under the stratified multistage sampling design in Subsection 1.1.1 and we assume that only item nonresponse occurs, for simplicity.

1.3.4 Mean Imputation

For a sample, s , of size n , suppose r elements, s_r , respond and m elements, s_m , do not respond to an item y . Let y_{hik}^* , for all $(hik) \in s_m$ be the imputed values for the missing data, based on the set of respondents (donors) $\{y_{hik}, (hik) \in s_r\}$. Assuming

we have one imputation class, the imputed estimator of the population total, Y , is given by

$$\begin{aligned}\hat{Y}_I &= \sum_{s_r} w_{hik} y_{hik} + \sum_{s_m} w_{hik} y_{hik}^* \\ &= \sum_s w_{hik} a_{hik} y_{hik} + \sum_s w_{hik} (1 - a_{hik}) y_{hik}^*,\end{aligned}\quad (1.3.1)$$

where a_{hik} is the nonresponse indicator variable. Mean imputation uses the mean of the respondents as the imputed values for the missing data. In other words, $y_{hik}^* = \hat{T}/\hat{R}$ for all $(hik) \in s_m$, where $\hat{T} = \sum_s w_{hik} a_{hik} y_{hik}$ and $\hat{R} = \sum_s w_{hik} a_{hik}$. In this case the imputed estimator of Y in (1.3.1) reduces to

$$\hat{Y}_I = \frac{\hat{T}}{\hat{R}} \hat{W}$$

where $\hat{W} = \sum_s w_{hik}$. This ratio estimator has no imputation variance because y_{hik}^* is deterministic given s_r .

For the calibration estimator of total under imputation, we follow the treatment in Yung and Rao (2000) where the calibration adjustment is performed first then followed by the imputation procedure. This approach should lead to more efficient imputation procedures because the calibration estimator of a total is more efficient than the basic estimator of a total. In this case, mean imputation uses

$$\tilde{y}_{hik}^* = \frac{\tilde{T}}{\tilde{R}} \text{ for all } (hik) \in s_m$$

as the imputed values for the missing data, where $\tilde{T} = \sum_s \tilde{w}_{hik} a_{hik} y_{hik}$ and $\tilde{R} = \sum_s \tilde{w}_{hik} a_{hik}$. We note that the calibration weights, $\tilde{w}_{hik} = w_{hik} g_{hik}$ where $g_{hik} = \mathbf{X}^T \tilde{\mathbf{A}}^{-1} \mathbf{x}_{hik}$ and $\tilde{\mathbf{A}} = \sum_s w_{hik} a_{hik} \mathbf{x}_{hik} \mathbf{x}_{hik}^T$, are used in the imputation procedure. The

calibration estimator of total under mean imputation is then given by

$$\begin{aligned}\hat{Y}_{IG} &= \sum_s \tilde{w}_{hik} a_{hik} y_{hik} + \sum_s \tilde{w}_{hik} (1 - a_{hik}) \tilde{y}_{hik}^* \\ &= \frac{\tilde{T}}{\tilde{R}} \tilde{W}\end{aligned}$$

where $\tilde{W} = \sum_s \tilde{w}_{hik}$. A standard jackknife variance estimator of $\hat{Y}_{IG}$ is given by

$$v_J(\hat{Y}_{IG}) = \sum_{g=1}^L \frac{n_g - 1}{n_g} \sum_{j=1}^{n_g} (\hat{Y}_{IG(gj)} - \hat{Y}_{IG})^2, \quad (1.3.2)$$

where

$$\hat{Y}_{IG} = \sum_s \tilde{w}_{hik(gj)} a_{hik} y_{hik} + \sum_s \tilde{w}_{hik(gj)} (1 - a_{hik}) \tilde{y}_{hik}^*.$$

We note that the variance estimator in (1.3.2) treats the imputed values as true values. Rao and Shao (1992) have shown that (1.3.2) leads to serious underestimation if the response rate is small, and they proposed a jackknife variance estimator that correctly estimates the variance of an estimator under imputation. Their method is based on making an adjustment to the imputed value whenever a respondent is to be deleted. The resulting jackknife variance estimator is design-consistent under uniform response. Yung and Rao (2000) extended the Rao-Shao adjusted jackknife to calibration (GREG) estimators that use poststratification information.

We now illustrate the Rao-Shao adjusted jackknife method under mean imputation. Let $\tilde{y}_{hik(gj)}^* = \tilde{T}_{(gj)} / \tilde{R}_{(gj)}$ for all $(hik) \in s_m$ when the $(gj)^{\text{th}}$ cluster is deleted, where $\tilde{T}_{(gj)} = \sum_s \tilde{w}_{hik(gj)} a_{hik} y_{hik}$ and $\tilde{R}_{(gj)} = \sum_s \tilde{w}_{hik(gj)} a_{hik}$. Then the adjusted imputed values

$$\tilde{z}_{hik(gj)}^* = \tilde{y}_{hik}^* + \tilde{y}_{hik(gj)}^* - \tilde{y}_{hik}^* = \frac{\tilde{T}_{(gj)}}{\tilde{R}_{(gj)}}, \quad (1.3.3)$$

reflect the fact that the donor set is changed when a sample cluster is deleted. Note that identification flags are needed when applying this method in order to identify observed and imputed values in the data file. Using the adjusted imputed values in (1.3.3), we define the adjusted imputed calibration estimator, when the $(gj)^{\text{th}}$ cluster is deleted, as

$$\hat{Y}_{IG(gj)}^a = \sum_s \tilde{w}_{hik(gj)} a_{hik} y_{hik} + \sum_s \tilde{w}_{hik(gj)} (1 - a_{hik}) \tilde{z}_{hik}^*$$

and a jackknife variance estimator is given by

$$v_J(\hat{Y}_{IG}) = \sum_{g=1}^L \frac{n_g - 1}{n_g} \sum_{j=1}^{n_g} (\hat{Y}_{IG(gj)}^a - \hat{Y}_{IG})^2.$$

1.3.5 Results of Chapter 5

In Chapter 5 we present jackknife variance estimators for the calibration estimator of a population total Y and the calibration estimator of a finite population vector θ_N , under mean imputation. The linearized jackknife variance estimator of $\hat{Y}_{IG}$, under one imputation class or multiple imputation classes (Yung and Rao, 2000), are given in Section 5.2. We then show that the EF jackknife variance estimator of $\hat{Y}_{IG}$ is approximately equal to the linearized jackknife variance estimator.

Section 5.3 discusses jackknife variance estimators for parameter estimates obtained from suitable estimating equations under mean imputation, for the case of one imputation class only. The EF jackknife variance estimator is also shown to be approximately equal to the linearized jackknife variance estimator of Yung, Hidioglou and Rao (2001).

1.4 Efficiency of the Balanced Repeated Replication Variance Estimator

The balanced repeated replication (BRR) method (McCarthy, 1969) is often used for estimating the variance of nonlinear statistics from stratified samples with two primary sampling units per stratum. If there are two or more primary units in each stratum, a common practice is to divide the units in the stratum into two equal or nearly equal groups and apply the BRR method to these groups (Kish and Frankel, 1970). The resulting variance estimators have been shown to be quite inefficient (Krewski, 1978).

In Chapter 6, we consider the simple case of stratified random sampling where the finite population of N units is partitioned into L non-overlapping strata of sizes $N_1, \dots, N_L$ with $N = \sum_{h=1}^L N_h$. A simple random sample of size $n_h \geq 2$ is drawn with replacement from the h^{th} stratum, independently for each h . An unbiased estimator of the population mean $\bar{Y} = \sum_{h=1}^L W_h \bar{Y}_h$ is given by

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

where $\bar{y}_h = n_h^{-1} \sum_{i=1}^{n_h} y_{hi}$ and $W_h = N_h/N$. A Taylor linearization variance estimator of $\bar{y}_{st}$ is

$$v_L(\bar{y}_{st}) = \sum_{h=1}^L W_h^2 \frac{s_h^2}{n_h}$$

where $s_h^2 = (n_h - 1)^{-1} \sum_{i=1}^{n_h} (y_{hi} - \bar{y}_h)^2$. When $n_h = 2$ for all h , McCarthy (1969) obtained a BRR variance estimator of $\bar{y}_{st}$ as

$$v_{BRR}(\bar{y}_{st}) = R^{-1} \sum_{r=1}^R \left(\bar{y}_{st}^{(r)} - \bar{y}_{st} \right)^2$$

where $\bar{y}_{st}^{(r)}$, $r = 1, \dots, R$, is the mean of the r^{th} half-sample or replicate. Krewski (1978) showed that when the sample in stratum h is randomly divided into two groups (i.e. $n_h = p = 2m_h$ where $m_h = m = p/2$ for all h) under some regularity conditions, then the efficiency due to the grouping is given by

$$\frac{V(v_{BRR}(\bar{y}_{st}))}{V(v_L(\bar{y}_{st}))} = p - 1, \quad (1.4.1)$$

where $V(v_{BRR}(\bar{y}_{st}))$ and $V(v_L(\bar{y}_{st}))$ are the variances of $v_{BRR}(\bar{y}_{st})$ and $v_L(\bar{y}_{st})$ respectively. We note that (1.4.1) increases markedly as the sample size within each stratum increases.

1.4.1 Results of Chapter 6

In Chapter 6, we use the BRR method for general $n_h \geq 2$, based on mixed orthogonal arrays (Gupta and Nigam, 1987), to extend the result in (1.4.1) to include the general case where two or more groups are employed to obtain the balanced replicates. Thus, we show that

$$\frac{V(v_{BRR}(\bar{y}_{st}))}{V(v_L(\bar{y}_{st}))} = \frac{p - 1}{t - 1} \quad (1.4.2)$$

where $t = p/m$ is the number of random groups assuming p is a multiple of t . It is worth noting that the efficiency in (1.4.2) is reduced as the number of groups increase, but this should be weighed against the increased cost of computations.

1.5 Coverage Accuracy of Confidence Intervals

In Part II of the thesis our interest in Chapter 7 is in obtaining confidence intervals for a parameter. These confidence intervals are often obtained by inverting the distribution of some approximate pivotal quantity of the form,

$$\left\{ \frac{(\text{estimate} - \text{parameter})}{(\text{estimated variance})^{1/2}} \right\}.$$

For example, to obtain confidence intervals for the population mean $\bar{Y}$, one may use the studentized t -statistic

$$S_t = \frac{\bar{y} - \bar{Y}}{s_{\bar{y}}}$$

as a pivotal statistic, where $s_{\bar{y}}^2$ is the variance of the sample mean $\bar{y}$. Note that $s_{\bar{y}}^2$ is an estimate of the unknown population variance $V(\bar{y})$. Godambe and Thompson (1999) raised the basic question as to “which of the many estimates of the variance would provide a pivot (or approximate pivot) leading to a set of plausible confidence intervals?”. To answer this question, Godambe and Thompson proposed a more direct method of constructing an approximate pivotal statistic using estimating function theory.

In Chapter 7, we compare the studentized t -statistic with this approximate pivotal statistic of Godambe and Thompson (1999) by examining Edgeworth expansions of these pivotal statistics.

1.6 Combining Classifiers

In Part III of the thesis we consider statistical classification problems, where we are interested in predicting the unknown nature of an observation, based on the available data. Chapter 8 gives a review of classification where we discuss the classification problem, misclassification error rates, different classification rules, error estimation methods, random partitioning of the data and other issues related to classification. In Chapter 9, we discuss the idea of combining several classifiers (or classification rules) in order to produce more effective classification rules, in terms of smallest error rates.

In a statistical classification setup, we consider the random pair $(\mathbf{X}, Y)$, where $\mathbf{X} \in \mathbb{R}^d$ is called the feature or predictor vector of observable measurements, and Y is called the class membership and takes values in $\{1, \dots, K\}$, assuming that a class structure already exists and there are K classes. The problem of classification is to predict Y based on $\mathbf{X}$, as accurately as possible, where accuracy means lower misclassification error rate. More specifically, we have to find a classifier ψ , which is any map of the form $\psi : \mathbb{R}^d \rightarrow \{1, \dots, K\}$, that we can use to predict Y given the feature vector $\mathbf{X}$. The problem is then to choose ψ in such a way that its misclassification error rate, $\text{err}(\psi)$, defined by

$$\text{err}(\psi) = \mathbf{P}(\psi(\mathbf{X}) \neq Y)$$

is as small as possible. If the distribution of $(\mathbf{X}, Y)$ is completely known, we would try to find the best possible classifier ψ_B , called the Bayes' classifier, with lowest error

rate, which is given by

$$\text{err}(\psi_B) = \mathbf{P}(\psi_B(\mathbf{X}) \neq Y) = \inf_{\psi: \mathbb{R}^d \rightarrow \{1, \dots, K\}} \mathbf{P}(\psi(\mathbf{X}) \neq Y).$$

Unfortunately, in practice, the underlying distribution of the random pair $(\mathbf{X}, Y)$ is unknown. However, we do have access to some data set $T_N = \{(\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_N, Y_N)\}$ where $(\mathbf{X}_i, Y_i)$, $i = 1, \dots, N$ are independent and identically distributed (i.i.d.) $\mathbb{R}^d \times \{1, \dots, K\}$ -valued random vectors, which may have been observed in the past through experiments or some other sources. We also assume that the random vectors in the data set $\{(\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_N, Y_N)\}$ have the same distribution as the new observation $(\mathbf{X}, Y)$. Suppose we can split the data set into a training sample of size n and a testing set of size m , where $n + m = N$. In particular, we define $T_n = \{(\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_n, Y_n)\}$ and $T_m = \{(\mathbf{X}_{n+1}, Y_{n+1}), \dots, (\mathbf{X}_{n+m}, Y_{n+m})\}$ as the training sample and the testing set respectively, where $(\mathbf{X}_{n+j}, Y_{n+j})$, $j = 1, \dots, m$ are independent of T_n and $(\mathbf{X}, Y)$.

Given that the distribution of $(\mathbf{X}, Y)$ is unknown in many practical problems, we have to form a classification rule $\{\psi_n, n \geq 1\}$, based on the training sample of size n , in such a way that the conditional error rate

$$\text{err}_n(\psi_n) = \mathbf{P}(\psi_n(\mathbf{X}) \neq Y \mid T_n)$$

is, in some sense, as small as possible. A classification rule based on the training sample $\{\psi_n, n \geq 1\}$ is said to be (weakly) consistent if

$$\text{err}_n(\psi_n) \longrightarrow_p \text{err}(\psi_B), \text{ as } n \rightarrow \infty.$$

There are a number of popular classification rules discussed in the literature.

These include Fisher's linear discriminant functions (Linear Discriminant Analysis), histogram classifiers, nearest-neighbour (NN) rules and tree classifiers. These classifiers (or classification rules) have different properties and their performance can depend on many different factors. For instance the Linear Discriminant Analysis (LDA) performs well for normal populations. In general, in a given situation, it is not clear at all as to how one should choose a classifier.

1.6.1 Results of Chapter 9

In Chapter 9, we consider methods for combining a number of individual classifiers. Suppose we have L individual classifiers available for a given problem. These individual classifiers could be, for example, Fisher's linear discriminant functions, nearest-neighbour classifiers, kernel rules, partitioning rules or tree classifiers. The approach often used is to find the single best classifier among the L individual classifiers. Some recent research (Wolpert, 1992, Breiman, 1995, Leblanc and Tibshirani, 1996, Mojirsheibani, 1997, 1999, 2000, and others) have looked at the idea of combining these individual classifiers rather than selecting the best classifier. They claim that the combined classifiers yield improvements in the overall error rate relative to the best individual classifiers.

In Section 9.2 we present methods for combining classifiers based on linear combinations of the estimated class conditional probabilities. Leblanc and Tibshirani (1996) considered a leave-one-out cross-validation procedure whereas Mojirsheibani (1997) used a data-splitting approach for combining classifiers. Section 9.3 presents a

discretization- or partitioning-based procedure for combining several classifiers, proposed by Mojirsheibani (1999). Kernel-based combined classification procedures (Mojirsheibani, 2000) are discussed in Section 9.4. Our contributions, which extend the results of Mojirsheibani (2000), involve kernel-based combined classification procedures with variable smoothing parameters.

Conclusions and Future Research

In Chapter 10 a brief summary of the main results is provided and suggestions for further research are also indicated.

Part I

Multistage Sampling

Chapter 2

Variance Estimation of a Total

2.1 Introduction

In complex large scale surveys, calibration methods such as poststratification and generalised regression are commonly employed to improve the precision of sample estimates or to benchmark to known population totals. In this chapter, we consider variance estimation for the generalised regression estimator of a total under stratified multistage sampling.

Section 2.2 outlines the stratified multistage sampling design for the estimation of a total, including the calibration estimator where auxiliary data is used to improve the precision of the sample estimates or to ensure consistency with known population totals. Asymptotic results for the Taylor linearization variance estimators are also given. Section 2.3 gives a review of the resampling variance estimators of the calibration estimator of a total; namely the jackknife, bootstrap and balanced repeated replication variance estimators. In Section 2.4 we apply the estimating func-

tion resampling method of Hu and Kalbfleisch (2000) to obtain a resampling variance estimator of the calibration estimator.

2.2 Stratified Multistage Sampling Design

Let us consider the following multistage sampling design. Suppose the population under consideration has been stratified into L nonoverlapping design strata with N_h first-stage clusters or primary sampling units (PSU's) in the h^{th} stratum. Let $M_{hi}, i = 1, \dots, N_h, h = 1, \dots, L$ represent the number of ultimate units (elements) in the $(hi)^{\text{th}}$ first-stage cluster, with the total number of elements in stratum h , which is usually unknown, being $M_h = \sum_{i=1}^{N_h} M_{hi}$, and $M = \sum_{h=1}^L M_h$ defining the total number of elements in the population. We will focus on stratified multistage designs with a large number of strata, L , and relatively few clusters, $n_h \geq 2$, selected from each stratum h , independently across the strata with $m_{hi}, i = 1, \dots, n_h, h = 1, \dots, L$ elements sampled, according to some probability sampling methods, if the $(hi)^{\text{th}}$ cluster is selected. Note that we do not specify the number of stages and the sampling methods used after the first-stage sampling. We will simply assume that subsampling within the sampled clusters is performed to ensure unbiased estimation of cluster totals, $Y_{hi}, i = 1, \dots, n_h, h = 1, \dots, L$. We define the total number of elements in the sample by $m = \sum_{h=1}^L \sum_{i=1}^{n_h} m_{hi}$ and let $n = \sum_{h=1}^L n_h$ denote the total number of first-stage sample clusters. Associated with the k^{th} sampled element in the i^{th} sampled cluster of the h^{th} stratum, denoted by element (hik) , is the variable or characteristic of interest $y_{hik}, k = 1, \dots, M_{hi}, i = 1, \dots, M_h, h = 1, \dots, L$.

2.2.1 Basic Estimator of a Total

From the specification of the sampling design, basic design weights $w_{hik}(> 0)$ attached to the sample elements $(hik) \in s$, where s denotes the sample of elements, are obtained. Using these basic weights, an unbiased estimator of the population total, Y , is defined by the basic estimator

$$\hat{Y} = \sum_s w_{hik} y_{hik}, \quad (2.2.1)$$

where y_{hik} represent the value of characteristic of interest associated with $(hik) \in s$ and $\sum_s$ denotes summation over $(hik) \in s$.

It is common practice to sample the first-stage clusters without replacement to avoid the selection of the same cluster more than once. But at the stage of variance estimation, the first-stage clusters are treated as if they were sampled with replacement. Under this assumption of with replacement sampling of clusters, the basic estimator in (2.2.1) can be expressed as

$$\hat{Y} = \sum_h n_h^{-1} \sum_i \sum_k (n_h w_{hik}) y_{hik} = \sum_h n_h^{-1} \sum_i y_{hi}, \quad (2.2.2)$$

where the cluster totals, $y_{hi} = \sum_k (n_h w_{hik}) y_{hik}$, are independent and identically distributed random variables within each stratum, but independent across strata. An estimator of the variance of $\hat{Y}$ is simply given by

$$v(\hat{Y}) = \sum_h n_h^{-1} (n_h - 1)^{-1} \sum_i (y_{hi} - \bar{y}_h)^2 = v(y_{hi}), \quad (2.2.3)$$

where $\bar{y}_h = n_h^{-1} \sum_i y_{hi}$. The operator notation $v(y_{hi})$ indicates that $v(\hat{Y})$ depends only on the weighted cluster totals y_{hi} . Note that under the assumption of without

replacement sampling of the clusters, the variance estimator defined in (2.2.3) will, in general, overestimate the true variance of $\hat{Y}$ but the calculations are greatly simplified. The relative bias due to assuming with replacement sampling of clusters will be negligible if the first-stage sampling fractions, n_h/N_h , are small.

2.2.2 Calibration Estimator of a Total

Auxiliary data is commonly used at the estimation stage, through calibration methods, to improve the precision of sample estimates or to ensure consistency with known population totals $\mathbf{X}$ of auxiliary variables $\mathbf{x}_{hik}$. The basic assumption behind the use of auxiliary variables $\mathbf{x}_{hik}$ is that they carry information about the variable of interest y_{hik} . We assume that $\mathbf{x}_{hik}$ is some auxiliary data vector known for all population units, or for all sampled units and the population total $\sum_U \mathbf{x}_{hik} = \mathbf{X}$ is known. The basic design weights, w_{hik} , are then adjusted by the calibration procedure to obtain calibration weights, $\tilde{w}_{hik} = w_{hik} g_{hik}$, where g_{hik} is the adjustment factor (also known as the estimation weight or g -weight), by minimizing a distance measure from w_{hik} subject to the restriction that $\sum_s \tilde{w}_{hik} \mathbf{x}_{hik} = \mathbf{X}$. The calibration estimator of the population total, Y , is given by

$$\hat{Y}_G = \sum_s \tilde{w}_{hik} y_{hik}. \quad (2.2.4)$$

Two commonly used calibration procedures are the poststratified estimator and the Generalised Regression (GREG) estimator. In the case of a single poststratifier, we assume that the population is partitioned into C nonoverlapping poststrata, that can cut across design strata, with known population counts ${}_cM$, $c = 1, \dots, C$. The

adjustment factor here is defined by

$$g_{hik} = \frac{{}_cM}{{}_c\widehat{M}} \text{ if } (hik) \in {}_cs, \quad (2.2.5)$$

where ${}_c\widehat{M} = \sum_s w_{hik} {}_c\delta_{hik}$ is an estimator of ${}_cM$, ${}_cs$ is the sample of elements from poststratum c , and ${}_c\delta_{hik}$ is a poststratum indicator variable defined as

$${}_c\delta_{hik} = \begin{cases} 1 & \text{if } (hik)^{\text{th}} \text{ element is in poststratum } c \\ 0 & \text{otherwise.} \end{cases}$$

The poststratified estimator is then given by

$$\widehat{Y}_{PS} = \sum_c \sum_s {}_cw_{hik} y_{hik} {}_c\delta_{hik} = \sum_c \frac{{}_cM}{{}_c\widehat{M}} {}_c\widehat{Y} \quad (2.2.6)$$

where ${}_cw_{hik} = ({}_cM/{}_c\widehat{M})w_{hik}$ is the poststratified weight and ${}_c\widehat{Y} = \sum_s w_{hik} y_{hik} {}_c\delta_{hik}$ is an estimator of the poststratum total ${}_cY$. We note that if the poststratified weights, ${}_cw_{hik}$, are used to estimate a poststratum total ${}_dM$, then replacing y_{hik} in (2.2.6) by the poststratum indicator variable ${}_d\delta_{hik}$, where ${}_d\delta_{hik} = 1$ if (hik) belongs to poststratum d and 0 otherwise, gives the poststratified estimator ${}_d\widehat{M}_{PS}$, which benchmarks to the known poststratum total ${}_dM$.

If we have two or more poststratifiers, we can apply the GREG estimator of the population total, Y , given by

$$\widehat{Y}_G = \widehat{Y} + (\mathbf{X} - \widehat{\mathbf{X}})^T \widehat{\mathbf{B}} \quad (2.2.7)$$

where $\mathbf{x}_{hik}$ is a P -dimensional vector of auxiliary variables with known population totals $\mathbf{X}$, $\widehat{\mathbf{X}} = \sum_s w_{hik} \mathbf{x}_{hik}$, $\widehat{\mathbf{B}} = \widehat{\mathbf{A}}^{-1} \widehat{\mathbf{U}}$ is the vector of estimated regression coefficients, $\widehat{\mathbf{A}} = \sum_s w_{hik} \mathbf{x}_{hik} \mathbf{x}_{hik}^T$ and $\widehat{\mathbf{U}} = \sum_s w_{hik} \mathbf{x}_{hik} y_{hik}$. The adjustment factor in

(2.2.7) can be written as

$$g_{hik} = 1 + (\mathbf{X} - \widehat{\mathbf{X}})^T \widehat{\mathbf{A}}^{-1} \mathbf{x}_{hik}. \quad (2.2.8)$$

In the special case of $\mathbf{x}_{hik}$ being a vector of indicator auxiliary variables corresponding to the categories of the poststratifiers with known marginal population totals $\mathbf{X}$, the GREG estimator in (2.2.7) reduces to the projection estimator (Särndal, Swensson and Wretman, 1989) given as

$$\widehat{Y}_G = \mathbf{X}^T \widehat{\mathbf{B}}, \quad (2.2.9)$$

and the adjustment factor becomes $g_{hik} = \mathbf{X}^T \widehat{\mathbf{A}}^{-1} \mathbf{x}_{hik}$. For the special case of the poststratified estimator, we have $\widehat{\mathbf{X}} = (\text{ }_1\widehat{M}, \dots, \text{ }_C\widehat{M})^T$, $\mathbf{X} = (\text{ }_1M, \dots, \text{ }_CM)^T$, $\widehat{\mathbf{B}} = (\text{ }_1\widehat{Y}/\text{ }_1\widehat{M}, \dots, \text{ }_C\widehat{Y}/\text{ }_C\widehat{M})^T$ and the adjustment factor then reduces to (2.2.5). We note that if the population totals $\mathbf{X}$ are estimated using the calibration weights, then $\widehat{\mathbf{X}}_G$ will benchmark to the known population totals.

2.2.3 Taylor Linearization Variance Estimator

Turning to variance estimation, the standard Taylor linearization variance estimator of $\widehat{Y}_G$ is given by

$$v_{L1}(\widehat{Y}_G) = \sum_h n_h^{-1} (n_h - 1)^{-1} \sum_i (e_{hi} - \bar{e}_h)^2 = v(e_{hi}), \quad (2.2.10)$$

where $e_{hi} = \sum_k (n_h w_{hik}) e_{hik}$, $e_{hik} = y_{hik} - \mathbf{x}_{hik}^T \widehat{\mathbf{B}}$ denotes the residuals obtained by regressing y_{hik} on the auxiliary variables $\mathbf{x}_{hik}$ and $\bar{e}_h = n_h^{-1} \sum_i e_{hi}$. Binder(1996) proposed a new linearization method where the partial derivatives are evaluated at

the estimates $\hat{Y}$, $\hat{\mathbf{X}}$ and $\hat{\mathbf{B}}$, rather than the population values Y , $\mathbf{X}$ and $\mathbf{B}$, as in the standard linearization method. This method is presented in this section.

Applying total differentials to both sides of equation (2.2.7), we obtain

$$\begin{aligned} (\partial \hat{Y}_G) &= (\partial \hat{Y}) - (\partial \hat{\mathbf{X}})^T \hat{\mathbf{B}} + (\mathbf{X} - \hat{\mathbf{X}})^T (\partial \hat{\mathbf{B}}) \\ &= (\partial \hat{Y}) - (\partial \hat{\mathbf{X}})^T \hat{\mathbf{B}} + (\mathbf{X} - \hat{\mathbf{X}})^T \hat{\mathbf{A}}^{-1} \{(\partial \hat{\mathbf{U}}) - (\partial \hat{\mathbf{A}}) \hat{\mathbf{B}}\}, \end{aligned} \quad (2.2.11)$$

where $\mathbf{X}$ is assumed to be fixed and known. Replacing total differentials of estimated quantities by deviations from their respective expected values, yields

$$\begin{aligned} \hat{Y}_G - Y &\approx (\hat{Y} - Y) - (\hat{\mathbf{X}} - \mathbf{X})^T \hat{\mathbf{B}} + (\mathbf{X} - \hat{\mathbf{X}})^T \hat{\mathbf{A}}^{-1} \{(\hat{\mathbf{U}} - \mathbf{U}) - (\hat{\mathbf{A}} - \mathbf{A}) \hat{\mathbf{B}}\} \\ &= \hat{Y} - \hat{\mathbf{X}}^T \hat{\mathbf{B}} + (\mathbf{X} - \hat{\mathbf{X}})^T \hat{\mathbf{A}}^{-1} \{\hat{\mathbf{U}} - \hat{\mathbf{A}} \hat{\mathbf{B}}\} + \text{other terms not depending} \\ &\quad \text{explicitly on } w_{hik} \\ &\approx \sum_s w_{hik} (y_{hik} - \mathbf{x}_{hik}^T \hat{\mathbf{B}}) + \sum_s w_{hik} (\mathbf{X} - \hat{\mathbf{X}})^T \hat{\mathbf{A}}^{-1} \mathbf{x}_{hik} (y_{hik} - \mathbf{x}_{hik}^T \hat{\mathbf{B}}) \\ &= \sum_s w_{hik} \left(1 + (\mathbf{X} - \hat{\mathbf{X}})^T \hat{\mathbf{A}}^{-1} \mathbf{x}_{hik}\right) e_{hik} \\ &= \sum_s w_{hik} g_{hik} e_{hik} \\ &= \sum_s \tilde{w}_{hik} e_{hik}. \end{aligned} \quad (2.2.12)$$

Using the approximation in (2.2.12), an alternative Taylor linearization variance estimator of $\hat{Y}_G$, using the calibration weights $\tilde{w}_{hik}$, is given by

$$v_{L2}(\hat{Y}_G) = \sum_h n_h^{-1} (n_h - 1)^{-1} \sum_i (\tilde{e}_{hi} - \bar{\tilde{e}}_h)^2 = v(\tilde{e}_{hi}), \quad (2.2.13)$$

where $\tilde{e}_{hi} = \sum_k (n_h \tilde{w}_{hik}) e_{hik}$ and $\bar{\tilde{e}}_h = n_h^{-1} \sum_i \tilde{e}_{hi}$. We note that $v_{L2}(\hat{Y}_G)$ is identical to a linearized jackknife variance estimator (Yung and Rao, 1996) discussed in Section 2.3.1. Rao (1985) proposed $v_{L2}(\hat{Y}_G)$ in the case of a single poststratifier, using

a heuristic argument. Valliant (1993) has shown that $v_{L2}(\hat{Y}_G)$ is design-consistent as well as model-consistent (under a model) for the special case of the poststratified estimator of a total defined in (2.2.6).

2.2.4 Large Sample Results

Krewski and Rao (1981) established the asymptotic consistency and normality for smooth functions of the basic estimator, $\hat{Y}$, and the consistency of variance estimators obtained using the linearization, jackknife and balanced repeated replication (BRR) methods in stratified sampling. Their asymptotic setup assumed that there is a sequence of populations $\{p_\nu, \nu = 1, 2, \dots\}$, where population p_ν contains L_ν strata and finitely many ($M = \sum_{h=1}^L \sum_{i=1}^{N_h} M_{hi}$) elements, although M may be very large, and the population under consideration is a member of this sequence of populations. Note that the quantities $L, M, N_h, n_T, n, w_{hik}, y_{hik}, v_{L1}, v_{L2}$ and so on depend on the population index ν , but ν is omitted for simplicity of notation. They also assumed that n_h is bounded and that $L \rightarrow \infty$ as $\nu \rightarrow \infty$. All limiting processes are understood to be as $\nu \rightarrow \infty$.

Yung (1996) established the asymptotic consistency and normality of the post-stratified estimator, $\hat{Y}_{PS}$, and the asymptotic consistency of the linearization variance estimator of $\hat{Y}_{PS}$. We will now present the large sample results of the calibration (GREG) estimator defined in equation (2.2.7). For developing the asymptotic theory, it is more convenient to work with means rather than with totals. The calibration

(GREG) estimator of the population mean $\bar{Y} = (Y/M)$ is

$$\bar{y}_G = \bar{y} + (\bar{\mathbf{X}} - \bar{\mathbf{x}})^T \hat{\mathbf{B}} = \sum_s \tilde{w}_{hik}^\mu y_{hik} \quad (2.2.14)$$

where $\tilde{w}_{hik}^\mu = w_{hik}^\mu g_{hik}$, $w_{hik}^\mu = (w_{hik}/M)$ are the normalized weights, $\bar{y} = \sum_s w_{hik}^\mu y_{hik}$ is the basic estimator of $\bar{Y}$, $\bar{\mathbf{x}} = \sum_s w_{hik}^\mu \mathbf{x}_{hik}$, $\hat{\mathbf{B}}$ and g_{hik} are defined in Section 2.2.2 with the basic design weights, w_{hik} , replaced by the normalized weights, w_{hik}^μ . Note that $\hat{\mathbf{B}}$ is an estimator of the vector of unknown population regression coefficients, $\mathbf{B}$. Since $\bar{Y} = \bar{\mathbf{X}}^T \mathbf{B}$, equation (2.2.14) can be expressed as

$$\begin{aligned} \bar{y}_G - \bar{Y} &= \bar{y} - \bar{\mathbf{X}}^T \mathbf{B} + (\bar{\mathbf{X}} - \bar{\mathbf{x}})^T \hat{\mathbf{B}} \\ &= \bar{y} - \bar{\mathbf{x}}^T \mathbf{B} + (\bar{\mathbf{X}} - \bar{\mathbf{x}})^T (\hat{\mathbf{B}} - \mathbf{B}) \\ &= \sum_s w_{hik}^\mu (y_{hik} - \mathbf{x}_{hik}^T \mathbf{B}) + (\bar{\mathbf{X}} - \bar{\mathbf{x}})^T (\hat{\mathbf{B}} - \mathbf{B}) \\ &= \bar{\epsilon} + (\bar{\mathbf{X}} - \bar{\mathbf{x}})^T (\hat{\mathbf{B}} - \mathbf{B}) \end{aligned} \quad (2.2.15)$$

where $\bar{\epsilon} = \sum_s w_{hik}^\mu \epsilon_{hik}$ and $\epsilon_{hik} = y_{hik} - \mathbf{x}_{hik}^T \mathbf{B}$ denotes the true residuals.

We assume the following regularity conditions to hold:

- C1. $n \max_{h,i,k} m_{hi} w_{hik}^\mu = O(1)$, where $n \rightarrow \infty$ as $L \rightarrow \infty$ and m_{hi} is the number of elements sampled in the $(hi)^{\text{th}}$ sampled cluster.
- C2. $n^{1+\delta} \sum_h \sum_i E|\epsilon_{hi} - E(\epsilon_{hi})|^{2+\delta} = O(1)$ for some $\delta > 0$, where $\epsilon_{hi} = \sum_k w_{hik}^\mu \epsilon_{hik}$ and $\epsilon_{hik} = y_{hik} - \mathbf{x}_{hik}^T \mathbf{B}$.
- C3. $n V(\bar{\epsilon}) \rightarrow \sigma_\epsilon^2$ as $n \rightarrow \infty$, where $\bar{\epsilon} = \sum_s w_{hik}^\mu \epsilon_{hik}$.
- C4. $n V(\bar{\mathbf{Z}}) = O(1)$ where $\bar{\mathbf{Z}} = \sum_s w_{hik}^\mu \mathbf{x}_{hik} \epsilon_{hik}$.

Condition C1 assumes that no survey weight is disproportionately large. Condition C2 is a standard Liapounov condition on the $2 + \delta$ moment of the cluster totals, ϵ_{hi} . Condition C3 assumes that the variance of $\bar{\epsilon}$ converges to a finite value σ_ϵ^2 , while condition C4 is needed to show $\sqrt{n}(\hat{\mathbf{B}} - \mathbf{B}) = O_p(1)$. Note that

$$\begin{aligned}
 \hat{\mathbf{B}} - \mathbf{B} &= \hat{\mathbf{A}}^{-1} \hat{\mathbf{U}} - \mathbf{B} \\
 &= \hat{\mathbf{A}}^{-1} \left\{ \sum_s w_{hik}^\mu \mathbf{x}_{hik} y_{hik} - \sum_s w_{hik}^\mu \mathbf{x}_{hik} \mathbf{x}_{hik}^T \mathbf{B} \right\} \\
 &= \hat{\mathbf{A}}^{-1} \left(\sum_s w_{hik}^\mu \mathbf{x}_{hik} \epsilon_{hik} \right) \\
 &= \hat{\mathbf{A}}^{-1} \bar{\mathbf{Z}}.
 \end{aligned} \tag{2.2.16}$$

Since $E(\bar{\mathbf{Z}}) = \mathbf{0}$, condition C4 and Chebyshev's inequality imply that $\bar{\mathbf{Z}} = O_p(n^{-\frac{1}{2}})$. Noting that $\hat{\mathbf{A}} = O_p(1)$ we can see that $\hat{\mathbf{B}} - \mathbf{B} = O_p(n^{-\frac{1}{2}})$. Condition C2 can be replaced by a weaker condition

$$C2'. \quad n^\delta \sum_h \sum_i E|\epsilon_{hi} - E(\epsilon_{hi})|^{1+\delta} = O(1) \text{ for some } \delta > 0$$

because C2 implies C2' (Yung, 1996). Before investigating the large sample theory, we will state some standard results.

Lemma 2.1 (Law of Large Numbers) *Let $\{X_i, i = 1, 2, \dots\}$ be a sequence of independent random variables with finite expectations. If*

$$\frac{1}{n} \sum_{i=1}^n E|X_i|^{1+\delta} = O(1)$$

as $n \rightarrow \infty$ for some $\delta > 0$, then

$$\frac{1}{n} \sum_{i=1}^n (X_i - E(X_i)) \rightarrow_p 0$$

where $\rightarrow_p$ denotes convergence in probability.

Lemma 2.2 (Central Limit Theorem) *Let $\{X_1, \dots, X_n\}$ be independent random variables with $E(X_i) = \mu_i$ and $V(X_i) = \sigma_i^2$. If $n^{-1} \sum_{i=1}^n \sigma_i^2 \rightarrow \sigma^2 > 0$ and $n^{-1} \sum_{i=1}^n E|X_i - \mu_i|^{2+\delta} = O(1)$ as $n \rightarrow \infty$ for some $\delta > 0$, then*

$$\sqrt{n}(\bar{X} - \bar{\mu}) \rightarrow_d N(0, \sigma^2)$$

where $\bar{X} = n^{-1} \sum_{i=1}^n X_i$, $\bar{\mu} = n^{-1} \sum_{i=1}^n \mu_i$, and $\rightarrow_d$ denotes convergence in distribution.

Lemma 2.3 (Slutsky's Theorem) *Let $X, X_1, X_2, \dots, Y_1, Y_2, \dots$ be random variables. If $X_n \rightarrow_d X$ and $Y_n \rightarrow_p c$, where c is a fixed real number, then*

$$(i) \ X_n + Y_n \rightarrow_d X + c.$$

$$(ii) \ X_n Y_n \rightarrow_d cX,$$

$$(iii) \ X_n / Y_n \rightarrow_d X/c \text{ if } c \neq 0.$$

Definition 2.1 *A sequence $\{Y_n\}_n$ is stochastically bounded if for each $\epsilon > 0$ there exists a constant c such that*

$$P\{|Y_n| > c\} < \epsilon$$

for all n sufficiently large.

A useful sufficient condition for $\{Y_n\}_n$ to be stochastically bounded is the uniform boundedness of $\{E|Y_n|\}_n$ which follows from the Markov inequality.

Lemma 2.4 *Let $\{X_n\}$ and $\{Y_n\}$ be sequences of random variables. If $X_n \rightarrow_p 0$ and $\{Y_n\}_n$ is stochastically bounded, then $X_n Y_n \rightarrow_p 0$ as $n \rightarrow \infty$.*

We will now present the large sample results for the calibration estimator of the mean, $\bar{y}_G$, and its variance estimator.

Theorem 2.1 (Consistency of the calibration estimator of the mean) *Under conditions C1, C2' and C3, $\bar{y}_G$ is consistent for $\bar{Y}$. That is,*

$$\bar{y}_G \rightarrow_p \bar{Y}.$$

Proof: It suffices to show that both terms in (2.2.15) converge to 0 in probability. To apply Lemma 2.1 to the first term of (2.2.15), let $X_{hi} = n(\epsilon_{hi} - E(\epsilon_{hi}))$ and from condition C2'

$$\frac{1}{n} \sum_h \sum_i E|X_{hi}|^{1+\delta} = n^\delta \sum_h \sum_i E|\epsilon_{hi} - E(\epsilon_{hi})|^{1+\delta} = O(1).$$

We then have

$$\frac{1}{n} \sum_h \sum_i X_{hi} \rightarrow_p \frac{1}{n} \sum_h \sum_i E(X_{hi}) = 0,$$

but $n^{-1} \sum_h \sum_i X_{hi} = \bar{\epsilon} - E(\bar{\epsilon})$. Now

$$\begin{aligned} E(\bar{\epsilon}) &= E\left(\sum_s w_{hik}^\mu y_{hik} - \sum_s w_{hik}^\mu \mathbf{x}_{hik}^T \mathbf{B}\right) \\ &= \bar{Y} - \bar{\mathbf{X}}^T \mathbf{B} \\ &= 0 \end{aligned}$$

where the last equality holds under the condition that $\mathbf{X}$ contains a row-vector of ones and therefore $\bar{\epsilon} \rightarrow_p 0$. From condition C4 and expression (2.2.16) we know that $\hat{\mathbf{B}} - \mathbf{B} \rightarrow_p 0$. Also $\bar{\mathbf{X}} - \bar{\mathbf{x}} \rightarrow_p 0$ (see Lemma 2.5) and therefore, the second term of (2.2.15) converges to 0 in probability.

■

Theorem 2.2 (Asymptotic normality of the calibration estimator of mean)

Under conditions C1-C4

$$\sqrt{n}(\bar{y}_G - \bar{Y}) \rightarrow_d N(0, \sigma_\epsilon^2).$$

Proof: Using equation (2.2.15), we will need to show:

$$(i) \quad \sqrt{n}\bar{\epsilon} \rightarrow_d N(0, \sigma_\epsilon^2), \quad (2.2.17)$$

$$(ii) \quad \sqrt{n}(\bar{\mathbf{X}} - \bar{\mathbf{x}})^T (\hat{\mathbf{B}} - \mathbf{B}) \rightarrow_p 0. \quad (2.2.18)$$

Defining $X_{hi} = n(\epsilon_{hi} - E(\epsilon_{hi}))$, we have

$$\begin{aligned} \frac{1}{n} \sum_h \sum_i V(X_{hi}) &= nV(\bar{\epsilon}) \rightarrow \sigma_\epsilon^2 \quad \text{by C3} \\ \text{and } \frac{1}{n} \sum_h \sum_i E|X_{hi}|^{2+\delta} &= n^{1+\delta} \sum_h \sum_i E|\epsilon_{hi} - E(\epsilon_{hi})|^{2+\delta} = O(1) \quad \text{by C2.} \end{aligned}$$

The Central Limit Theorem gives the result in (2.2.17) since the conditions of Lemma 2.2 are satisfied. To show (2.2.18) we define

$$\sqrt{n}(\bar{\mathbf{X}} - \bar{\mathbf{x}})^T (\hat{\mathbf{B}} - \mathbf{B}) = \sum_t \sqrt{n}(\bar{X}_t - \bar{x}_t) (\hat{B}_t - B_t)$$

where $\hat{B}_t$, B_t , $\bar{x}_t$ and $\bar{X}_t$ represent the t^{th} elements of the vectors $\hat{\mathbf{B}}$, $\mathbf{B}$, $\bar{\mathbf{x}}$ and $\bar{\mathbf{X}}$ respectively. Let $\bar{x}_t = \sum_s w_{hik}^\mu \delta_{hik}$ be the estimator of $\bar{X}_t = {}_tM / M$ where ${}_tM$ is the population count in category t with indicator variable δ_{hik} . To complete the proof we will need the following result.

Lemma 2.5 *Under condition C1 $nV(\bar{x}_t) = O(1)$ and $\bar{x}_t \rightarrow_p \bar{X}_t$ for each t .*

Proof:

$$nV(\bar{x}_t) = nV\left(\sum_s w_{hik}^\mu \delta_{hik}\right)$$

$$\begin{aligned}
&= n \sum_h \sum_i V \left(\sum_k w_{hik}^\mu \epsilon_{hik} \right) \\
&\leq n \sum_h \sum_i E \left(\sum_k w_{hik}^\mu \epsilon_{hik} \right)^2 \\
&\leq n \sum_h \sum_i E \left(\sum_k (w_{hik}^\mu)^2 \right) \left(\sum_k \epsilon_{hik}^2 \right) \\
&\quad \text{by the Cauchy – Schwarz inequality} \\
&\leq n \sum_h \sum_i E \left(\sum_k w_{hik}^\mu w_{hik}^\mu \right) m_{hi} \\
&\leq n \max_{h,i,k} w_{hik}^\mu m_{hi} E \left(\sum_s w_{hik}^\mu \right) \\
&= O(1),
\end{aligned}$$

since $E(\sum_s w_{hik}^\mu) = 1$. By Chebyshev's inequality, we have

$$P \left\{ |\bar{x}_t - \bar{X}_t| \geq \epsilon \right\} \leq \frac{1}{\epsilon^2} V(\bar{x}_t) \rightarrow 0$$

and thus $\bar{x}_t \rightarrow_p \bar{X}_t$ for all t .

■

Since $\sqrt{n}(\hat{\mathbf{B}} - \mathbf{B}) = O_p(1)$ and $(\bar{\mathbf{X}} - \bar{\mathbf{x}}) \rightarrow_p 0$, we can appeal to Lemma 2.4 and confirm the result in (2.2.18). Applying Slutsky's Theorem (Lemma 2.3), we have

$$\begin{aligned}
\sqrt{n}(\bar{y}_G - \bar{Y}) &= \sqrt{n}\bar{\epsilon} + \sqrt{n}(\bar{\mathbf{X}} - \bar{\mathbf{x}})^T (\hat{\mathbf{B}} - \mathbf{B}) \\
&\rightarrow_d N(0, \sigma_\epsilon^2),
\end{aligned}$$

which completes the proof of Theorem 2.2.

■

Theorem 2.3 (Consistency of Taylor linearization variance estimator) *Under conditions C1-C4,*

$$n(v_{L1}(\bar{y}_G) - V(\bar{y}_G)) \rightarrow_p 0.$$

Proof: By adding and subtracting $\mathbf{x}_{hik}^T \mathbf{B}$ to the estimated residuals, e_{hik} , we have

$$\begin{aligned} e_{hik} &= y_{hik} - \mathbf{x}_{hik}^T \mathbf{B} - \mathbf{x}_{hik}^T (\hat{\mathbf{B}} - \mathbf{B}) \\ &= \epsilon_{hik} - \mathbf{x}_{hik}^T (\hat{\mathbf{B}} - \mathbf{B}). \end{aligned}$$

Then

$$e_{hi} = \sum_k w_{hik}^\mu e_{hik} = \epsilon_{hi} - \mathbf{x}_{hi}^T (\hat{\mathbf{B}} - \mathbf{B})$$

where $\epsilon_{hi} = \sum_k w_{hik}^\mu \epsilon_{hik}$ and $\mathbf{x}_{hi} = \sum_k w_{hik}^\mu \mathbf{x}_{hik}$. Therefore, we have

$$e_{hi} - \bar{e}_h = \epsilon_{hi} - \bar{\epsilon}_h - (\mathbf{x}_{hi} - \bar{\mathbf{x}}_h)^T (\hat{\mathbf{B}} - \mathbf{B}). \quad (2.2.19)$$

A Taylor linearization variance estimator for $\bar{y}_G$ is given by

$$\begin{aligned} v_{L1}(\bar{y}_G) &= \sum_h \frac{n_h}{n_h - 1} \sum_i (e_{hi} - \bar{e}_h)^2 \\ &= \sum_h \frac{n_h}{n_h - 1} \sum_i \left[\epsilon_{hi} - \bar{\epsilon}_h - (\mathbf{x}_{hi} - \bar{\mathbf{x}}_h)^T (\hat{\mathbf{B}} - \mathbf{B}) \right]^2 \\ &= \sum_h \frac{n_h}{n_h - 1} \sum_i (\epsilon_{hi} - \bar{\epsilon}_h)^2 + \\ &\quad \sum_h \frac{n_h}{n_h - 1} \sum_i \left[(\mathbf{x}_{hi} - \bar{\mathbf{x}}_h)^T (\hat{\mathbf{B}} - \mathbf{B}) \right]^2 - \\ &\quad 2 \sum_h \frac{n_h}{n_h - 1} \sum_i (\epsilon_{hi} - \bar{\epsilon}_h) \left[(\mathbf{x}_{hi} - \bar{\mathbf{x}}_h)^T (\hat{\mathbf{B}} - \mathbf{B}) \right] \\ &= v_L^\epsilon(\bar{y}_G) + v_2 + v_3, \end{aligned}$$

where

$$\begin{aligned} v_L^\epsilon(\bar{y}_G) &= \sum_h \frac{n_h}{n_h - 1} \sum_i (\epsilon_{hi} - \bar{\epsilon}_h)^2, \\ v_2 &= \sum_h \frac{n_h}{n_h - 1} \sum_i \left[(\mathbf{x}_{hi} - \bar{\mathbf{x}}_h)^T (\hat{\mathbf{B}} - \mathbf{B}) \right]^2 \text{ and} \\ v_3 &= -2 \sum_h \frac{n_h}{n_h - 1} \sum_i (\epsilon_{hi} - \bar{\epsilon}_h) \left[(\mathbf{x}_{hi} - \bar{\mathbf{x}}_h)^T (\hat{\mathbf{B}} - \mathbf{B}) \right]. \end{aligned}$$

We note that $v_L^\epsilon(\bar{y}_G)$ is not a true variance because it involves the unknown residuals

$\epsilon_{hik} = y_{hik} - \mathbf{x}_{hik}^T \mathbf{B}$. To show the consistency of $v_{L1}(\bar{y}_G)$ we will need to show

$$n(v_L^\epsilon(\bar{y}_G) - V(\bar{y}_G)) \rightarrow_p 0,$$

$$n v_2 \rightarrow_p 0 \text{ and } n v_3 \rightarrow_p 0.$$

We will now show the consistency of $v_L^\epsilon(\bar{y}_G)$. By adding and subtracting $E(\epsilon_{hi})$ we can rewrite $v_L^\epsilon(\bar{y}_G)$ as

$$\begin{aligned} v_L^\epsilon(\bar{y}_G) &= \sum_h \frac{n_h}{n_h - 1} \sum_i [(\epsilon_{hi} - E(\epsilon_{hi})) - (\bar{\epsilon}_h - E(\epsilon_{hi}))]^2 \\ &= \sum_h \frac{n_h}{n_h - 1} \sum_i (\epsilon_{hi} - E(\epsilon_{hi}))^2 - \sum_h \frac{n_h^2}{n_h - 1} (\bar{\epsilon}_h - E(\epsilon_{hi}))^2 \quad (2.2.20) \end{aligned}$$

We can use Theorem 2.3 of Yung (1996, p.29), with $y_{hi} = \sum_k w_{hik}^\mu y_{hik}$ replaced by

$\epsilon_{hi} = \sum_k w_{hik}^\mu \epsilon_{hik}$, to show that

$$\begin{aligned} n v_L^\epsilon(\bar{y}_G) &\rightarrow_p n \sum_h \frac{n_h^2}{n_h - 1} \sigma_{\epsilon h}^2 - n \sum_h \frac{n_h}{n_h - 1} \sigma_{\epsilon h}^2 \\ &= n \sum_h n_h \sigma_{\epsilon h}^2 = n V(\bar{\epsilon}), \end{aligned}$$

where $\sigma_{\epsilon h}^2 = V(\epsilon_{hi}) = E(\epsilon_{hi} - E(\epsilon_{hi}))^2$ and $V(\bar{\epsilon}) = \sum_h n_h \sigma_{\epsilon h}^2$. We note from equation

(4.3) of Särndal et. al. (1989) that $V(\bar{y}_G) \approx V(\bar{\epsilon})$ and therefore

$$n v_L^\epsilon(\bar{y}_G) \rightarrow_p n V(\bar{y}_G).$$

To show that $n v_2 \rightarrow_p 0$, let

$$\begin{aligned} n v_2 &= n \sum_h \frac{n_h}{n_h - 1} \sum_i [(\mathbf{x}_{hi} - \bar{\mathbf{x}}_h)^T (\hat{\mathbf{B}} - \mathbf{B})]^2 \\ &\leq n \sum_h \frac{n_h}{n_h - 1} \sum_i \|\mathbf{x}_{hi} - \bar{\mathbf{x}}_h\|^2 \|\hat{\mathbf{B}} - \mathbf{B}\|^2 \end{aligned}$$

by the Cauchy – Schwarz inequality

$$\begin{aligned}
&= \|\hat{\mathbf{B}} - \mathbf{B}\|^2 n \sum_h \frac{n_h}{n_h - 1} \sum_i \|\mathbf{x}_{hi} - \bar{\mathbf{x}}_h\|^2 \\
&= \sum_t |\hat{B}_t - B_t|^2 n \sum_h \frac{n_h}{n_h - 1} \sum_i |x_{thi} - \bar{x}_{th}|^2 \\
&= \sum_t |\hat{B}_t - B_t|^2 n v_L(\bar{x}_t). \tag{2.2.21}
\end{aligned}$$

We note that $n v_L(\bar{x}_t)$ is stochastically bounded for all t since

$$\mathbb{E}(|n v_L(\bar{x}_t)|) = n \mathbb{E}(v_L(\bar{x}_t)) = n V(\bar{x}_t) = O(1)$$

by Lemma 2.5, and $|\hat{B}_t - B_t| \rightarrow_p 0$. Applying Lemma 2.4 to equation (2.2.21) gives

$$n v_2 \rightarrow_p 0.$$

Finally, we can apply the Cauchy-Schwarz inequality to $n v_3$ to show that

$$n v_3 \rightarrow_p 0,$$

which completes the proof of Theorem 2.3. ■

We can show the consistency of $v_{L2}(\bar{y}_G)$ by noting that

$$g_{hik} = \bar{\mathbf{X}}^T \hat{\mathbf{A}}^{-1} \mathbf{x}_{hik} \rightarrow_p 1$$

(Särndal et. al., 1989) where $\hat{\mathbf{A}} = \sum_s w_{hik}^\mu \mathbf{x}_{hik} \mathbf{x}_{hik}^T$. This will mean that

$$\tilde{e}_{hi} = \sum_k w_{hik}^\mu g_{hik} e_{hik} \rightarrow_p \sum_k w_{hik}^\mu e_{hik} = e_{hi}$$

and therefore

$$v_{L2}(\bar{y}_G) \approx v_{L1}(\bar{y}_G).$$

So $v_{L2}(\bar{y}_G)$ is also consistent for $V(\bar{y}_G)$.

2.3 Resampling Variance Estimators

In this section we discuss three resampling techniques that are commonly used in survey sampling. These resampling techniques are known as the jackknife, balance repeated replication (BRR) and the bootstrap. An advantage of these resampling techniques is that they employ a single variance formula for all parameter estimates, unlike the Taylor linearization method where a separate variance formula is derived for each parameter estimate. However, the Taylor linearization method is applicable to general sampling designs. The jackknife and the BRR methods are applicable only to those stratified multistage designs in which clusters within strata are sampled with replacement or the first-stage sampling fraction is negligible. The bootstrap method is applicable to more general designs (Rao, Wu and Yue, 1992) but its properties have not been fully studied.

2.3.1 Linearized Jackknife

To apply the jackknife method in stratified multistage sampling, we delete one first-stage cluster at a time. We note that the total number of first-stage clusters, n , is large because we are dealing with stratified multistage designs with a large number of strata, L . Therefore, deleting clusters does not lose much efficiency but saves a large number of computations. We will first define the jackknife basic weights,

$w_{hik(gj)} = w_{hik} b_{gj}$, where

$$b_{gj} = \begin{cases} 0 & \text{if } hi = gj \\ \frac{n_g}{(n_g-1)} & \text{if } h = g; i \neq j, \\ 1 & \text{if } h \neq g \end{cases}$$

when the $(gj)^{\text{th}}$ first-stage cluster is deleted. The jackknife calibration weights are then defined by

$$\tilde{w}_{hik(gj)} = w_{hik(gj)} g_{hik(gj)}, \quad (2.3.1)$$

where the jackknife adjustment factor $g_{hik(gj)} = \mathbf{X}^T \widehat{\mathbf{A}}_{(gj)}^{-1} \mathbf{x}_{hik}$ needs to be recalculated for every cluster (gj) that is deleted. It is also possible that $\widehat{\mathbf{A}}_{(gj)}^{-1}$ may not exist for some (gj) . After deleting the $(gj)^{\text{th}}$ first-stage cluster, the jackknife calibration estimator of the population total, Y , is given by

$$\hat{Y}_{G(gj)} = \sum_s \tilde{w}_{hik(gj)} y_{hik}, \quad (2.3.2)$$

and a customary jackknife variance estimator of $\hat{Y}_G$ is defined by

$$v_J(\hat{Y}_G) = \sum_{g=1}^L \frac{n_g - 1}{n_g} \sum_{j=1}^{n_g} (\hat{Y}_{G(gj)} - \hat{Y}_G)^2. \quad (2.3.3)$$

An advantage of the jackknife variance estimator v_J is that it generalizes readily to more complex estimators. For a general population parameter, θ , estimated from the sample by $\hat{\theta}$, we express the jackknife variance estimator of the sampling variance, $V(\hat{\theta})$, as

$$v_J(\hat{\theta}) = \sum_{g=1}^L \frac{n_g - 1}{n_g} \sum_{j=1}^{n_g} (\hat{\theta}_{(gj)} - \hat{\theta})^2 \quad (2.3.4)$$

where $\hat{\theta}_{(gj)}$ is obtained in the same way as $\hat{\theta}$, but using the jackknife basic weights, $w_{hik(gj)}$, in place of w_{hik} .

By linearizing (2.3.3), Yung and Rao (1996) obtained a linearized jackknife variance estimator of $\hat{Y}_G$ given by

$$v_{JL}(\hat{Y}_G) = v(\tilde{e}_{gj}), \quad (2.3.5)$$

which is identical to the Taylor linearization variance estimator $v_{L2}(\hat{Y}_G)$.

2.3.2 Bootstrap

The bootstrap method (Efron, 1979) for the independent and identically distributed (i.i.d.) case has been extensively studied. In this section, we will describe extensions of the standard bootstrap to complex survey data, especially those obtained from stratified multistage sampling.

The Naive Bootstrap

An extension of the bootstrap method to survey problems is to apply the standard bootstrap in each stratum, since the y_{hi} 's are i.i.d. within each stratum, under the assumption of with replacement sampling of first-stage clusters. The naive bootstrap works as follows:

1. Take a simple random sample $\{y_{hi}^*, i = 1, \dots, n_h\}$ with replacement from the given sample $\{y_{hi}, i = 1, \dots, n_h\}$ in stratum h , independently across the strata. Let m_{hi}^* be the number of times the $(hi)^{\text{th}}$ sample cluster is selected and define the naive bootstrap weights associated with the bootstrap data y_{hik}^* by

$$w_{hik}^* = m_{hi}^* w_{hik}.$$

A bootstrap estimate of the population total Y is

$$\begin{aligned}\hat{Y}^* &= \sum_s w_{hik}^* y_{hik}^* \\ &= \sum_h n_h^{-1} \sum_i \sum_k (n_h w_{hik}^*) y_{hik}^* \\ &= \sum_h n_h^{-1} \sum_i y_{hi}^*\end{aligned}$$

where $y_{hi}^* = \sum_k (n_h w_{hik}^*) y_{hik}^*$.

2. Independently replicate step 1 a large number of times, B , and calculate the corresponding estimates $\hat{Y}^{*1}, \dots, \hat{Y}^{*B}$.

The bootstrap variance estimator of $\hat{Y}$ is given by

$$\begin{aligned}v_B(\hat{Y}) &= E_*(\hat{Y}^* - E_*(\hat{Y}^*))^2 \\ &= \text{var}_*(\hat{Y}^*) \\ &= \sum_h n_h^{-1} \sum_i \text{var}_*(y_{hi}^*) \\ &= \sum_h \frac{1}{n_h} \left(\frac{n_h - 1}{n_h} \right) s_h^2\end{aligned}\tag{2.3.6}$$

since $\text{var}_*(y_{hi}^*) = n_h^{-1} \sum_i (y_{hi} - \bar{y}_h)^2$ for any i and h , where $s_h^2 = (n_h - 1)^{-1} \sum_i (y_{hi} - \bar{y}_h)^2$, and E_* and var_* refer to the expectation and variance with respect to the bootstrap sampling. We note that the existence of the scale factor $\frac{n_h - 1}{n_h}$ in (2.3.6) results in a biased and inconsistent variance estimator for the important practical case of bounded n_h , because in many survey problems, a large number of n_h may be small. The bootstrap variance estimator in (2.3.6) is approximated by a Monte Carlo estimator

$$v_{B(a)}(\hat{Y}) = \frac{1}{B} \sum_{b=1}^B (\hat{Y}^{*b} - \hat{Y}^*)^2\tag{2.3.7}$$

where $\hat{Y}^* = B^{-1} \sum_{b=1}^B \hat{Y}^{*b}$. Note that $E_*(\hat{Y}^*)$ and $\hat{Y}^*$ in (2.3.6) and (2.3.7) respectively can be replaced by $\hat{Y}$.

As with the jackknife, the usefulness of the bootstrap method is for more complex sample estimators, $\hat{\theta}$, for which straightforward variance estimators do not exist. Thus for general estimators $\hat{\theta}$, the Monte Carlo estimator is simply given by

$$v_{B(a)}(\hat{\theta}) = \frac{1}{B} \sum_{b=1}^B (\hat{\theta}^{*b} - \hat{\theta}^*)^2 \quad (2.3.8)$$

where $\hat{\theta}^{*b}$ is obtained in the same way as $\hat{\theta}$, but using the weights w_{hik}^* for the b^{th} replicate in place of w_{hik} with $\hat{\theta}^* = B^{-1} \sum_{b=1}^B \hat{\theta}^{*b}$. In using the bootstrap method, one must determine the number of replicates, B , to be used. Indeed, the precision of the variance estimator increases with increasing B but the resources needed to carry out the bootstrap method obviously increase also.

The Rescaling Bootstrap

Recognizing the scaling problem in (2.3.6), Rao and Wu (1988) provided a bootstrap method for smooth statistics $\hat{\theta} = g(\hat{Y})$, herein called *the rescaling bootstrap*, whereby some scale adjustments are made on the standard bootstrap. These scale adjustments are made on the y_{hik} values directly, in order to have valid variance estimation in the practical case of small n_h . A more general bootstrap method that covers smooth as well as non-smooth statistics for arbitrary sizes n_h is given in Rao, Wu and Yue (1992). This method, which makes the adjustments on the basic design weights w_{hik} rather than the y_{hik} 's, works as follows:

1. Draw a simple random sample of m_h first-stage clusters with replacement from

the n_h first-stage sample clusters, independently across the strata. Let m_{hi}^* be the number of times the $(hi)^{\text{th}}$ sample cluster is selected ($\sum_i m_{hi}^* = m_{hi}$). Define the bootstrap weights by

$$w_{hik}^{(*)} = \left[\left\{ 1 - \left(\frac{m_h}{n_h - 1} \right)^{\frac{1}{2}} \right\} + \left\{ \left(\frac{m_h}{n_h - 1} \right)^{\frac{1}{2}} \left(\frac{n_h}{m_h} \right) m_{hi}^* \right\} \right] w_{hik}. \quad (2.3.9)$$

Calculate the rescaling bootstrap estimator of the population total Y by

$$\begin{aligned} \hat{Y}^{(*)} &= \sum_s w_{hik}^{(*)} y_{hik} \\ &= \sum_h m_h^{-1} \sum_{i=1}^{m_h} y_{hi}^{(*)} \end{aligned} \quad (2.3.10)$$

where $y_{hi}^{(*)} = \sum_k (n_h w_{hik}^{(*)}) y_{hik}$ and the bootstrap weights are defined in (2.3.9).

We note that, if the $(hi)^{\text{th}}$ cluster is not selected in the bootstrap sample, $m_{hi}^* = 0$ and the second term of (2.3.9) vanishes. If $m_h \leq n_h - 1$ then the bootstrap weights w_{hik}^* are all positive if $w_{hik} > 0$ for all $(hik) \in s$.

2. Independently replicate step 1 a large number of times, B , and calculate the corresponding estimates $\hat{Y}^{(*)1}, \dots, \hat{Y}^{(*)B}$.

The rescaling bootstrap variance estimator of $\hat{Y}$ and its Monte Carlo approximation are given by (2.3.6) and (2.3.7) respectively, with $\hat{Y}^*$ replaced by $\hat{Y}^{(*)}$. In the linear case, the rescaling bootstrap variance estimator of $\hat{Y}$ reduces to the correct variance estimator, defined by equation (2.2.3).

To incorporate auxiliary information in the rescaling bootstrap, we define the bootstrap calibration weights by

$$\tilde{w}_{hik}^{(*)} = w_{hik}^{(*)} g_{hik}^{(*)}$$

where $g_{hik}^{(*)} = \mathbf{X}^T \widehat{\mathbf{A}}^{(*)^{-1}} \mathbf{x}_{hik}$ denote the bootstrap adjustment factor and $\widehat{\mathbf{A}}^{(*)} = \sum_s w_{hik}^{(*)} \mathbf{x}_{hik} \mathbf{x}_{hik}^T$. We note that the bootstrap adjustment factor, $g_{hik}^{(*)}$, needs to be recalculated for every bootstrap sample. A bootstrap calibration estimator of the population total, Y , is given by

$$\hat{Y}_G^{(*)} = \sum_s w_{hik}^{(*)} y_{hik}, \quad (2.3.11)$$

and the rescaling bootstrap variance estimator of $\hat{Y}_G$ is defined by

$$v_B(\hat{Y}_G) = E_*(\hat{Y}_G^{(*)} - E_*(\hat{Y}_G^{(*)}))^2. \quad (2.3.12)$$

We note that the inverse of $\widehat{\mathbf{A}}^{(*)}$ may not exist in the Monte Carlo version of (2.3.12).

2.3.3 Balanced Repeated Replication

The balanced repeated replication (BRR) or balanced half-sample method was first proposed by McCarthy (1969) for the important special case where $n_h = 2$ first-stage clusters per stratum are sampled with replacement. A set of R half-samples or replicates is formed by deleting one first-stage cluster from the sample in each stratum. Let

$$\delta_{hi}^{(r)} = \begin{cases} 2 & \text{if the } (hi)^{\text{th}} \text{ cluster is in the } r^{\text{th}} \text{ half-sample} \\ 0 & \text{otherwise} \end{cases}$$

for $i = 1, 2$ and $r = 1, \dots, R$. This set may be defined by an $R \times L$ matrix $\Gamma = (\gamma_h^{(r)})$ where

$$\begin{aligned} \gamma_h^{(r)} &= \delta_{h1}^{(r)} - 1 \\ &= \begin{cases} +1 & \text{if the first cluster is in the } r^{\text{th}} \text{ half-sample} \\ -1 & \text{if the second cluster is in the } r^{\text{th}} \text{ half-sample} \end{cases} \end{aligned}$$

such that $\sum_{r=1}^R \gamma_h^{(r)} = 0$ for all h . A set of R half-samples is said to be balanced if

$$\sum_{r=1}^R \gamma_h^{(r)} \gamma_{h'}^{(r)} = 0 \quad \text{for all } h \neq h',$$

which means that the columns of the matrix are orthogonal. A minimal set of R balanced half-samples can be constructed from an $R \times R$ Hadamard matrix by choosing any L columns excluding the column of all +1's, where $L + 1 \leq R \leq L + 4$. Let $\hat{Y}_G^{(r)}$ be the calibration estimator of Y based on the r^{th} half-sample. Define the BRR basic weights by $w_{hik}^{(r)} = w_{hik} \delta_{hi}^{(r)}$ and BRR adjustment factors by $g_{hik}^{(r)} = \mathbf{X}^T \hat{\mathbf{A}}^{(r)-1} \mathbf{x}_{hik}$ using the projection estimator in (2.2.9), where $\hat{\mathbf{A}}^{(r)} = \sum_s w_{hik}^{(r)} \mathbf{x}_{hik} \mathbf{x}_{hik}^T$. We then have

$$\hat{Y}_G^{(r)} = \sum_s w_{hik}^{(r)} g_{hik}^{(r)} y_{hik} = \sum_s \tilde{w}_{hik}^{(r)} y_{hik} \quad (2.3.13)$$

where $\tilde{w}_{hik}^{(r)} = w_{hik}^{(r)} g_{hik}^{(r)}$ denotes the BRR calibration weights. Yung (1996) showed that the BRR variance estimator of $\hat{Y}_G$ is approximately equal to the Taylor linearization variance estimator in (2.2.13) for the special case of $n_h = 2$ for all h . That is,

$$\begin{aligned} v_{BRR}(\hat{Y}_G) &= \frac{1}{R} \sum_{r=1}^R (\hat{Y}_G^{(r)} - \hat{Y}_G)^2 \\ &\approx \frac{1}{R} \sum_{r=1}^R \left(\sum_h \gamma_h^{(r)} \Delta_{\tilde{e}_h} \right)^2 \\ &= \frac{1}{4} \sum_h (\tilde{e}_{h1} - \tilde{e}_{h2})^2 \\ &= v_{L2}(\hat{Y}_G) \end{aligned} \quad (2.3.14)$$

where $\Delta_{\tilde{e}_h} = \frac{1}{2}(\tilde{e}_{h1} - \tilde{e}_{h2})$ and $\tilde{e}_{hi} = \sum_k (n_h \tilde{w}_{hik}) e_{hik}$, $i = 1, 2$.

For more general estimators $\hat{\theta}$, a BRR variance estimator is given by

$$v_{BRR}(\hat{\theta}) = \frac{1}{R} \sum_{r=1}^R (\hat{\theta}^{(r)} - \hat{\theta})^2 \quad (2.3.15)$$

where $\hat{\theta}^{(r)}$ is the estimate obtained from the r^{th} half-sample or replicate.

2.3.4 Modified Balanced Repeated Replication

The standard BRR method defined in Section 2.3.3 deletes a half-sample for each replicate. Deleting half the sample is the same as increasing the weights of one half-sample by 100% and decreasing the weights of the other half-sample to zero. Noting this sharp perturbation of the weights, Robert Fay of the U.S. Census Bureau proposed a modified method obtained by perturbing the weights by $\pm 100\epsilon\%$ for the half-sample and its complement, where $0 < \epsilon < 1$ (see Dippo, Fay and Morganstein, 1984). We note here that, unlike the standard BRR method described in Section 2.3.3, this method, denoted by $\text{BRR}(\epsilon)$, does not delete any sampled units.

For the special case of $n_h = 2$ for all h , $m_h = n_h/2$ and a fixed ϵ , let

$$\begin{aligned} \delta_{hi}^{(r\epsilon)} &= \begin{cases} 1 + \epsilon \left(\frac{n_h - m_h}{m_h} \right)^{1/2} & \text{if the } (hi)^{\text{th}} \text{ cluster is selected in the } r^{\text{th}} \text{ replicate} \\ 1 - \epsilon \left(\frac{m_h}{n_h - m_h} \right)^{1/2} & \text{otherwise} \end{cases} \\ &= \begin{cases} 1 + \epsilon & \text{if the } (hi)^{\text{th}} \text{ cluster is selected in the } r^{\text{th}} \text{ replicate} \\ 1 - \epsilon & \text{otherwise} \end{cases} \end{aligned}$$

and define

$$\begin{aligned} \gamma_h^{(r\epsilon)} &= \delta_{h1}^{(r\epsilon)} - 1 \\ &= \begin{cases} \epsilon & \text{if the first cluster is selected in the } r^{\text{th}} \text{ replicate} \\ -\epsilon & \text{if the second cluster is selected in the } r^{\text{th}} \text{ replicate} \end{cases} \end{aligned}$$

where $\delta_{h1}^{(r\epsilon)} + \delta_{h2}^{(r\epsilon)} = 2$. Note that if $\epsilon = 1$ we have the standard BRR method. Define the $\text{BRR}(\epsilon)$ basic weights by $w_{hik}^{(r\epsilon)} = w_{hik} \delta_{hi}^{(r\epsilon)}$ and $\text{BRR}(\epsilon)$ adjustment factors by $g_{hik}^{(r\epsilon)} = \mathbf{X}^T \widehat{\mathbf{A}}^{(r\epsilon)-1} \mathbf{x}_{hik}$ using the projection estimator, where $\widehat{\mathbf{A}}^{(r\epsilon)} = \sum_s w_{hik}^{(r\epsilon)} \mathbf{x}_{hik} \mathbf{x}_{hik}^T$.

Then a $BRR(\epsilon)$ calibration estimator of Y , based on the r^{th} replicate, is given by

$$\hat{Y}_G^{(r\epsilon)} = \sum_s w_{hik}^{(r\epsilon)} g_{hik}^{(r\epsilon)} y_{hik} = \sum_s \tilde{w}_{hik}^{(r\epsilon)} y_{hik} \quad (2.3.16)$$

where $\tilde{w}_{hik}^{(r\epsilon)} = w_{hik}^{(r\epsilon)} g_{hik}^{(r\epsilon)}$ represent the $BRR(\epsilon)$ calibration weights. The $BRR(\epsilon)$ variance estimator of $\hat{Y}_G$ is given in Rao and Shao (1999) as

$$\begin{aligned} v_{BRR(\epsilon)}(\hat{Y}_G) &= \frac{1}{\epsilon^2 R} \sum_{r=1}^R \left(\hat{Y}_G^{(r\epsilon)} - \hat{Y}_G \right)^2 \\ &\approx \frac{1}{\epsilon^2 R} \sum_{r=1}^R \left(\sum_h \gamma_h^{(r\epsilon)} \Delta_{\tilde{e}_h} \right)^2 \\ &= v_{L2}(\hat{Y}_G). \end{aligned} \quad (2.3.17)$$

2.4 Estimating Function Variance Estimators

In this section we present Estimating Function (EF) resampling techniques, based on the method proposed by Hu and Kalbfleisch (2000) for the bootstrap, which can be applied to problems where the data are independent but not necessarily identically distributed. The EF bootstrap method seems to have substantial advantages, both in computation and accuracy, over the more traditional bootstrap method. We will apply the EF method to the resampling procedures discussed in Section 2.3.

2.4.1 Estimating Function Jackknife

For the projection estimator in (2.2.9), $\hat{\mathbf{B}}$ is obtained as the solution of the sample estimating function

$$\hat{\mathbf{S}}(\mathbf{B}) = \sum_s w_{hik} \mathbf{x}_{hik} \left(y_{hik} - \mathbf{x}_{hik}^T \mathbf{B} \right) = 0. \quad (2.4.1)$$

That is,

$$\sum_s w_{hik} \mathbf{x}_{hik} e_{hik} = \mathbf{0}$$

where $e_{hik} = y_{hik} - \mathbf{x}_{hik}^T \hat{\mathbf{B}}$. We now apply the EF method of Hu and Kalbfleisch (2000) to the jackknife procedure. First we replace the basic design weights and $\mathbf{B}$ in (2.4.1) by the jackknife basic weights and $\hat{\mathbf{B}}$ respectively, which yields

$$\hat{\mathbf{S}}_{(gj)}(\hat{\mathbf{B}}) = \sum_s w_{hik(gj)} \mathbf{x}_{hik} (y_{hik} - \mathbf{x}_{hik}^T \hat{\mathbf{B}}) = \sum_s w_{hik(gj)} \mathbf{x}_{hik} e_{hik}. \quad (2.4.2)$$

We then solve $\hat{\mathbf{S}}(\mathbf{B}) = \hat{\mathbf{S}}_{(gj)}(\hat{\mathbf{B}})$ for $\mathbf{B}$ to obtain the EF jackknife estimator of $\mathbf{B}$, which is given by

$$\check{\mathbf{B}}_{(gj)} = \hat{\mathbf{B}} - \hat{\mathbf{A}}^{-1} \sum_s w_{hik(gj)} \mathbf{x}_{hik} e_{hik}. \quad (2.4.3)$$

An EF jackknife calibration estimator of the population total Y is given by

$$\check{Y}_{G(gj)} = \mathbf{X}^T \check{\mathbf{B}}_{(gj)} = \sum_s \check{w}_{hik(gj)} y_{hik}, \quad (2.4.4)$$

where $\check{w}_{hik(gj)}$ denote the EF jackknife calibration weights:

$$\begin{aligned} \check{w}_{hik(gj)} &= \tilde{w}_{hik} - w_{hik(gj)} g_{hik} \\ &\quad + w_{hik} \left(\sum_s w_{hik(gj)} g_{hik} \mathbf{x}_{hik}^T \right) \hat{\mathbf{A}}^{-1} \mathbf{x}_{hik}. \end{aligned} \quad (2.4.5)$$

It is important to note that the EF jackknife calibration weights may preserve confidentiality in the data in the sense that we do not get zero weights if the $(gj)^{\text{th}}$ cluster is deleted. Furthermore, the EF jackknife method is computationally simpler to implement because we only have to invert the full matrix $\hat{\mathbf{A}}$ thereby avoiding the inversion of possibly illconditioned matrices $\hat{\mathbf{A}}_{(gj)}$, that may be obtained from the ordinary jackknife method.

An EF jackknife variance estimator of $\hat{Y}_G$ is then given by

$$v_{EFJ}(\hat{Y}_G) = \sum_{g=1}^L \frac{n_g - 1}{n_g} \sum_{j=1}^{n_g} (\check{Y}_{G(gj)} - \hat{Y}_G)^2. \quad (2.4.6)$$

Theorem 2.4 *The EF jackknife variance estimator of $\hat{Y}_G$ is identical to the linearized jackknife variance estimator. That is,*

$$v_{EFJ}(\hat{Y}_G) = v(\bar{e}_{gj}) = v_{JL}(\hat{Y}_G). \quad (2.4.7)$$

Proof: From equations (2.4.4) and (2.4.5), we have

$$\begin{aligned} \check{Y}_{G(gj)} - \hat{Y}_G &= -\sum_s w_{hik(gj)} g_{hik} y_{hik} + \left(\sum_s w_{hik(gj)} g_{hik} \mathbf{x}_{hik}^T \right) \hat{\mathbf{A}}^{-1} \sum_s w_{hik} \mathbf{x}_{hik} y_{hik} \\ &= -\sum_s w_{hik(gj)} g_{hik} y_{hik} + \left(\sum_s w_{hik(gj)} g_{hik} \mathbf{x}_{hik}^T \right) \hat{\mathbf{B}} \\ &= -\sum_s w_{hik(gj)} g_{hik} (y_{hik} - \mathbf{x}_{hik}^T \hat{\mathbf{B}}) \\ &= -\sum_s w_{hik(gj)} g_{hik} e_{hik} \\ &= -\sum_s g_{hik} e_{hik} (w_{hik(gj)} - w_{hik}), \end{aligned}$$

by noting that $\sum_s w_{hik} \mathbf{x}_{hik} e_{hik} = 0$, which implies that $\sum_s w_{hik} (\mathbf{X}^T \hat{\mathbf{A}}^{-1} \mathbf{x}_{hik}) e_{hik} = 0$.

From the definition of the jackknife basic weights we have

$$w_{hik(gj)} - w_{hik} = \begin{cases} -w_{hik} & \text{if } hi = gj \\ \frac{1}{(n_g - 1)} w_{hik} & \text{if } h = g; i \neq j. \\ 0 & \text{if } h \neq g \end{cases}$$

Then

$$\begin{aligned} \check{Y}_{G(gj)} - \hat{Y}_G &= -\left\{ -\sum_k w_{gjk} g_{gjk} e_{gjk} + \frac{1}{n_g - 1} \sum_{i \neq j} \sum_k w_{gik} g_{gik} e_{gik} \right\} \\ &= \frac{1}{n_g} \left\{ \sum_k n_g \bar{w}_{gjk} e_{gjk} - \frac{1}{n_g - 1} \left(\sum_i \sum_k n_g \bar{w}_{gik} e_{gik} - \sum_k n_g \bar{w}_{gjk} e_{gjk} \right) \right\} \end{aligned}$$

$$\begin{aligned}
&= \frac{1}{n_g} \left\{ \bar{e}_{gj} - \frac{1}{n_g - 1} \left(\sum_i \bar{e}_{gi} - \bar{e}_{gj} \right) \right\} \\
&= \frac{1}{n_g - 1} (\bar{e}_{gj} - \bar{\bar{e}}_g)
\end{aligned} \tag{2.4.8}$$

where $\bar{e}_{gj} = \sum_k (n_g \bar{w}_{gjk}) e_{gjk}$ and $\bar{\bar{e}}_g = n_g^{-1} \sum_i \bar{e}_{gi}$. Substituting (2.4.8) into (2.4.6) proves the result in Theorem 2.4. ■

For more general calibration estimators $\hat{\theta}_G$, we can use the EF jackknife variance estimator in (2.4.6) with $\check{Y}_{G(gj)}$ and $\hat{Y}_G$ replaced by $\check{\theta}_{G(gj)}$ and $\hat{\theta}_G$ respectively. Thus we have

$$v_{EFJ}(\hat{\theta}_G) = \sum_{g=1}^L \frac{n_g - 1}{n_g} \sum_{j=1}^{n_g} (\check{\theta}_{G(gj)} - \hat{\theta}_G)^2$$

where $\check{\theta}_{G(gj)}$ involve the EF jackknife calibration weights.

2.4.2 Estimating Function Bootstrap

We will now present the EF bootstrap method (Hu and Kalbfleish, 2000). For the sample estimating function in (2.4.1), replace w_{hik} and $\mathbf{B}$ by $w_{hik}^{(*)}$ defined in (2.3.9) and $\hat{\mathbf{B}}$ respectively. Define

$$\hat{\mathbf{S}}^{(*)}(\hat{\mathbf{B}}) = \sum_s w_{hik}^{(*)} \mathbf{x}_{hik} (y_{hik} - \mathbf{x}_{hik}^T \hat{\mathbf{B}}) = \sum_s w_{hik}^{(*)} \mathbf{x}_{hik} e_{hik} \tag{2.4.9}$$

and solve $\hat{\mathbf{S}}(\mathbf{B}) = \hat{\mathbf{S}}^{(*)}(\hat{\mathbf{B}})$ for $\mathbf{B}$ to obtain the EF bootstrap estimator of $\mathbf{B}$, denoted by $\check{\mathbf{B}}^{(*)} = \hat{\mathbf{B}} - \hat{\mathbf{A}}^{-1} \sum_s w_{hik}^{(*)} \mathbf{x}_{hik} e_{hik}$. Then an EF bootstrap calibration estimator of Y is given by

$$\check{Y}_G^{(*)} = \mathbf{X}^T \check{\mathbf{B}}^{(*)} = \sum_s \check{w}_{hik}^{(*)} y_{hik}, \tag{2.4.10}$$

where the EF bootstrap calibration weights, $\check{w}_{hik}^{(*)}$, are defined by

$$\begin{aligned}\check{w}_{hik}^{(*)} &= \bar{w}_{hik} - w_{hik}^{(*)} g_{hik} \\ &\quad + w_{hik} \left(\sum_s w_{hik}^{(*)} g_{hik} \mathbf{x}_{hik}^T \right) \widehat{\mathbf{A}}^{-1} \mathbf{x}_{hik}.\end{aligned}\quad (2.4.11)$$

We note that the EF bootstrap method involves the inversion of the matrix $\widehat{\mathbf{A}}$ only, and not $\widehat{\mathbf{A}}^{(*)} = \sum_s w_{hik}^{(*)} \mathbf{x}_{hik} \mathbf{x}_{hik}^T$ for each bootstrap sample as in the usual bootstrap method. Therefore, we avoid the inversion of possibly illconditioned matrices $\widehat{\mathbf{A}}^{(*)}$. An EF bootstrap variance estimator of $\widehat{Y}_G$ is given by

$$v_{EFB}(\widehat{Y}_G) = \mathbb{E}_* \left(\check{Y}_G^{(*)} - \widehat{Y}_G \right)^2. \quad (2.4.12)$$

Theorem 2.5 *For $m_h = n_h - 1$, the EF bootstrap variance estimator of $\widehat{Y}_G$ is identical to the Taylor linearization variance estimator $v_{L2}(\widehat{Y}_G)$. That is,*

$$v_{EFB}(\widehat{Y}_G) = v(\bar{e}_{hi}) = v_{L2}(\widehat{Y}_G). \quad (2.4.13)$$

Proof: From (2.4.10) and (2.4.11) we have

$$\check{Y}_G^{(*)} - \widehat{Y}_G = - \sum_s w_{hik}^{(*)} g_{hik} e_{hik}$$

and note the fact that

$$\mathbb{E}_* \left\{ \sum_s w_{hik}^{(*)} g_{hik} e_{hik} \right\} = \sum_s w_{hik} g_{hik} e_{hik} = 0. \quad (2.4.14)$$

Then

$$\begin{aligned}v_{EFB}(\widehat{Y}_G) &= \mathbb{E}_* \left(\check{Y}_G^{(*)} - \widehat{Y}_G \right)^2 \\ &= \mathbb{E}_* \left(- \sum_s w_{hik}^{(*)} g_{hik} e_{hik} \right)^2\end{aligned}$$

$$\begin{aligned}
&= E_* \left(\sum_s w_{hik}^{(*)} g_{hik} e_{hik} \right)^2 \\
&= \text{var}_* \left\{ \sum_s w_{hik}^{(*)} g_{hik} e_{hik} \right\} \quad \text{by (2.4.14)} \\
&= \text{var}_* \left\{ \sum_h \frac{1}{m_h} \sum_{i=1}^{m_h} \sum_k (n_h w_{hik}^{(*)}) g_{hik} e_{hik} \right\} \\
&= \text{var}_* \left\{ \sum_h \bar{e}_h^{(*)} \right\} \quad \text{where } \bar{e}_h^{(*)} = m_h^{-1} \sum_{i=1}^{m_h} e_{hi}^{(*)}, \quad e_{hi}^{(*)} = \sum_k (n_h w_{hik}^{(*)}) g_{hik} e_{hik} \\
&= \sum_h \frac{\text{var}_*(e_{hi}^{(*)})}{m_h} \\
&= \sum_h \frac{1}{m_h} \frac{1}{n_h} \sum_i (\bar{e}_{hi} - \bar{e}_h)^2 \\
&= \sum_h n_h^{-1} (n_h - 1)^{-1} \sum_i (\bar{e}_{hi} - \bar{e}_h)^2 \tag{2.4.15}
\end{aligned}$$

which is identical to $v_{L2}(\hat{Y}_G)$ defined by equation (2.2.13).

■

2.4.3 Estimating Function Balanced Repeated Replication

To apply the estimating function method to the BRR procedure, we replace the basic design weights and $\mathbf{B}$ in (2.4.1) by the BRR basic weights and $\hat{\mathbf{B}}$ respectively, and define

$$\hat{\mathbf{S}}^{(r)}(\hat{\mathbf{B}}) = \sum_s w_{hik}^{(r)} \mathbf{x}_{hik} (y_{hik} - \mathbf{x}_{hik}^T \hat{\mathbf{B}}) = \sum_s w_{hik}^{(r)} \mathbf{x}_{hik} e_{hik}. \tag{2.4.16}$$

Then we solve $\hat{\mathbf{S}}(\mathbf{B}) = \hat{\mathbf{S}}^{(r)}(\hat{\mathbf{B}})$ for $\mathbf{B}$ to obtain an EF BRR estimator, say $\check{\mathbf{B}}^{(r)}$, of $\mathbf{B}$ and define an EF calibration estimator of the population total Y based on the r^{th} half-sample by

$$\check{Y}_G^{(r)} = \mathbf{X}^T \check{\mathbf{B}}^{(r)} = \sum_s \check{w}_{hik}^{(r)} y_{hik}, \tag{2.4.17}$$

where the EF BRR weights for the r^{th} half-sample have the same form as (2.4.5) with the jackknife basic weights, $w_{hik(gj)}$, replaced by the BRR basic weights, $w_{hik}^{(r)}$. An EF BRR variance estimator of $\hat{Y}_G$ is defined by

$$\check{v}_{BRR}(\hat{Y}_G) = \frac{1}{R} \sum_{r=1}^R \left(\check{Y}_G^{(r)} - \hat{Y}_G \right)^2, \quad (2.4.18)$$

which is identical to $v_{L2}(\hat{Y}_G)$ as shown below. For the r^{th} half-sample in the special case of $n_h = 2$ for all h , define

$$\begin{aligned} \check{Y}_G^{(r)} - \hat{Y}_G &= - \sum_s w_{hik}^{(r)} g_{hik} e_{hik} \\ &= - \left\{ \sum_h \frac{1}{2} \sum_i \delta_{hi}^{(r)} \sum_k (n_h w_{hik}) g_{hik} e_{hik} \right\} \\ &= - \left\{ \sum_h \frac{1}{2} \sum_i \delta_{hi}^{(r)} \tilde{e}_{hi} \right\} \quad \text{since } \tilde{e}_{hi} = \sum_k (n_h \tilde{w}_{hik}) e_{hik} \\ &= - \left\{ \sum_h \frac{1}{2} (\delta_{h1}^{(r)} \tilde{e}_{h1} + \delta_{h2}^{(r)} \tilde{e}_{h2}) \right\} \\ &= - \left\{ \sum_h \frac{1}{2} \delta_{h1}^{(r)} (\tilde{e}_{h1} - \tilde{e}_{h2}) + \sum_h \tilde{e}_{h2} \right\} \quad \text{since } \delta_{h1}^{(r)} + \delta_{h2}^{(r)} = 2 \\ &= - \left\{ \sum_h \gamma_h^{(r)} \frac{1}{2} (\tilde{e}_{h1} - \tilde{e}_{h2}) + \sum_h \frac{1}{2} (\tilde{e}_{h1} + \tilde{e}_{h2}) \right\} \quad \text{since } \gamma_h^{(r)} = \delta_{h1}^{(r)} - 1 \\ &= - \left\{ \sum_h \gamma_h^{(r)} \Delta_{\tilde{e}_h} \right\} \end{aligned} \quad (2.4.19)$$

because $\sum_h \frac{1}{2} (\tilde{e}_{h1} + \tilde{e}_{h2}) = \sum_s \tilde{w}_{hik} e_{hik} = 0$. Substituting (2.4.19) into (2.4.18) gives the Taylor linearization variance estimator $v_{L2}(\hat{Y}_G)$.

2.4.4 Estimating Function Modified Balanced Repeated Replication

In this section we will apply the estimating function method to the BRR(ϵ) procedure.

Rather than repeating all the steps involved, we will simply mention the relevant

changes to the EF BRR method discussed in Section 2.4.3. An EF BRR(ϵ) estimator of $\mathbf{B}$, denoted by $\check{\mathbf{B}}^{(r\epsilon)}$, is obtained as the solution of $\hat{\mathbf{S}}(\mathbf{B}) = \hat{\mathbf{S}}^{(r\epsilon)}(\hat{\mathbf{B}})$, where $\hat{\mathbf{S}}^{(r\epsilon)}(\hat{\mathbf{B}})$ takes the form of (2.4.16) with $w_{hik}^{(r)}$ replaced by $w_{hik}^{(r\epsilon)}$. Similarly, an EF BRR(ϵ) estimator of Y based on the r^{th} replicate is defined by

$$\check{Y}_G^{(r\epsilon)} = \mathbf{X}^T \check{\mathbf{B}}^{(r\epsilon)} = \sum_s \check{w}_{hik}^{(r\epsilon)} y_{hik}, \quad (2.4.20)$$

and an EF BRR(ϵ) variance estimator of $\hat{Y}_G$ is given by

$$\check{v}_{BRR(\epsilon)}(\hat{Y}_G) = \frac{1}{\epsilon^2 R} \sum_{r=1}^R \left(\check{Y}_G^{(r\epsilon)} - \hat{Y}_G \right)^2. \quad (2.4.21)$$

Now

$$\begin{aligned} \check{Y}_G^{(r\epsilon)} - \hat{Y}_G &= - \sum_s w_{hik}^{(r\epsilon)} g_{hik} e_{hik} \\ &= - \left\{ \sum_h \gamma_h^{(r\epsilon)} \Delta_{\tilde{e}_h} \right\} \end{aligned} \quad (2.4.22)$$

following the same steps in (2.4.19) with $\delta_{hi}^{(r)}$, $i = 1, 2$ and $\gamma_h^{(r)}$ replaced by $\delta_{hi}^{(r\epsilon)}$ and $\gamma_h^{(r\epsilon)}$ respectively. Again, by substituting (2.4.22) into (2.4.21) we can see that the EF BRR(ϵ) variance estimator of $\hat{Y}_G$ is identical to the alternative Taylor linearization variance estimator in (2.2.13) for the special case of $n_h = 2$ for all h .

Chapter 3

Variance Estimation of General Parameters θ

3.1 Introduction

In this chapter we will extend the estimating function jackknife method to cover finite population parameters that are defined as solutions of census estimating equations. The notion of testing hypotheses on a subset of the vector of parameters is also investigated. For simplicity, we focus on the jackknife but similar results hold for the BRR and bootstrap, using the results of Chapter 2.

An outline of the generation of parameters from the population estimating equations and the estimators obtained from the sample estimating equations is given in Section 3.2. Three methods of estimating variances for the calibration estimators that result from the sample estimating equations are discussed in Section 3.3. Section 3.4 extends the results to Wald and quasi-score tests where we are interested in testing

hypotheses on a subset of the vector of parameters.

3.2 Parameter Estimation

Suppose that the population under consideration, U , consists of M elements. Associated with the $(hik)^{\text{th}}$ element, we have data $(y_{hik}, \mathbf{z}_{hik}^T), k = 1, \dots, M_{hi}, i = 1, \dots, M_h, h = 1, \dots, L$, where the $\mathbf{z}_{hik}$'s are Q -dimensional explanatory variables and the y_{hik} 's are response variables. We assume that for a given value of $\mathbf{z}$, the y -value is generated by some random process with mean

$$E_m(y_{hik}) = \mu_{hik} = \mu(\mathbf{z}_{hik}, \boldsymbol{\theta}), \quad (3.2.1)$$

where $\boldsymbol{\theta}$ is a Q -dimensional vector of parameters. We also assume that we have in mind a working model for the variance of y_{hik} , which is denoted by

$$V_m(y_{hik}) = V_{0hik} = \sigma^2 V_0(\mu_{hik}) \quad (3.2.2)$$

for $(hik) \in U$ and known $V(\cdot)$, where E_m and V_m denote model expectation and model variance respectively. We note that the model for the mean in (3.2.1) is assumed to be valid, but the working variance in (3.2.2) may only be a rough approximation. If we had values for the whole finite population U , we can define a census parameter $\boldsymbol{\theta}_N$ as the solution to the population estimating equations

$$\mathbf{S}(\boldsymbol{\theta}) = \sum_U \mathbf{u}_{hik}(\boldsymbol{\theta}) = \mathbf{0}, \quad (3.2.3)$$

where $\sum_U$ denotes the summation over the finite population U and the q^{th} component of the vector $\mathbf{u}_{hik}(\boldsymbol{\theta})$ is defined by

$$u_{hikq}(\boldsymbol{\theta}) = \frac{\partial \mu_{hik}}{\partial \theta_q} \left\{ \frac{y_{hik} - \mu_{hik}}{V_{0hik}} \right\}, \quad q = 1, \dots, Q. \quad (3.2.4)$$

For example, the model for generating the mean of a random variable y is given by $E_m(y_{hik}) = \mu_{hik} = \theta$, $V_m(y_{hik}) = \sigma^2$, and $\text{Cov}_m(y_{hik}, y_{h'i'k'}) = 0$ for $(hik) \in U$ and $(h'i'k') \in U$ with $(h'i'k') \neq (hik)$. Using (3.2.3) and (3.2.4), this model leads to $u_{hik}(\theta) = y_{hik} - \theta$,

$$S(\theta) = \sum_{(hik) \in U} (y_{hik} - \theta)$$

and

$$\theta_N = M^{-1} \sum_{(hik) \in U} y_{hik} = \bar{Y}$$

is the population mean of the variable y , where M is the total number of elements in the population. Other examples of commonly used census parameters θ_N such as the ratio of two population totals, linear regression and logistic regression coefficients, generated from (3.2.3) are given in Binder (1983), Rao, Yung and Hidirolou (2001) and others.

Census parameters, θ_N , may be motivated through a model for the finite population. However, our interest, in Sections 3.2 and 3.3, is in making design-based inferences on the finite population parameters, θ_N , defined as the solution to the population estimating equations.

In reality, we do not know the values for the whole finite population U but only for those elements in a sample drawn from the population, according to some well-defined sampling scheme. We assume that the sample design provides consistent, asymptotically normal estimators of population totals and associated standard errors. An estimator, $\hat{\theta}$, of the census parameter, θ_N , is then obtained as the solution to the

sample estimating equations

$$\hat{\mathbf{S}}(\boldsymbol{\theta}) = \sum_s w_{hik} \mathbf{u}_{hik}(\boldsymbol{\theta}) = \mathbf{0}, \quad (3.2.5)$$

where $\hat{\mathbf{S}}(\boldsymbol{\theta})$ is the estimator of the population estimating function defined in (3.2.3), which is a vector of population totals for fixed $\boldsymbol{\theta}$ (Binder, 1983).

If we have auxiliary information available at the estimation stage, then we can use the calibration methods, discussed in Section 2.2.2, to improve the precision of sample estimates or to benchmark to known population totals. A calibration estimator, $\hat{\boldsymbol{\theta}}_G$, of the census parameter $\boldsymbol{\theta}_N$ is defined as the solution of

$$\hat{\mathbf{S}}_G(\boldsymbol{\theta}) = \sum_s \tilde{w}_{hik} \mathbf{u}_{hik}(\boldsymbol{\theta}) = \mathbf{0}, \quad (3.2.6)$$

where $\hat{\mathbf{S}}_G(\boldsymbol{\theta})$ is the calibration estimator of $\mathbf{S}(\boldsymbol{\theta})$ and $\tilde{w}_{hik} = w_{hik} g_{hik}$ is the calibration weight for the GREG estimator defined in Section 2.2.2.

We note that for simple parameters of interest, such as totals or means, the solution to the estimating equations has a closed form and the calculation of $\hat{\boldsymbol{\theta}}$ or $\hat{\boldsymbol{\theta}}_G$ is therefore straightforward. For more complex parameters, such as logistic regression, we may have to solve the sample estimating equations iteratively, for example, using the Newton-Raphson algorithm in (3.2.5) whose r^{th} iteration is given by

$$\hat{\boldsymbol{\theta}}_r = \hat{\boldsymbol{\theta}}_{r-1} + \hat{\mathbf{J}}^{-1}(\hat{\boldsymbol{\theta}}_{r-1}) \hat{\mathbf{S}}(\hat{\boldsymbol{\theta}}_{r-1}), \quad (3.2.7)$$

where $\hat{\boldsymbol{\theta}}_{r-1}$ is the value of $\hat{\boldsymbol{\theta}}$ obtained at the $(r-1)^{\text{th}}$ iteration, $\hat{\mathbf{J}}(\hat{\boldsymbol{\theta}}_{r-1}) = -\partial \hat{\mathbf{S}}(\boldsymbol{\theta}) / \partial \boldsymbol{\theta}^T$ evaluated at $\hat{\boldsymbol{\theta}}_{r-1}$ and $\hat{\mathbf{S}}(\hat{\boldsymbol{\theta}}_{r-1})$ is $\hat{\mathbf{S}}(\boldsymbol{\theta})$ evaluated at $\hat{\boldsymbol{\theta}}_{r-1}$. Iterating the Newton-Raphson algorithm to convergence produces the estimate $\hat{\boldsymbol{\theta}}$.

3.3 Variance Estimators

We will now present three methods of estimating variances for the calibration estimators $\hat{\boldsymbol{\theta}}_G$ that result from the sample estimating equations. One method uses the Taylor linearization approach while the other two methods are based on the jackknife procedure. We will also note that the Taylor linearization approach is applicable to general designs, whereas the jackknife procedure is restricted to some particular sample designs.

3.3.1 Taylor linearization

Binder (1983) established general results on the Taylor linearization method for estimates $\hat{\boldsymbol{\theta}}$. This Taylor linearization approach assumes that,

- (i) $\hat{\mathbf{S}}(\boldsymbol{\theta})$ is asymptotically normal with mean $\mathbf{S}(\boldsymbol{\theta})$ and covariance matrix $\mathbf{V}(\hat{\mathbf{S}}(\boldsymbol{\theta}))$,
- (ii) a consistent estimator of $\mathbf{V}(\hat{\mathbf{S}}(\boldsymbol{\theta}))$ is given by $\hat{\mathbf{V}}(\hat{\mathbf{S}}(\boldsymbol{\theta}))$.

To show consistency of $\hat{\boldsymbol{\theta}}$ we will state the following theorem from Rubin-Bleuer (1997).

Theorem 3.1 (Consistency of $\hat{\boldsymbol{\theta}}$) *We assume:*

- (i) *There exists $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}$, the parameter space, and a real valued differentiable function*

$\mathbf{W}(\boldsymbol{\theta})$ with continuous partial derivatives such that $\mathbf{W}(\boldsymbol{\theta}_0) = \mathbf{0}$ and the matrix

$\mathbf{W}'(\boldsymbol{\theta}_0)$ is of full rank where

$$\mathbf{W}'(\boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}} \mathbf{W}(\boldsymbol{\theta}).$$

(ii) There exists a compact neighbourhood $\mathbf{K}(\boldsymbol{\theta}_0)$ of $\boldsymbol{\theta}_0$ on which we have

$$\frac{1}{N} \frac{\partial \hat{\mathbf{S}}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = O_{p_d}(1)$$

as $N \rightarrow \infty$, uniformly in $\boldsymbol{\theta} \in \mathbf{K}(\boldsymbol{\theta}_0)$, where p_d denotes design probability.

(iii) $\hat{\mathbf{S}}(\boldsymbol{\theta})/N \rightarrow \mathbf{W}(\boldsymbol{\theta})$ in design probability for every $\boldsymbol{\theta} \in \boldsymbol{\Theta}$.

Then

$$\hat{\boldsymbol{\theta}} \rightarrow \boldsymbol{\theta}_0$$

in design probability.

Remark 3.1 Conditions for the Laws of Large Numbers in the finite population imply conditions (ii) and (iii) of Theorem 3.1 if

$$\sup_{\boldsymbol{\theta} \in \mathbf{K}(\boldsymbol{\theta}_0)} \left| \frac{\partial \mathbf{u}_{hik}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right| \leq H(y),$$

where $H(y)$ is independent of $\boldsymbol{\theta}$.

Corollary 3.1 Under the conditions of Theorem 3.1, and assuming $\boldsymbol{\theta}_N \rightarrow \boldsymbol{\theta}_0$ (as a sequence of numbers), we have

$$\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_N \rightarrow 0$$

in design probability.

We now assume that the function $\mathbf{u}$ is differentiable and define the finite population Jacobian $\mathbf{J}_N(\boldsymbol{\theta})$ and the sample Jacobian $\hat{\mathbf{J}}(\boldsymbol{\theta})$ by

$$\mathbf{J}_N(\boldsymbol{\theta}) = \frac{1}{N} \frac{\partial \mathbf{S}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$$

and

$$\hat{\mathbf{J}}(\boldsymbol{\theta}) = \frac{1}{N} \frac{\partial \hat{\mathbf{S}}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$$

respectively. We will also use the following conditions to show the asymptotic distribution of $\hat{\boldsymbol{\theta}}$.

Definition 3.1 *Conditions for the asymptotic distribution of $\hat{\boldsymbol{\theta}}$:*

(A) $\hat{\mathbf{J}}(\boldsymbol{\theta}) \longrightarrow \mathbf{J}(\boldsymbol{\theta}_0)$ in design probability as $N \rightarrow \infty$ and $\boldsymbol{\theta} \rightarrow \boldsymbol{\theta}_0$, where $\mathbf{J}(\boldsymbol{\theta}_0)$ is a matrix of full rank.

(B)

$$n^{\frac{1}{2}}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_N) = \hat{\mathbf{J}}^{-1}(\boldsymbol{\xi}_N) \frac{n^{\frac{1}{2}}}{N} \{ \hat{\mathbf{S}}(\hat{\boldsymbol{\theta}}) - \hat{\mathbf{S}}(\boldsymbol{\theta}_N) \}$$

where

$$|\boldsymbol{\xi}_N - \boldsymbol{\theta}_0| \leq |\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_N| + |\boldsymbol{\theta}_N - \boldsymbol{\theta}_0| \longrightarrow 0.$$

For condition (B) to hold, we require conditions (i)-(iii) of Theorem 3.1 and

$\boldsymbol{\theta}_N \rightarrow \boldsymbol{\theta}_0$. Condition (A) ensures $\hat{\mathbf{J}}^{-1}(\boldsymbol{\xi}_N) \longrightarrow \mathbf{J}^{-1}(\boldsymbol{\theta}_0)$. We now define

$$\hat{\mathbf{T}}(\boldsymbol{\theta}) = \frac{n^{\frac{1}{2}}}{N} \{ \hat{\mathbf{S}}(\boldsymbol{\theta}) - \mathbf{S}(\boldsymbol{\theta}) \} \quad (3.3.1)$$

and note that

$$\begin{aligned} n^{\frac{1}{2}}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_N) &= \hat{\mathbf{J}}^{-1}(\boldsymbol{\xi}_N) \frac{n^{\frac{1}{2}}}{N} \{ \hat{\mathbf{S}}(\hat{\boldsymbol{\theta}}) - \hat{\mathbf{S}}(\boldsymbol{\theta}_N) + \mathbf{S}(\boldsymbol{\theta}_N) \} \\ &= -\hat{\mathbf{J}}^{-1}(\boldsymbol{\xi}_N) \frac{n^{\frac{1}{2}}}{N} \{ \hat{\mathbf{S}}(\boldsymbol{\theta}_N) - \mathbf{S}(\boldsymbol{\theta}_N) \} \\ &= -\hat{\mathbf{J}}^{-1}(\boldsymbol{\xi}_N) \hat{\mathbf{T}}(\boldsymbol{\theta}_N) \end{aligned} \quad (3.3.2)$$

since $\hat{\mathbf{S}}(\hat{\boldsymbol{\theta}}) = \mathbf{0}$ and $\mathbf{S}(\boldsymbol{\theta}_N) = \mathbf{0}$. We also have

$$\hat{\mathbf{T}}(\boldsymbol{\theta}_N) = \hat{\mathbf{T}}(\boldsymbol{\theta}_N) - \hat{\mathbf{T}}(\boldsymbol{\theta}_0) + \hat{\mathbf{T}}(\boldsymbol{\theta}_0). \quad (3.3.3)$$

(C) $\hat{\mathbf{T}}(\boldsymbol{\theta}_N) - \hat{\mathbf{T}}(\boldsymbol{\theta}_0) \rightarrow \mathbf{0}$ in design probability as $N \rightarrow \infty$. Rubin-Bleuer (1997)

gave conditions for (C) under two different sampling designs: simple random sampling without replacement (SRSWOR) and sampling with replacement with probability proportional to size (PPSWR). For condition (C) to hold, we require the design variance of $\hat{\mathbf{T}}(\boldsymbol{\theta}_N) - \hat{\mathbf{T}}(\boldsymbol{\theta}_0)$ to converge to zero. Write

$$\begin{aligned}
 & \hat{\mathbf{T}}(\boldsymbol{\theta}_N) - \hat{\mathbf{T}}(\boldsymbol{\theta}_0) \\
 &= \frac{n^{\frac{1}{2}}}{N} \{ \hat{\mathbf{S}}(\boldsymbol{\theta}_N) - \mathbf{S}(\boldsymbol{\theta}_N) \} - \frac{n^{\frac{1}{2}}}{N} \{ \hat{\mathbf{S}}(\boldsymbol{\theta}_0) - \mathbf{S}(\boldsymbol{\theta}_0) \} \\
 &= \frac{n^{\frac{1}{2}}}{N} \mathbf{S}(\boldsymbol{\theta}_0) + \frac{n^{\frac{1}{2}}}{N} \{ \hat{\mathbf{S}}(\boldsymbol{\theta}_N) - \hat{\mathbf{S}}(\boldsymbol{\theta}_0) \} \quad \text{because } \mathbf{S}(\boldsymbol{\theta}_N) = \mathbf{0} \\
 &= \frac{n^{\frac{1}{2}}}{N} \mathbf{S}(\boldsymbol{\theta}_0) + \frac{n^{\frac{1}{2}}}{N} \sum_U w_{hik} \delta_{hik} \{ \mathbf{u}_{hik}(\boldsymbol{\theta}_N) - \mathbf{u}_{hik}(\boldsymbol{\theta}_0) \}, \quad (3.3.4)
 \end{aligned}$$

where δ_{hik} is the sample indicator. We use the following equicontinuity condition on the $\mathbf{u}_{hik}$'s so that the design variance of $\hat{\mathbf{T}}(\boldsymbol{\theta}_N) - \hat{\mathbf{T}}(\boldsymbol{\theta}_0)$ converges to zero.

(D) The function $\mathbf{u}$ is such that for every finite population U , given $\epsilon > 0$ there exists $\delta > 0$ independent of N such that if $|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2| < \delta$ then

$$\sup_{(hik) \in U} |\mathbf{u}_{hik}(\boldsymbol{\theta}_1) - \mathbf{u}_{hik}(\boldsymbol{\theta}_2)| \leq \epsilon.$$

Hence for $\epsilon > 0$ and for $|\boldsymbol{\theta}_N - \boldsymbol{\theta}_0| < \delta$ we have,

$$\mathbf{V}_d(\hat{\mathbf{T}}(\boldsymbol{\theta}_N) - \hat{\mathbf{T}}(\boldsymbol{\theta}_0)) \leq O(1) \epsilon$$

and therefore $\mathbf{V}_d(\hat{\mathbf{T}}(\boldsymbol{\theta}_N) - \hat{\mathbf{T}}(\boldsymbol{\theta}_0)) = o(1)$ as $N \rightarrow \infty$ and as $\boldsymbol{\theta}_N \rightarrow \boldsymbol{\theta}_0$.

(E)

$$\hat{\mathbf{T}}(\boldsymbol{\theta}_0) = \frac{n^{\frac{1}{2}}}{N} \{ \hat{\mathbf{S}}(\boldsymbol{\theta}_0) - \mathbf{S}(\boldsymbol{\theta}_0) \}$$

is asymptotically normal in the design probability space with mean $\mathbf{0}$ and variance $\Sigma(\boldsymbol{\theta}_0)$. Condition (E) follows directly from Krewski and Rao (1981) and requires their conditions for the asymptotic normality of the sample mean.

Theorem 3.2 (Binder, 1983) *Under the regularity conditions in Definition 3.1, then*

$$n^{\frac{1}{2}}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_N)$$

is asymptotically normal in the design probability space with mean $\mathbf{0}$ and variance

$$\mathbf{J}^{-1}(\boldsymbol{\theta}_0)\Sigma(\boldsymbol{\theta}_0)\left[\mathbf{J}^{-1}(\boldsymbol{\theta}_0)\right]^T, \quad (3.3.5)$$

where $\Sigma(\boldsymbol{\theta}_0)$ is the limiting variance of $\hat{\mathbf{S}}(\boldsymbol{\theta}_N)$.

Proof: From equations (3.3.2) and (3.3.3) we have

$$\begin{aligned} n^{\frac{1}{2}}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_N) &= \hat{\mathbf{J}}^{-1}(\boldsymbol{\xi}_N) \left\{ \hat{\mathbf{T}}(\boldsymbol{\theta}_N) - \hat{\mathbf{T}}(\boldsymbol{\theta}_0) + \hat{\mathbf{T}}(\boldsymbol{\theta}_0) \right\} \\ &= \hat{\mathbf{J}}^{-1}(\boldsymbol{\xi}_N) \left\{ \hat{\mathbf{T}}(\boldsymbol{\theta}_N) - \hat{\mathbf{T}}(\boldsymbol{\theta}_0) \right\} + \hat{\mathbf{J}}^{-1}(\boldsymbol{\xi}_N) \hat{\mathbf{T}}(\boldsymbol{\theta}_0), \end{aligned} \quad (3.3.6)$$

where

$$\hat{\mathbf{J}}^{-1}(\boldsymbol{\xi}_N) \left\{ \hat{\mathbf{T}}(\boldsymbol{\theta}_N) - \hat{\mathbf{T}}(\boldsymbol{\theta}_0) \right\} \longrightarrow \mathbf{0} \quad (3.3.7)$$

in design probability by condition (C). Combining conditions (A) and (E) with multivariate versions of Slutsky's Theorem in Lemma 2.3 (ii), we can show that

$$\hat{\mathbf{J}}^{-1}(\boldsymbol{\xi}_N) \hat{\mathbf{T}}(\boldsymbol{\theta}_0) \longrightarrow_d N\left(\mathbf{0}, \mathbf{J}^{-1}(\boldsymbol{\theta}_0)\Sigma(\boldsymbol{\theta}_0)\left[\mathbf{J}^{-1}(\boldsymbol{\theta}_0)\right]^T\right). \quad (3.3.8)$$

Substituting (3.3.7) and (3.3.8) into (3.3.6) and applying Lemma 2.3 (i) completes the proof of Theorem 3.2.

A consistent estimator of the variance of $\hat{\boldsymbol{\theta}}$ is given in equation (3.4) of Binder (1983) as

$$\widehat{\mathbf{V}}_L(\hat{\boldsymbol{\theta}}) = \hat{\mathbf{J}}^{-1}(\hat{\boldsymbol{\theta}}) \widehat{\mathbf{V}}(\hat{\mathbf{S}}(\hat{\boldsymbol{\theta}})) \hat{\mathbf{J}}^{-1}(\hat{\boldsymbol{\theta}}), \quad (3.3.9)$$

where

$$\hat{\mathbf{J}}(\hat{\boldsymbol{\theta}}) = - \left. \frac{\partial \hat{\mathbf{S}}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T} \right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} = \sum_{(hik) \in s} w_{hik} \left. \frac{-\partial \mathbf{u}_{hik}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T} \right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \quad (3.3.10)$$

and $\widehat{\mathbf{V}}(\hat{\mathbf{S}}(\hat{\boldsymbol{\theta}}))$ is equal to $\widehat{\mathbf{V}}(\hat{\mathbf{S}}(\boldsymbol{\theta}))$ evaluated at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$. We note that $\widehat{\mathbf{V}}(\hat{\mathbf{S}}(\boldsymbol{\theta}))$ is obtained from the standard survey variance estimator for a total since $\hat{\mathbf{S}}(\boldsymbol{\theta})$, defined in (3.2.5), is the estimator of the vector of population totals $\mathbf{S}(\boldsymbol{\theta})$ given by (3.2.3).

Rao, Yung and Hidirolou (2002) extended Binder's (1983) Taylor linearization method to obtain variance estimators of parameter estimates, $\hat{\boldsymbol{\theta}}_G$, that can be obtained by solving sample estimating equations in (3.2.6), which involve the design weights, w_{hik} , as well as the adjustment factors, g_{hik} , resulting from poststratification or regression adjustments. In this method we assume that,

- (i) $\hat{\mathbf{S}}_G(\boldsymbol{\theta})$ is asymptotically normal with mean $\mathbf{S}(\boldsymbol{\theta})$ and covariance matrix $\mathbf{V}(\hat{\mathbf{S}}_G(\boldsymbol{\theta}))$,
- (ii) a consistent estimator of $\mathbf{V}(\hat{\mathbf{S}}_G(\boldsymbol{\theta}))$ is given by $\widehat{\mathbf{V}}(\hat{\mathbf{S}}_G(\boldsymbol{\theta}))$.

A consistent estimator of the variance of $\hat{\boldsymbol{\theta}}_G$ is then given by

$$\widehat{\mathbf{V}}_L(\hat{\boldsymbol{\theta}}_G) = \hat{\mathbf{J}}_G^{-1}(\hat{\boldsymbol{\theta}}_G) \widehat{\mathbf{V}}(\hat{\mathbf{S}}_G(\hat{\boldsymbol{\theta}}_G)) \hat{\mathbf{J}}_G^{-1}(\hat{\boldsymbol{\theta}}_G), \quad (3.3.11)$$

where

$$\hat{\mathbf{J}}_G(\hat{\boldsymbol{\theta}}_G) = - \left. \frac{\partial \hat{\mathbf{S}}_G(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T} \right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_G} = \sum_{(hik) \in s} \tilde{w}_{hik} \left. \frac{-\partial \mathbf{u}_{hik}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T} \right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_G} \quad (3.3.12)$$

and $\widehat{\mathbf{V}}(\widehat{\mathbf{S}}_G(\widehat{\boldsymbol{\theta}}_G))$ is an estimator of $\widehat{\mathbf{V}}(\widehat{\mathbf{S}}_G(\boldsymbol{\theta}))$. Since $\widehat{\mathbf{S}}_G(\boldsymbol{\theta})$ is the calibration (GREG) estimator of the total $\mathbf{S}(\boldsymbol{\theta}) = \sum_U \mathbf{u}_{hik}(\boldsymbol{\theta})$, the variance estimator $\widehat{\mathbf{V}}(\widehat{\mathbf{S}}_G(\widehat{\boldsymbol{\theta}}_G))$ is obtained from the GREG results for a total, given in Section 2.2.2. For example, an estimator of the covariance matrix of the GREG estimator

$$\widehat{\mathbf{Y}}_G = \sum_s \tilde{w}_{hik} \mathbf{y}_{hik}, \quad (3.3.13)$$

under with replacement sampling of the first-stage clusters, is given by

$$\begin{aligned} \widehat{\mathbf{V}}_L(\widehat{\mathbf{Y}}_G) &= \sum_h n_h^{-1} (n_h - 1)^{-1} \sum_i (\tilde{\mathbf{e}}_{hi} - \bar{\tilde{\mathbf{e}}}_h) (\tilde{\mathbf{e}}_{hi} - \bar{\tilde{\mathbf{e}}}_h)^T \\ &= v(\tilde{\mathbf{e}}_{hi}) \end{aligned} \quad (3.3.14)$$

where $\mathbf{y}_{hik} = (y_{hik1}, \dots, y_{hikQ})^T$, $\tilde{\mathbf{e}}_{hi} = \sum_k (n_h \tilde{w}_{hik}) \mathbf{e}_{hik}$, $\bar{\tilde{\mathbf{e}}}_h = n_h^{-1} \sum_i \tilde{\mathbf{e}}_{hi}$, $\mathbf{e}_{hik} = (e_{hik1}, \dots, e_{hikQ})^T$ with $e_{hikq} = y_{hikq} - \mathbf{x}_{hik}^T \hat{\mathbf{B}}_q$, $q = 1, \dots, Q$ denoting the residual obtained by regressing the q^{th} component y_{hikq} on the auxiliary variables $\mathbf{x}_{hik}$. Note that $v(\tilde{\mathbf{e}}_{hi})$ indicates that the estimator of the covariance matrix $\widehat{\mathbf{V}}_L(\widehat{\mathbf{Y}}_G)$ depends only on the vector of weighted cluster residuals. Using the expression in (3.3.14), an estimator of the covariance matrix of $\widehat{\mathbf{S}}_G(\widehat{\boldsymbol{\theta}}_G)$ is then given by

$$\widehat{\mathbf{V}}_L(\widehat{\mathbf{S}}_G(\widehat{\boldsymbol{\theta}}_G)) = v(\tilde{\mathbf{e}}_{hi}(\widehat{\boldsymbol{\theta}}_G)), \quad (3.3.15)$$

where $\tilde{\mathbf{e}}_{hi}(\widehat{\boldsymbol{\theta}}_G) = \sum_k (n_h \tilde{w}_{hik}) \mathbf{e}_{hik}(\widehat{\boldsymbol{\theta}}_G)$ and $\mathbf{e}_{hik}(\widehat{\boldsymbol{\theta}}_G) = (e_{hik1}(\widehat{\boldsymbol{\theta}}_G), \dots, e_{hikQ}(\widehat{\boldsymbol{\theta}}_G))^T$. The residuals $e_{hikq}(\widehat{\boldsymbol{\theta}}_G)$, obtained by regressing the q^{th} component of $\mathbf{u}_{hik}(\widehat{\boldsymbol{\theta}}_G)$ on the auxiliary variables $\mathbf{x}_{hik}$, are given by

$$e_{hikq}(\widehat{\boldsymbol{\theta}}_G) = u_{hikq}(\widehat{\boldsymbol{\theta}}_G) - \mathbf{x}_{hik}^T \hat{\mathbf{B}}_q(\widehat{\boldsymbol{\theta}}_G), \quad q = 1, \dots, Q, \quad (3.3.16)$$

where

$$\hat{\mathbf{B}}_q(\widehat{\boldsymbol{\theta}}_G) = \left[\sum_s w_{hik} \mathbf{x}_{hik} \mathbf{x}_{hik}^T \right]^{-1} \left\{ \sum_s w_{hik} \mathbf{x}_{hik} u_{hikq}(\widehat{\boldsymbol{\theta}}_G) \right\}$$

$$= \widehat{\mathbf{A}}^{-1} \widehat{\mathbf{U}}_q(\widehat{\boldsymbol{\theta}}_G). \quad (3.3.17)$$

We note from (3.3.16) that the calibration estimation may not lead to gains in efficiency if the model residuals $u_{hikq}(\widehat{\boldsymbol{\theta}}_G)$ are unrelated to the auxiliary variables $\mathbf{x}_{hik}$. By substituting (3.3.15) into the variance estimator of $\widehat{\boldsymbol{\theta}}_G$ in equation (3.3.11), we get

$$\widehat{\mathbf{V}}_L(\widehat{\boldsymbol{\theta}}_G) = v(\widehat{\mathbf{J}}_G^{-1}(\widehat{\boldsymbol{\theta}}_G) \bar{\mathbf{e}}_{hi}(\widehat{\boldsymbol{\theta}}_G)) \quad (3.3.18)$$

in operator notation. We note that $\widehat{\mathbf{V}}_L(\widehat{\boldsymbol{\theta}}_G)$ incorporates the calibration adjustment factors as well as synthetic residuals obtained by regressing the components of the sample estimating equations on the auxiliary variables.

Following the discussion in Rao, Yung and Hidirolou (2002), suppose we are interested in confidence intervals on the components of $\boldsymbol{\theta}_N$, for example, the first component θ_1 . We can use equation (3.3.18) to find the variance estimator of $\widehat{\theta}_{G1}$ as

$$\widehat{V}_L(\widehat{\theta}_{G1}) = v(\bar{e}_{hi1}^*(\widehat{\boldsymbol{\theta}}_G)), \quad (3.3.19)$$

where $\bar{e}_{hi1}^*(\widehat{\boldsymbol{\theta}}_G)$ is the first element of $\bar{\mathbf{e}}_{hi}^*(\widehat{\boldsymbol{\theta}}_G) = \widehat{\mathbf{J}}_G^{-1}(\widehat{\boldsymbol{\theta}}_G) \bar{\mathbf{e}}_{hi}(\widehat{\boldsymbol{\theta}}_G)$. Then a $(1 - \alpha)$ -level normal theory confidence interval on θ_1 is given by

$$I_{1-\alpha} = \left[\widehat{\theta}_{G1} - z_{\alpha/2} v^{\frac{1}{2}}(\bar{e}_{hi1}^*(\widehat{\boldsymbol{\theta}}_G)), \widehat{\theta}_{G1} + z_{\alpha/2} v^{\frac{1}{2}}(\bar{e}_{hi1}^*(\widehat{\boldsymbol{\theta}}_G)) \right] \quad (3.3.20)$$

where $z_{\alpha/2}$ is the upper $\alpha/2$ -point of a $N(0, 1)$ variable. This interval is asymptotically correct under regularity conditions.

For the confidence interval in (3.3.20), we denote $\mathbf{u}_{hik}(\boldsymbol{\theta}) = [u_{hik1}(\boldsymbol{\theta}), \mathbf{u}_{hik2}(\boldsymbol{\theta})^T]^T$, where $\mathbf{u}_{hik2}(\boldsymbol{\theta})$ is a $(Q - 1)$ vector. Suppose that $\mathbf{u}_{hik2}(\boldsymbol{\theta})$ does not depend on θ_1 and the first derivative of $u_{hik1}(\boldsymbol{\theta})$ with respect to θ_1 is independent of $\boldsymbol{\theta}_2$, where

$\boldsymbol{\theta} = (\theta_1, \boldsymbol{\theta}_2^T)^T$. Then expressing $\hat{\mathbf{J}}_G(\boldsymbol{\theta})$ as a partitioned matrix, we have

$$\begin{aligned}\hat{\mathbf{J}}_G(\boldsymbol{\theta}) &= -\frac{\partial \hat{\mathbf{S}}_G(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T} \\ &= \begin{bmatrix} \hat{J}_{G11}(\boldsymbol{\theta}) & \hat{\mathbf{J}}_{G12}(\boldsymbol{\theta}) \\ \hat{\mathbf{J}}_{G21}(\boldsymbol{\theta}) & \hat{\mathbf{J}}_{G22}(\boldsymbol{\theta}) \end{bmatrix}\end{aligned}$$

where $\hat{\mathbf{J}}_{G12}(\boldsymbol{\theta}) = \mathbf{0}$, $\hat{J}_{G11}(\boldsymbol{\theta}) = \hat{J}_{G11}(\theta_1)$ and $\tilde{e}_{hi1}^*(\hat{\boldsymbol{\theta}}_G)$ reduces to

$$\tilde{e}_{hi1}^*(\hat{\boldsymbol{\theta}}_G) = \hat{J}_{G11}^{-1}(\hat{\theta}_{G1}) \left[\hat{\mathbf{J}}_{G12}(\hat{\boldsymbol{\theta}}_G) \hat{\mathbf{J}}_{G22}^{-1}(\hat{\boldsymbol{\theta}}_G) \tilde{e}_{hi2}(\hat{\boldsymbol{\theta}}_G) - \tilde{e}_{hi1}(\hat{\boldsymbol{\theta}}_G) \right]. \quad (3.3.21)$$

3.3.2 Linearized jackknife

In stratified multistage sampling, we delete one first-stage cluster at a time. When the $(gj)^{\text{th}}$ cluster is deleted, we define the sample estimating equations by

$$\hat{\mathbf{S}}_{G(gj)}(\boldsymbol{\theta}) = \sum_s \tilde{w}_{hik(gj)} \mathbf{u}_{hik}(\boldsymbol{\theta}) \quad (3.3.22)$$

where $\tilde{w}_{hik(gj)}$ are the jackknife calibration weights defined in equation (2.3.1). A jackknife calibration estimator of $\boldsymbol{\theta}$, say $\hat{\boldsymbol{\theta}}_{G(gj)}$, can be obtained by either solving $\hat{\mathbf{S}}_{G(gj)}(\boldsymbol{\theta}) = \mathbf{0}$ using the Newton-Raphson algorithm or use the one-step jackknife method proposed by Lipsitz, Dear and Zhao (1994). The one-step jackknife method uses the full sample estimate, $\hat{\boldsymbol{\theta}}_G$, as the starting value and performs only one step of the Newton-Raphson algorithm with the calibration weights, $\tilde{w}_{hik}$, replaced by the jackknife calibration weights, $\tilde{w}_{hik(gj)}$. That is,

$$\hat{\boldsymbol{\theta}}_{G(gj)} = \hat{\boldsymbol{\theta}}_G + \hat{\mathbf{J}}_{G(gj)}^{-1}(\hat{\boldsymbol{\theta}}_G) \hat{\mathbf{S}}_{G(gj)}(\hat{\boldsymbol{\theta}}_G) \quad (3.3.23)$$

where $\hat{\mathbf{J}}_{G(gj)}(\hat{\boldsymbol{\theta}}_G)$ and $\hat{\mathbf{S}}_{G(gj)}(\hat{\boldsymbol{\theta}}_G)$ are obtained from $\hat{\mathbf{J}}_G(\hat{\boldsymbol{\theta}}_G)$ and $\hat{\mathbf{S}}_G(\hat{\boldsymbol{\theta}}_G)$ with $\tilde{w}_{hik}$ replaced by $\tilde{w}_{hik(gj)}$. Then a jackknife estimator of the covariance matrix of $\hat{\boldsymbol{\theta}}_G$ is

given by

$$\widehat{\mathbf{V}}_J(\widehat{\boldsymbol{\theta}}_G) = \sum_g \frac{n_g - 1}{n_g} \sum_j \left(\widehat{\boldsymbol{\theta}}_{G(gj)} - \widehat{\boldsymbol{\theta}}_G \right) \left(\widehat{\boldsymbol{\theta}}_{G(gj)} - \widehat{\boldsymbol{\theta}}_G \right)^T. \quad (3.3.24)$$

Noting that $\widehat{\mathbf{J}}_{G(gj)}^{-1}(\widehat{\boldsymbol{\theta}}_G) \approx \widehat{\mathbf{J}}_G^{-1}(\widehat{\boldsymbol{\theta}}_G)$ and $\widehat{\mathbf{S}}_G(\widehat{\boldsymbol{\theta}}_G) = \mathbf{0}$, we can express equation (3.3.23) as

$$\widehat{\boldsymbol{\theta}}_{G(gj)} \approx \widehat{\boldsymbol{\theta}}_G + \widehat{\mathbf{J}}_G^{-1}(\widehat{\boldsymbol{\theta}}_G) \left\{ \widehat{\mathbf{S}}_{G(gj)}(\widehat{\boldsymbol{\theta}}_G) - \widehat{\mathbf{S}}_G(\widehat{\boldsymbol{\theta}}_G) \right\} \quad (3.3.25)$$

and equation (3.3.24) becomes

$$\widehat{\mathbf{V}}_J(\widehat{\boldsymbol{\theta}}_G) \approx \widehat{\mathbf{J}}_G^{-1}(\widehat{\boldsymbol{\theta}}_G) \widehat{\mathbf{V}}_J(\widehat{\mathbf{S}}_G(\widehat{\boldsymbol{\theta}}_G)) \widehat{\mathbf{J}}_G^{-1}(\widehat{\boldsymbol{\theta}}_G), \quad (3.3.26)$$

where $\widehat{\mathbf{V}}_J(\widehat{\mathbf{S}}_G(\widehat{\boldsymbol{\theta}}_G))$ is the jackknife variance estimator of $\widehat{\mathbf{S}}_G(\widehat{\boldsymbol{\theta}}_G) = \sum_s \tilde{w}_{hik} \mathbf{u}_{hik}(\widehat{\boldsymbol{\theta}}_G)$ given by

$$\widehat{\mathbf{V}}_J(\widehat{\mathbf{S}}_G(\widehat{\boldsymbol{\theta}}_G)) = \sum_g \frac{n_g - 1}{n_g} \sum_j \left(\widehat{\mathbf{S}}_{G(gj)}(\widehat{\boldsymbol{\theta}}_G) - \widehat{\mathbf{S}}_G(\widehat{\boldsymbol{\theta}}_G) \right) \left(\widehat{\mathbf{S}}_{G(gj)}(\widehat{\boldsymbol{\theta}}_G) - \widehat{\mathbf{S}}_G(\widehat{\boldsymbol{\theta}}_G) \right)^T. \quad (3.3.27)$$

Since $\widehat{\mathbf{S}}_G(\widehat{\boldsymbol{\theta}}_G)$ is a GREG estimator of total $\mathbf{S}(\boldsymbol{\theta}) = \sum_U \mathbf{u}_{hik}(\boldsymbol{\theta})$, we can apply the techniques in Yung and Rao (1996) to obtain a linearized jackknife variance estimator defined by

$$\widehat{\mathbf{V}}_{JL}(\widehat{\mathbf{S}}_G(\widehat{\boldsymbol{\theta}}_G)) = v(\tilde{\mathbf{e}}_{gj}(\widehat{\boldsymbol{\theta}}_G)), \quad (3.3.28)$$

where $\tilde{\mathbf{e}}_{gj}(\widehat{\boldsymbol{\theta}}_G) = \sum_k (n_g \tilde{w}_{gjk}) \mathbf{e}_{gjk}(\widehat{\boldsymbol{\theta}}_G)$. It now follows from (3.3.26) and (3.3.28) that $\widehat{\mathbf{V}}_J(\widehat{\boldsymbol{\theta}}_G)$ can be approximated by the linearized jackknife variance estimator (Rao, Yung and Hidirolou, 2002)

$$\widehat{\mathbf{V}}_{JL}(\widehat{\boldsymbol{\theta}}_G) = v(\widehat{\mathbf{J}}_G^{-1}(\widehat{\boldsymbol{\theta}}_G) \tilde{\mathbf{e}}_{gj}(\widehat{\boldsymbol{\theta}}_G)), \quad (3.3.29)$$

which is identical to the Taylor linearization variance estimator in (3.3.18).

3.3.3 Estimating function jackknife

In this section we will apply the EF jackknife method to find EF jackknife estimators of the census parameters $\boldsymbol{\theta}_N$. Since the sample estimating equations in (3.2.6) involve the calibration adjustments, g_{hik} , we can use two options for the EF jackknife method. The first option is to replace the calibration weights, $\bar{w}_{hik}$, in (3.2.6) by the EF jackknife calibration weights, $\check{w}_{hik(gj)}$, defined by (2.4.5). Let

$$\check{\mathbf{S}}_{G(gj)}(\boldsymbol{\theta}) = \sum_s \check{w}_{hik(gj)} \mathbf{u}_{hik}(\boldsymbol{\theta}) \quad (3.3.30)$$

be an EF jackknife calibration estimator of $\mathbf{S}(\boldsymbol{\theta})$. Denote the solution of $\check{\mathbf{S}}_{G(gj)}(\boldsymbol{\theta}) = \mathbf{0}$ by $\check{\boldsymbol{\theta}}_{G1(gj)}$, which is an EF jackknife calibration estimator of $\boldsymbol{\theta}_N$. We now use the one-step jackknife method of Lipsitz et. al. (1994), with $\hat{\boldsymbol{\theta}}_G$ as the starting value, to derive

$$\check{\boldsymbol{\theta}}_{G1(gj)} = \hat{\boldsymbol{\theta}}_G + \check{\mathbf{J}}_{G(gj)}^{-1}(\hat{\boldsymbol{\theta}}_G) \check{\mathbf{S}}_{G(gj)}(\hat{\boldsymbol{\theta}}_G) \quad (3.3.31)$$

where

$$\check{\mathbf{J}}_{G(gj)}(\hat{\boldsymbol{\theta}}_G) = - \left. \frac{\partial \check{\mathbf{S}}_{G(gj)}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T} \right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_G} = \sum_s \check{w}_{hik(gj)} \left. \frac{-\partial \mathbf{u}_{hik}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T} \right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_G}$$

and $\check{\mathbf{S}}_{G(gj)}(\hat{\boldsymbol{\theta}}_G)$ is $\check{\mathbf{S}}_{G(gj)}(\boldsymbol{\theta})$ evaluated at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_G$. An EF jackknife estimator of $\text{cov}(\hat{\boldsymbol{\theta}}_G)$ is given by

$$\widehat{\mathbf{V}}_{EFJ}(\hat{\boldsymbol{\theta}}_G) = \sum_g \frac{n_g - 1}{n_g} \sum_j \left(\check{\boldsymbol{\theta}}_{G1(gj)} - \hat{\boldsymbol{\theta}}_G \right) \left(\check{\boldsymbol{\theta}}_{G1(gj)} - \hat{\boldsymbol{\theta}}_G \right)^T. \quad (3.3.32)$$

Theorem 3.3 *The EF jackknife estimator of $\text{cov}(\hat{\boldsymbol{\theta}}_G)$ is approximately equal to the jackknife linearization variance estimator. That is*

$$\widehat{\mathbf{V}}_{EFJ}(\hat{\boldsymbol{\theta}}_G) \approx v \left(\hat{\mathbf{J}}_G^{-1}(\hat{\boldsymbol{\theta}}_G) \bar{\mathbf{e}}_{gj}(\hat{\boldsymbol{\theta}}_G) \right) = \widehat{\mathbf{V}}_{JL}(\hat{\boldsymbol{\theta}}_G).$$

Proof: We note that

$$\begin{aligned}\check{\mathbf{J}}_{G(gj)}(\hat{\boldsymbol{\theta}}_G) &= \hat{\mathbf{J}}_G(\hat{\boldsymbol{\theta}}_G) + \hat{\mathbf{E}}_{(gj)}(\hat{\boldsymbol{\theta}}_G) \\ &= \hat{\mathbf{J}}_G(\hat{\boldsymbol{\theta}}_G) \left\{ \mathbf{I} + \hat{\mathbf{J}}_G^{-1}(\hat{\boldsymbol{\theta}}_G) \hat{\mathbf{E}}_{(gj)}(\hat{\boldsymbol{\theta}}_G) \right\}\end{aligned}$$

where

$$\hat{\mathbf{E}}_{(gj)}(\hat{\boldsymbol{\theta}}_G) = \sum_s w_{hik(gj)} g_{hik} \left. \frac{\partial \mathbf{e}_{hik}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T} \right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_G}.$$

Using the matrix equality

$$(\mathbf{I} + \mathbf{PQ})^{-1} = \mathbf{I} - \mathbf{P}(\mathbf{I} + \mathbf{QP})^{-1}\mathbf{Q},$$

we can show that

$$\begin{aligned}\check{\mathbf{J}}_{G(gj)}^{-1}(\hat{\boldsymbol{\theta}}_G) &= \left[\hat{\mathbf{J}}_G(\hat{\boldsymbol{\theta}}_G) \left\{ \mathbf{I} + \hat{\mathbf{J}}_G^{-1}(\hat{\boldsymbol{\theta}}_G) \hat{\mathbf{E}}_{(gj)}(\hat{\boldsymbol{\theta}}_G) \right\} \right]^{-1} \\ &\approx \hat{\mathbf{J}}_G^{-1}(\hat{\boldsymbol{\theta}}_G)\end{aligned}$$

by ignoring the higher order terms. Noting that $\hat{\mathbf{S}}_G(\hat{\boldsymbol{\theta}}_G) = \mathbf{0}$, (3.3.31) becomes

$$\check{\boldsymbol{\theta}}_{G1(gj)} \approx \hat{\boldsymbol{\theta}}_G + \hat{\mathbf{J}}_G^{-1}(\hat{\boldsymbol{\theta}}_G) \left\{ \check{\mathbf{S}}_{G(gj)}(\hat{\boldsymbol{\theta}}_G) - \hat{\mathbf{S}}_G(\hat{\boldsymbol{\theta}}_G) \right\}. \quad (3.3.33)$$

Using (3.3.33) in the EF jackknife estimator of $\text{cov}(\hat{\boldsymbol{\theta}}_G)$, we have

$$\begin{aligned}\widehat{\mathbf{V}}_{EFJ}(\hat{\boldsymbol{\theta}}_G) &\approx \hat{\mathbf{J}}_G^{-1}(\hat{\boldsymbol{\theta}}_G) \sum_g \frac{n_g - 1}{n_g} \sum_j \left(\check{\mathbf{S}}_{G(gj)}(\hat{\boldsymbol{\theta}}_G) - \hat{\mathbf{S}}_G(\hat{\boldsymbol{\theta}}_G) \right) \\ &\quad \left(\check{\mathbf{S}}_{G(gj)}(\hat{\boldsymbol{\theta}}_G) - \hat{\mathbf{S}}_G(\hat{\boldsymbol{\theta}}_G) \right)^T \hat{\mathbf{J}}_G^{-1}(\hat{\boldsymbol{\theta}}_G) \\ &= \hat{\mathbf{J}}_G^{-1}(\hat{\boldsymbol{\theta}}_G) \widehat{\mathbf{V}}_{EFJ} \left(\hat{\mathbf{S}}_G(\hat{\boldsymbol{\theta}}_G) \right) \hat{\mathbf{J}}_G^{-1}(\hat{\boldsymbol{\theta}}_G)\end{aligned} \quad (3.3.34)$$

where we then use the results in Section 2.4.1 to obtain $\widehat{\mathbf{V}}_{EFJ} \left(\hat{\mathbf{S}}_G(\hat{\boldsymbol{\theta}}_G) \right)$. We now replace y_{hik} in the proof of Theorem 2.4 by a Q -dimensional vector $\mathbf{u}_{hik}(\hat{\boldsymbol{\theta}}_G)$ and

obtain

$$\begin{aligned}
\check{S}_{G(gj)}(\hat{\theta}_G) - \hat{S}_G(\hat{\theta}_G) &= - \sum_s w_{hik(gj)} g_{hik} \mathbf{e}_{hik}(\hat{\theta}_G) \\
&= - \sum_s g_{hik} \mathbf{e}_{hik}(\hat{\theta}_G) (w_{hik(gj)} - w_{hik}) \\
&= \frac{1}{n_g - 1} (\bar{\mathbf{e}}_{gj}(\hat{\theta}_G) - \bar{\mathbf{e}}_g(\hat{\theta}_G)) \quad (3.3.35)
\end{aligned}$$

following the steps in (2.4.8), where $\bar{\mathbf{e}}_{gj}(\hat{\theta}_G) = \sum_k (n_g \tilde{w}_{gjk}) \mathbf{e}_{gjk}(\hat{\theta}_G)$ and $\bar{\mathbf{e}}_g(\hat{\theta}_G) = n_g^{-1} \sum_j \bar{\mathbf{e}}_{gj}(\hat{\theta}_G)$. Then substituting

$$\widehat{\mathbf{V}}_{EFJ}(\hat{S}_G(\hat{\theta}_G)) = v(\bar{\mathbf{e}}_{gj}(\hat{\theta}_G))$$

into (3.3.34) completes the proof. ■

The second option is to replace $\tilde{w}_{hik}$ and θ in (3.2.6) by $\check{w}_{hik(gj)}$ and $\hat{\theta}_G$ respectively, and define

$$\check{S}_{G(gj)}(\hat{\theta}_G) = \sum_s \check{w}_{hik(gj)} \mathbf{u}_{hik}(\hat{\theta}_G). \quad (3.3.36)$$

We now solve $\hat{S}_G(\theta) = \check{S}_{G(gj)}(\hat{\theta}_G)$ for θ to get another EF jackknife calibration estimator, denoted by $\check{\theta}_{G(gj)}$. This option avoids repeated inversions in the Newton-Raphson steps (3.3.23). For this option we will examine the linear regression case, where the model for generating the parameters of a linear regression of y on a Q -dimensional vector of explanatory variables $\mathbf{z}$ is given by $E(y_{hik}) = \mathbf{z}_{hik}^T \theta$, $V(y_{hik}) = \sigma^2$ and $\text{Cov}(y_{hik}, y_{h'i'k'}) = 0$, $hik \neq h'i'k'$. In this case we have $\mathbf{u}_{hik}(\theta) = \mathbf{z}_{hik}(y_{hik} - \mathbf{z}_{hik}^T \theta)$,

$$\mathbf{S}(\theta) = \sum_U \mathbf{z}_{hik}(y_{hik} - \mathbf{z}_{hik}^T \theta),$$

and the census parameter vector is given by

$$\boldsymbol{\theta}_N = \left(\sum_U \mathbf{z}_{hik} \mathbf{z}_{hik}^T \right)^{-1} \sum_U \mathbf{z}_{hik} y_{hik}.$$

Now the calibration estimator of $\mathbf{S}(\boldsymbol{\theta})$ is defined as

$$\hat{\mathbf{S}}_G(\boldsymbol{\theta}) = \sum_s \tilde{w}_{hik} \mathbf{z}_{hik} (y_{hik} - \mathbf{z}_{hik}^T \boldsymbol{\theta}) \quad (3.3.37)$$

and the calibration estimator of $\boldsymbol{\theta}_N$ is

$$\hat{\boldsymbol{\theta}}_G = \hat{\mathbf{A}}_z^{-1} \sum_s \tilde{w}_{hik} \mathbf{z}_{hik} y_{hik} \quad (3.3.38)$$

where $\hat{\mathbf{A}}_z = \sum_s \tilde{w}_{hik} \mathbf{z}_{hik} \mathbf{z}_{hik}^T$. A solution of $\hat{\mathbf{S}}_G(\boldsymbol{\theta}) = \check{\mathbf{S}}_{G(gj)}(\hat{\boldsymbol{\theta}}_G)$ is given by

$$\begin{aligned} \check{\boldsymbol{\theta}}_{G(gj)} &= \hat{\mathbf{A}}_z^{-1} \sum_s \tilde{w}_{hik} \mathbf{z}_{hik} y_{hik} - \hat{\mathbf{A}}_z^{-1} \check{\mathbf{S}}_{G(gj)}(\hat{\boldsymbol{\theta}}_G) \\ &= \hat{\boldsymbol{\theta}}_G - \hat{\mathbf{A}}_z^{-1} \check{\mathbf{S}}_{G(gj)}(\hat{\boldsymbol{\theta}}_G). \end{aligned} \quad (3.3.39)$$

Note that $\check{\boldsymbol{\theta}}_{G(gj)}$ does not require the repeated inversions of $\hat{\mathbf{A}}_z$ -type matrices, unlike the usual jackknife or other resampling methods. From the definition of the EF jackknife calibration weights $\check{w}_{hik(gj)}$ we can re-write (3.3.36) as

$$\begin{aligned} \check{\mathbf{S}}_{G(gj)}(\hat{\boldsymbol{\theta}}_G) &= \hat{\mathbf{S}}_G(\hat{\boldsymbol{\theta}}_G) - \sum_s w_{hik(gj)} g_{hik} \mathbf{e}_{hik}(\hat{\boldsymbol{\theta}}_G) \\ &= \hat{\mathbf{S}}_G(\hat{\boldsymbol{\theta}}_G) - \sum_s g_{hik} \mathbf{e}_{hik}(\hat{\boldsymbol{\theta}}_G) \{w_{hik(gj)} - w_{hik}\} \end{aligned} \quad (3.3.40)$$

since $\sum_s w_{hik} \mathbf{x}_{hik} e_{hikq}(\hat{\boldsymbol{\theta}}_G) = 0$, $q = 1, \dots, Q$ by regressing the q^{th} component of $\mathbf{u}_{hik}(\hat{\boldsymbol{\theta}}_G)$ on the auxiliary indicator variables $\mathbf{x}_{hik}$. Then

$$\check{\mathbf{S}}_{G(gj)}(\hat{\boldsymbol{\theta}}_G) - \hat{\mathbf{S}}_G(\hat{\boldsymbol{\theta}}_G) = \frac{1}{n_g - 1} (\bar{\mathbf{e}}_{gj}(\hat{\boldsymbol{\theta}}_G) - \bar{\mathbf{e}}_g(\hat{\boldsymbol{\theta}}_G)) \quad (3.3.41)$$

following the steps in (2.4.8) where $\tilde{\mathbf{e}}_{gj}(\hat{\boldsymbol{\theta}}_G) = \sum_k (n_g \tilde{w}_{gjk}) \mathbf{e}_{gjk}(\hat{\boldsymbol{\theta}}_G)$. Using (3.3.39) and (3.3.41) we obtain an EF jackknife estimator of $\text{Cov}(\hat{\boldsymbol{\theta}}_G)$ as

$$\begin{aligned} \widehat{\mathbf{V}}_{EFJ2}(\hat{\boldsymbol{\theta}}_G) &= \sum_g \frac{n_g - 1}{n_g} \sum_j (\tilde{\boldsymbol{\theta}}_{G(gj)} - \hat{\boldsymbol{\theta}}_G) (\tilde{\boldsymbol{\theta}}_{G(gj)} - \hat{\boldsymbol{\theta}}_G)^T \\ &= v \left(\widehat{\mathbf{A}}_z^{-1} \tilde{\mathbf{e}}_{gj}(\hat{\boldsymbol{\theta}}_G) \right). \end{aligned} \quad (3.3.42)$$

Note that (3.3.42) is the same as $\widehat{\mathbf{V}}_L(\hat{\boldsymbol{\theta}}_G)$ defined in (3.3.18) because $\hat{\mathbf{J}}_G(\hat{\boldsymbol{\theta}}_G) = \widehat{\mathbf{A}}_z$ for the linear regression model.

We do not know if the relationship in (3.3.42) will hold in general. Further work is needed.

3.4 Wald and Quasi-Score Tests

In this section we will discuss design-model inferences on a model parameter vector $\boldsymbol{\theta}$ assuming a super-population model. In particular, we are interested in testing a null hypothesis of the form

$$H_0 : \boldsymbol{\theta}_2 = \boldsymbol{\theta}_{20}$$

using the sample data, where $\boldsymbol{\theta}$ is partitioned as $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^T, \boldsymbol{\theta}_2^T)^T$ with $\boldsymbol{\theta}_2$ being a $R \times 1$ vector, and $\boldsymbol{\theta}_1$ a $(Q - R) \times 1$ vector. Corresponding to this partition of $\boldsymbol{\theta}$, we also partition $\hat{\boldsymbol{\theta}}_G$ as $\hat{\boldsymbol{\theta}}_G = (\hat{\boldsymbol{\theta}}_{G1}^T, \hat{\boldsymbol{\theta}}_{G2}^T)^T$ and a consistent variance estimator of $\hat{\boldsymbol{\theta}}_G$ as

$$\widehat{\mathbf{V}}(\hat{\boldsymbol{\theta}}_G) = \begin{pmatrix} \widehat{\mathbf{V}}_{11} & \widehat{\mathbf{V}}_{12} \\ \widehat{\mathbf{V}}_{21} & \widehat{\mathbf{V}}_{22} \end{pmatrix}, \quad (3.4.1)$$

where $\widehat{\mathbf{V}}_{11} = \widehat{\mathbf{V}}_{11}(\hat{\boldsymbol{\theta}}_G)$ is the upper diagonal block of the matrix $\widehat{\mathbf{V}}(\hat{\boldsymbol{\theta}}_G)$ with similar expressions for $\widehat{\mathbf{V}}_L(\hat{\boldsymbol{\theta}}_G)$, $\widehat{\mathbf{V}}_J(\hat{\boldsymbol{\theta}}_G)$, $\widehat{\mathbf{V}}_{JL}(\hat{\boldsymbol{\theta}}_G)$ and $\widehat{\mathbf{V}}_{EFJ}(\hat{\boldsymbol{\theta}}_G)$.

We assume that the census y -values are generated by a random process with model expectation and model variance given by (3.2.1) and (3.2.2) respectively. Using the $\mathbf{u}$ -function defined in (3.2.4), we can derive the census parameter, $\boldsymbol{\theta}_N$, and the calibration estimator, $\hat{\boldsymbol{\theta}}_G$, from equations (3.2.3) and (3.2.6) respectively.

3.4.1 Wald Tests

One approach to testing H_0 is to use the Wald statistic

$$X_W^2 = (\hat{\boldsymbol{\theta}}_{G2} - \boldsymbol{\theta}_{20})^T \widehat{\mathbf{V}}_{22}^{-1} (\hat{\boldsymbol{\theta}}_{G2} - \boldsymbol{\theta}_{20}) \quad (3.4.2)$$

based on the calibration estimator, $\hat{\boldsymbol{\theta}}_G$, where $\widehat{\mathbf{V}}_{22}$ is the lower diagonal block of the matrix $\widehat{\mathbf{V}}(\hat{\boldsymbol{\theta}}_G)$, given by (3.4.1).

For general sampling designs, we can use the Taylor linearization variance estimator $\widehat{\mathbf{V}}_L(\hat{\boldsymbol{\theta}}_G)$. For stratified random sampling or stratified multistage sampling, we can use jackknife variance estimators $\widehat{\mathbf{V}}_J(\hat{\boldsymbol{\theta}}_G)$, $\widehat{\mathbf{V}}_{JL}(\hat{\boldsymbol{\theta}}_G)$ or $\widehat{\mathbf{V}}_{EFJ}(\hat{\boldsymbol{\theta}}_{G2})$. As discussed in Rao, Yung and Hidirolou (2002), suppose we use $\widehat{\mathbf{V}}_{L22}$ in (3.4.1) and we write $n_e^{\frac{1}{2}}(\hat{\boldsymbol{\theta}}_{G2} - \boldsymbol{\theta}_{20})$ as

$$\begin{aligned} \hat{\mathbf{T}} &= n_e^{\frac{1}{2}}(\hat{\boldsymbol{\theta}}_{G2} - \boldsymbol{\theta}_{20}) \\ &= n_e^{\frac{1}{2}}(\hat{\boldsymbol{\theta}}_{G2} - \boldsymbol{\theta}_{2N}) + \frac{n_e^{\frac{1}{2}}}{N_e^{\frac{1}{2}}} N_e^{\frac{1}{2}}(\boldsymbol{\theta}_{2N} - \boldsymbol{\theta}_{20}), \end{aligned} \quad (3.4.3)$$

where n_e and N_e are the “effective” sample and population sizes respectively and $\boldsymbol{\theta}_N = (\boldsymbol{\theta}_{1N}^T, \boldsymbol{\theta}_{2N}^T)^T$ is the census parameter. For example, under two-stage cluster sampling, we may take n_e and N_e as the number of sample clusters and the number of population clusters respectively, assuming that the population is generated by a model that induces correlated model errors within clusters.

Under regularity conditions, Rao, Yung and Hidirolou (2002) noted that:

- (i) $\hat{\mathbf{T}}_1 = n_e^{\frac{1}{2}}(\hat{\boldsymbol{\theta}}_{G2} - \boldsymbol{\theta}_{2N})$ converges in design distribution to $N_R(\mathbf{0}, \boldsymbol{\Sigma}_{22N})$,
- (ii) $n_e \widehat{\mathbf{V}}_{L22}$ converges in design probability to $\boldsymbol{\Sigma}_{22N}$, and
- (iii) $\mathbf{T}_{2N} = N_e^{\frac{1}{2}}(\boldsymbol{\theta}_{2N} - \boldsymbol{\theta}_{20})$ converges in model distribution, under H_0 , to $N_R(\mathbf{0}, \boldsymbol{\Sigma}_{22}^*)$.

Rubin-Bleuer and Shiopu-Kratina (2001) have shown that $\hat{\mathbf{T}}_1$ and $\mathbf{T}_{2N}$ are asymptotically independent so that $\hat{\mathbf{T}}$, given by (3.4.3), converges in distribution to $N_R(\mathbf{0}, \boldsymbol{\Sigma}_{22} + f_e \boldsymbol{\Sigma}_{22}^*)$, where $f_e = \lim(n_e/N_e)$ and $\boldsymbol{\Sigma}_{22N}$ converges in model probability to $\boldsymbol{\Sigma}_{22}$. Assuming that $f_e = 0$, which is typically satisfied in large scale surveys if the model errors associated with the finite population are regarded as independent, then the Wald statistic $X_{W'}^2$ is asymptotically χ_R^2 under H_0 .

3.4.2 Quasi-Score Tests

Wald tests have the following drawbacks: (i) The full model has to be fitted to get the parameter estimate $\hat{\boldsymbol{\theta}}_G$; (ii) Wald statistics are not invariant to nonlinear transformations of $\boldsymbol{\theta}$ (Boos, 1992). An alternative approach is to use C.R.Rao's (1948) score tests, which require only fitting the simpler null model H_0 . Rao's score tests have considerable advantages if the full model contains a large number of parameters, and they are free of both limitations associated with Wald tests. Rao, Scott and Skinner (1998) extended Rao's score test to survey data by developing an analogue, which they call the quasi-score test, that take account of survey design features. Their development of the quasi-score test and its properties is based on the sample estimating equations

in (3.2.5) which involve the basic design weights only, and parallels the development of Boos (1992) for the case of random sampling from an infinite population. Rao, Yung and Hidirolou (2002) discussed quasi-score tests when the calibration weights, $\tilde{w}_{hik}$, are used. Let $\hat{\mathbf{S}}_G(\boldsymbol{\theta}) = (\hat{\mathbf{S}}_{G1}(\boldsymbol{\theta})^T, \hat{\mathbf{S}}_{G2}(\boldsymbol{\theta})^T)^T$ be the partition of $\hat{\mathbf{S}}_G(\boldsymbol{\theta})$ corresponding to the partition of $\boldsymbol{\theta}$ where

$$\hat{\mathbf{S}}_{Gt}(\boldsymbol{\theta}) = \sum_s \tilde{w}_{hik} \mathbf{u}_{thik}(\boldsymbol{\theta}), \quad t = 1, 2 \quad (3.4.4)$$

and $\mathbf{u}_{hik}(\boldsymbol{\theta}) = (\mathbf{u}_{1hik}(\boldsymbol{\theta})^T, \mathbf{u}_{2hik}(\boldsymbol{\theta})^T)^T$. Denote the solution of $\hat{\mathbf{S}}_{G1}(\boldsymbol{\theta}') = \mathbf{0}$, where $\boldsymbol{\theta}' = (\boldsymbol{\theta}_1^T, \boldsymbol{\theta}_{20}^T)^T$, by $\tilde{\boldsymbol{\theta}}_G = (\tilde{\boldsymbol{\theta}}_{G1}^T, \boldsymbol{\theta}_{20}^T)^T$. The quasi-score test is based on the statistic

$$X_S^2 = \tilde{\mathbf{S}}_{G2} [\hat{\mathbf{V}}(\tilde{\mathbf{S}}_{G2})]^{-1} \tilde{\mathbf{S}}_{G2}, \quad (3.4.5)$$

where $\tilde{\mathbf{S}}_{G2} = \hat{\mathbf{S}}_{G2}(\tilde{\boldsymbol{\theta}}_G)$ and $\hat{\mathbf{V}}(\tilde{\mathbf{S}}_{G2})$ is a consistent estimator of $\mathbf{V}(\tilde{\mathbf{S}}_{G2})$. The asymptotic distribution of $\tilde{\mathbf{S}}_{G2}$ under H_0 can be obtained, as in Boos (1992), by expanding $\hat{\mathbf{S}}_{G1}(\tilde{\boldsymbol{\theta}}_G)$ and $\hat{\mathbf{S}}_{G2}(\tilde{\boldsymbol{\theta}}_G)$ around the true value $\boldsymbol{\theta}^* = (\boldsymbol{\theta}^{*T}, \boldsymbol{\theta}_{20}^T)^T$. That is,

$$\mathbf{0} = \hat{\mathbf{S}}_{G1}(\tilde{\boldsymbol{\theta}}_G) \approx \hat{\mathbf{S}}_{G1}(\boldsymbol{\theta}^*) - \hat{\mathbf{J}}_{G11}^*(\tilde{\boldsymbol{\theta}}_{G1} - \boldsymbol{\theta}_1^*) \quad (3.4.6)$$

$$\text{and } \hat{\mathbf{S}}_{G2}(\tilde{\boldsymbol{\theta}}_G) \approx \hat{\mathbf{S}}_{G2}(\boldsymbol{\theta}^*) - \hat{\mathbf{J}}_{G21}^*(\tilde{\boldsymbol{\theta}}_{G1} - \boldsymbol{\theta}_1^*), \quad (3.4.7)$$

where $\hat{\mathbf{J}}_G^* = \hat{\mathbf{J}}_G(\boldsymbol{\theta}^*)$, which is defined by (3.3.12) and evaluated at $\boldsymbol{\theta} = \boldsymbol{\theta}^*$, can be partitioned as

$$\hat{\mathbf{J}}_G^* = \begin{bmatrix} \hat{\mathbf{J}}_{G11}^* & \hat{\mathbf{J}}_{G12}^* \\ \hat{\mathbf{J}}_{G21}^* & \hat{\mathbf{J}}_{G22}^* \end{bmatrix}.$$

Then replacing $\hat{\mathbf{J}}_G^*$ by its expected value, $\mathbf{I}_G^* = \mathbf{E}[\hat{\mathbf{J}}_G(\boldsymbol{\theta}^*)]$ say, and substituting $\tilde{\boldsymbol{\theta}}_{G1} - \boldsymbol{\theta}_1^*$ from (3.4.6) into (3.4.7) yields

$$\tilde{\mathbf{S}}_{G2} = \hat{\mathbf{S}}_{G2}(\tilde{\boldsymbol{\theta}}_G) \approx \hat{\mathbf{S}}_{G2}(\boldsymbol{\theta}^*) - \mathbf{I}_{G21}^* \mathbf{I}_{G11}^{*-1} \hat{\mathbf{S}}_{G1}(\boldsymbol{\theta}^*)$$

$$= \sum_s \tilde{w}_{hik} \mathbf{u}_{2.1hik}(\boldsymbol{\theta}^*), \quad (3.4.8)$$

where $\mathbf{u}_{2.1hik}(\boldsymbol{\theta}^*) = \mathbf{u}_{2hik}(\boldsymbol{\theta}^*) - \mathbf{I}_{G21}^* \mathbf{I}_{G11}^{*-1} \mathbf{u}_{1hik}(\boldsymbol{\theta}^*)$. Using equation (3.4.8), we may write

$$\begin{aligned} \bar{\mathbf{T}} &= \left(\frac{n_e^{\frac{1}{2}}}{N} \right) \bar{\mathbf{S}}_{G2} \\ &\approx \left(\frac{n_e^{\frac{1}{2}}}{N} \right) \left[\sum_s \tilde{w}_{hik} \mathbf{u}_{2.1hik}(\boldsymbol{\theta}^*) - \sum_U \mathbf{u}_{2.1hik}(\boldsymbol{\theta}^*) \right] + \left(\frac{n_e}{N_e} \right)^{\frac{1}{2}} \left(\frac{N_e^{\frac{1}{2}}}{N} \right) \sum_U \mathbf{u}_{2.1hik}(\boldsymbol{\theta}^*) \\ &= \bar{\mathbf{T}}_1 + \left(\frac{n_e}{N_e} \right)^{\frac{1}{2}} \bar{\mathbf{T}}_{2N}, \end{aligned} \quad (3.4.9)$$

where $\bar{\mathbf{T}}_{2N}$ converges in model distribution to $N_R(\mathbf{0}, \bar{\mathbf{\Sigma}}_{22}^*)$, say, and $\bar{\mathbf{T}}_1$ converges in design distribution to $N_R(\mathbf{0}, \bar{\mathbf{\Sigma}}_{22N})$, say. Assuming that $f_e = 0$, then $n_e \mathbf{V}(\bar{\mathbf{S}}_{G2})$ is a consistent estimator of $\bar{\mathbf{\Sigma}}_{22N}$ (Rao, Yung and Hidirolou, 2002) and the quasi-score statistic X_S^2 is asymptotically χ_R^2 under H_0 .

Calculation of the quasi-score statistic X_S^2 requires an estimator of the covariance matrix $\mathbf{V}(\bar{\mathbf{S}}_{G2})$. We will present the Taylor linearization, linearized jackknife and estimating function jackknife variance estimators.

3.4.3 Taylor linearization variance estimator

A Taylor linearization estimator of $\mathbf{V}(\bar{\mathbf{S}}_{G2})$ can be obtained by applying the survey variance estimator for a total, defined in (3.3.14), to the representation in (3.4.8), replacing the R -dimensional vector $\mathbf{u}_{2.1hik}(\boldsymbol{\theta}^*)$ by

$$\bar{\mathbf{u}}_{2.1hik}(\bar{\boldsymbol{\theta}}_G) = \mathbf{u}_{2hik}(\bar{\boldsymbol{\theta}}_G) - \bar{\mathbf{D}}(\bar{\boldsymbol{\theta}}_G) \mathbf{u}_{1hik}(\bar{\boldsymbol{\theta}}_G) \quad (3.4.10)$$

where $\bar{\mathbf{D}}(\bar{\boldsymbol{\theta}}_G)$ is an estimator of $\mathbf{D}^* = \mathbf{I}_{G21}^* \mathbf{I}_{G11}^{*-1}$. A natural estimator of $\mathbf{D}^*$ is based on $\hat{\mathbf{J}}_G(\bar{\boldsymbol{\theta}}_G)$, but the resulting quasi-score test does not have the desired invariance

property in general (Boos, 1992). As suggested in Rao et. al. (2001), we can get an invariant test by taking the expectation of $\hat{\mathbf{J}}_G(\boldsymbol{\theta}^*)$ under the mean specification in (3.2.1), giving

$$\tilde{\mathbf{D}}(\tilde{\boldsymbol{\theta}}_G) = \sum_s \tilde{w}_{hik} \frac{\mathbf{d}_{hik}(\tilde{\boldsymbol{\theta}}_G) \mathbf{d}_{hik}(\tilde{\boldsymbol{\theta}}_G)^T}{V_{0hik}(\tilde{\mu}_{hik})} \quad (3.4.11)$$

as an estimator of $\mathbf{D}^*$, where

$$\mathbf{d}_{hik}(\boldsymbol{\theta}) = \frac{\partial \mu(\mathbf{z}_{hik}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$$

and $\tilde{\mu}_{hik} = \mu(\mathbf{z}_{hik}, \tilde{\boldsymbol{\theta}}_G)$. Since $\tilde{\mathbf{S}}_{G2}$ is approximated by a calibration estimator of a vector of totals, we can use the results in Section 3.3.1 to obtain a Taylor linearization variance estimator. Using the expression in (3.3.14) and $\tilde{\mathbf{u}}_{2.1hik}(\tilde{\boldsymbol{\theta}}_G)$, a Taylor linearization estimator of $\mathbf{V}(\tilde{\mathbf{S}}_{G2})$ is given by

$$\widehat{\mathbf{V}}_L(\tilde{\mathbf{S}}_{G2}) = v(\tilde{\mathbf{e}}_{2.1hi}(\tilde{\boldsymbol{\theta}}_G)) \quad (3.4.12)$$

where $\tilde{\mathbf{e}}_{2.1hi}(\tilde{\boldsymbol{\theta}}_G) = \sum_k (n_h \tilde{w}_{hik}) \tilde{\mathbf{e}}_{2.1hik}(\tilde{\boldsymbol{\theta}}_G)$, $\tilde{\mathbf{e}}_{2.1hik}(\tilde{\boldsymbol{\theta}}_G)$ is a R -dimensional vector with elements $\tilde{e}_{2.1hikq}(\tilde{\boldsymbol{\theta}}_G) = u_{2.1hikq}(\tilde{\boldsymbol{\theta}}_G) - \mathbf{x}_{hik}^T \tilde{\mathbf{B}}_q(\tilde{\boldsymbol{\theta}}_G)$ and

$$\begin{aligned} \tilde{\mathbf{B}}_q(\tilde{\boldsymbol{\theta}}_G) &= \left[\sum_s w_{hik} \mathbf{x}_{hik} \mathbf{x}_{hik}^T \right]^{-1} \left\{ \sum_s w_{hik} \mathbf{x}_{hik} u_{2.1hikq}(\tilde{\boldsymbol{\theta}}_G) \right\} \\ &= \widehat{\mathbf{A}}^{-1} \tilde{\mathbf{U}}_q(\tilde{\boldsymbol{\theta}}_G). \end{aligned} \quad (3.4.13)$$

3.4.4 Linearized jackknife variance estimator

To obtain a linearized jackknife estimator of $\mathbf{V}(\tilde{\mathbf{S}}_{G2})$ we use the results in Rao et. al. (1998) and Yung and Rao (1996). Replacing the basic design weights, w_{hik} , in (3.4.4) by the jackknife weights, $w_{hik(gj)}$, defined in Section 2.3.1 we get

$$\hat{\mathbf{S}}_{Gt(gj)}(\boldsymbol{\theta}) = \sum_s \tilde{w}_{hik(gj)} \mathbf{u}_{thik}(\boldsymbol{\theta}), \quad t = 1, 2$$

and jackknife estimates $\tilde{\boldsymbol{\theta}}_{G(gj)}$ and $\tilde{\mathbf{S}}_{Gt(gj)} = \hat{\mathbf{S}}_{Gt(gj)}(\tilde{\boldsymbol{\theta}}_{G(gj)})$. We then expand $\tilde{\mathbf{S}}_{G1(gj)}$ and $\tilde{\mathbf{S}}_{G2(gj)}$ around $\tilde{\boldsymbol{\theta}}_G$ to get

$$\mathbf{0} = \tilde{\mathbf{S}}_{G1(gj)} \approx \hat{\mathbf{S}}_{G1(gj)}(\tilde{\boldsymbol{\theta}}_G) - \bar{\mathbf{I}}_{G11}(\tilde{\boldsymbol{\theta}}_{G1(gj)} - \tilde{\boldsymbol{\theta}}_{G1}) \quad (3.4.14)$$

$$\text{and } \tilde{\mathbf{S}}_{G2(gj)} \approx \hat{\mathbf{S}}_{G2(gj)}(\tilde{\boldsymbol{\theta}}_G) - \bar{\mathbf{I}}_{G21}(\tilde{\boldsymbol{\theta}}_{G1(gj)} - \tilde{\boldsymbol{\theta}}_{G1}), \quad (3.4.15)$$

where $\bar{\mathbf{I}}_G = \mathbf{E}[\mathbf{J}_G(\tilde{\boldsymbol{\theta}}_G)]$. Substituting for $\tilde{\boldsymbol{\theta}}_{G1(gj)} - \tilde{\boldsymbol{\theta}}_{G1}$ from (3.4.14) into (3.4.15) yields

$$\tilde{\mathbf{S}}_{G2(gj)} \approx \sum_s \tilde{w}_{hik(gj)} \bar{\mathbf{u}}_{2.1hik}(\tilde{\boldsymbol{\theta}}_G) \quad (3.4.16)$$

assuming $\bar{\mathbf{D}}(\tilde{\boldsymbol{\theta}}_G) = \bar{\mathbf{I}}_{G21} \bar{\mathbf{I}}_{G11}^{-1}$. A customary jackknife estimator of $\mathbf{V}(\tilde{\mathbf{S}}_{G2})$ is given by

$$\widehat{\mathbf{V}}_J(\tilde{\mathbf{S}}_{G2}) = \sum_g \frac{n_g - 1}{n_g} \sum_j (\tilde{\mathbf{S}}_{G2(gj)} - \tilde{\mathbf{S}}_{G2}) (\tilde{\mathbf{S}}_{G2(gj)} - \tilde{\mathbf{S}}_{G2})^T. \quad (3.4.17)$$

Applying the techniques in Yung and Rao (1996) we can obtain a linearized jackknife variance estimator

$$\widehat{\mathbf{V}}_{JL}(\tilde{\mathbf{S}}_{G2}) = v(\bar{\mathbf{e}}_{2.1gj}(\tilde{\boldsymbol{\theta}}_G)), \quad (3.4.18)$$

which is identical to the Taylor linearization variance estimator in (3.4.12).

3.4.5 Estimating function jackknife variance estimator

To apply the EF jackknife method we partition the EF jackknife calibration estimator of $\mathbf{S}(\boldsymbol{\theta})$, given in (3.3.30), as

$$\check{\mathbf{S}}_{Gt(gj)}(\boldsymbol{\theta}) = \sum_s \check{w}_{hik(gj)} \mathbf{u}_{thik}(\boldsymbol{\theta}), \quad t = 1, 2. \quad (3.4.19)$$

For $\boldsymbol{\theta}' = (\boldsymbol{\theta}_1^T, \boldsymbol{\theta}_{20}^T)^T$ we solve $\check{\mathbf{S}}_{G1(gj)}(\boldsymbol{\theta}') = \mathbf{0}$ and obtain $\check{\boldsymbol{\theta}}_{G1(gj)}$. Hence an EF jackknife estimator of $\boldsymbol{\theta}_N$ is given by

$$\check{\boldsymbol{\theta}}_{G(gj)} = \begin{pmatrix} \check{\boldsymbol{\theta}}_{G1(gj)} \\ \boldsymbol{\theta}_{20} \end{pmatrix}.$$

We will obtain an EF jackknife estimator of $\mathbf{V}(\tilde{\mathbf{S}}_{G2})$ by expanding $\check{\mathbf{S}}_{G1(gj)}(\check{\boldsymbol{\theta}}_{G(gj)})$ and $\check{\mathbf{S}}_{G2(gj)}(\check{\boldsymbol{\theta}}_{G(gj)})$ around $\tilde{\boldsymbol{\theta}}_G = (\tilde{\boldsymbol{\theta}}_{G1}^T, \boldsymbol{\theta}_{20}^T)^T$. That is,

$$\mathbf{0} = \check{\mathbf{S}}_{G1(gj)}(\check{\boldsymbol{\theta}}_{G(gj)}) \approx \check{\mathbf{S}}_{G1(gj)}(\tilde{\boldsymbol{\theta}}_G) - \check{\mathbf{J}}_{G1(gj)}(\tilde{\boldsymbol{\theta}}_G)(\check{\boldsymbol{\theta}}_{G1(gj)} - \tilde{\boldsymbol{\theta}}_{G1}) \quad (3.4.20)$$

where

$$\check{\mathbf{J}}_{G1(gj)}(\tilde{\boldsymbol{\theta}}_G) = - \left. \frac{\partial \check{\mathbf{S}}_{G1(gj)}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_1^T} \right|_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}_G}.$$

Using the definition of the EF jackknife calibration weights in (2.4.5) and the fact that $\hat{\mathbf{S}}_{G1}(\tilde{\boldsymbol{\theta}}_G) = \mathbf{0}$ and $\sum_s w_{hik} \mathbf{x}_{hik} e_{hikq}(\tilde{\boldsymbol{\theta}}_G) = \mathbf{0}$, where $e_{hikq}(\tilde{\boldsymbol{\theta}}_G) = (u_{hikq}(\tilde{\boldsymbol{\theta}}_G) - \mathbf{x}_{hik}^T \hat{\mathbf{B}}_q(\tilde{\boldsymbol{\theta}}_G))$, $q = 1, \dots, Q$, and

$$\hat{\mathbf{B}}_q(\tilde{\boldsymbol{\theta}}_G) = \left[\sum_s w_{hik} \mathbf{x}_{hik} \mathbf{x}_{hik}^T \right]^{-1} \left\{ \sum_s w_{hik} \mathbf{x}_{hik} u_{hikq}(\tilde{\boldsymbol{\theta}}_G) \right\},$$

we can show that

$$\begin{aligned} \check{\mathbf{S}}_{G1(gj)}(\tilde{\boldsymbol{\theta}}_G) &= \sum_s \check{w}_{hik(gj)} u_{1hik}(\tilde{\boldsymbol{\theta}}_G) \\ &= \hat{\mathbf{S}}_{G1}(\tilde{\boldsymbol{\theta}}_G) - \sum_s w_{hik(gj)} g_{hik}(\mathbf{u}_{1hik}(\tilde{\boldsymbol{\theta}}_G) - \mathbf{x}_{hik}^T \hat{\mathbf{B}}_1(\tilde{\boldsymbol{\theta}}_G)) \\ &= - \sum_s w_{hik(gj)} g_{hik} \mathbf{e}_{1hik}(\tilde{\boldsymbol{\theta}}_G) \\ &= - \sum_s g_{hik} \mathbf{e}_{1hik}(\tilde{\boldsymbol{\theta}}_G) \{w_{hik(gj)} - w_{hik}\} \end{aligned}$$

where $\hat{\mathbf{B}}(\tilde{\boldsymbol{\theta}}_G)$ is partitioned as $\hat{\mathbf{B}}(\tilde{\boldsymbol{\theta}}_G) = (\hat{\mathbf{B}}_1^T(\tilde{\boldsymbol{\theta}}_G), \hat{\mathbf{B}}_2^T(\tilde{\boldsymbol{\theta}}_G))^T$. From the results in (2.4.8) we get

$$\check{\mathbf{S}}_{G1(gj)}(\tilde{\boldsymbol{\theta}}_G) = \frac{1}{n_g - 1} (\bar{\mathbf{e}}_{1gj}(\tilde{\boldsymbol{\theta}}_G) - \bar{\mathbf{e}}_{1g}(\tilde{\boldsymbol{\theta}}_G)). \quad (3.4.21)$$

Similarly, we expand $\check{\mathbf{S}}_{G2(gj)}(\check{\boldsymbol{\theta}}_{G(gj)})$ around $\tilde{\boldsymbol{\theta}}_G$ and obtain

$$\begin{aligned} &\check{\mathbf{S}}_{G2(gj)}(\check{\boldsymbol{\theta}}_{G(gj)}) \\ &\approx \check{\mathbf{S}}_{G2(gj)}(\tilde{\boldsymbol{\theta}}_G) - \check{\mathbf{J}}_{G2(gj)}(\tilde{\boldsymbol{\theta}}_G)(\check{\boldsymbol{\theta}}_{G1(gj)} - \tilde{\boldsymbol{\theta}}_{G1}) \\ &= \check{\mathbf{S}}_{G2(gj)}(\tilde{\boldsymbol{\theta}}_G) - \check{\mathbf{J}}_{G2(gj)}(\tilde{\boldsymbol{\theta}}_G) \check{\mathbf{J}}_{G1(gj)}^{-1}(\tilde{\boldsymbol{\theta}}_G) \check{\mathbf{S}}_{G1(gj)}(\tilde{\boldsymbol{\theta}}_G). \end{aligned} \quad (3.4.22)$$

From Section 3.3.3 we can see that $\check{\mathbf{J}}_{G21(gj)}(\bar{\boldsymbol{\theta}}_G) \approx \hat{\mathbf{J}}_{G21}(\bar{\boldsymbol{\theta}}_G)$ and $\check{\mathbf{J}}_{G11(gj)}^{-1}(\bar{\boldsymbol{\theta}}_G) \approx \hat{\mathbf{J}}_{G11}^{-1}(\bar{\boldsymbol{\theta}}_G)$. If we replace $\hat{\mathbf{J}}_G(\bar{\boldsymbol{\theta}}_G)$ by $\bar{\mathbf{I}}_G = \mathbf{E}[\hat{\mathbf{J}}_G(\bar{\boldsymbol{\theta}}_G)]$ and let $\bar{\mathbf{D}}(\bar{\boldsymbol{\theta}}_G) = \bar{\mathbf{I}}_{G21} \bar{\mathbf{I}}_{G11}^{-1}$ then combining (3.4.21) and (3.4.22) gives

$$\check{\mathbf{S}}_{G2(gj)}(\check{\boldsymbol{\theta}}_{G(gj)}) \approx \check{\mathbf{S}}_{G2(gj)}(\bar{\boldsymbol{\theta}}_G) - \frac{1}{n_g - 1} \bar{\mathbf{D}}(\bar{\boldsymbol{\theta}}_G) (\bar{\mathbf{e}}_{1gj}(\bar{\boldsymbol{\theta}}_G) - \bar{\bar{\mathbf{e}}}_{1g}(\bar{\boldsymbol{\theta}}_G)). \quad (3.4.23)$$

But

$$\begin{aligned} \check{\mathbf{S}}_{G2(gj)}(\bar{\boldsymbol{\theta}}_G) &= \hat{\mathbf{S}}_{G2}(\bar{\boldsymbol{\theta}}_G) - \sum_s w_{hik(gj)} g_{hik} (\mathbf{u}_{2hik}(\bar{\boldsymbol{\theta}}_G) - \mathbf{x}_{hik}^T \hat{\mathbf{B}}_2(\bar{\boldsymbol{\theta}}_G)) \\ &= \hat{\mathbf{S}}_{G2}(\bar{\boldsymbol{\theta}}_G) - \sum_s w_{hik(gj)} g_{hik} \mathbf{e}_{2hik}(\bar{\boldsymbol{\theta}}_G) \\ &= \hat{\mathbf{S}}_{G2}(\bar{\boldsymbol{\theta}}_G) - \sum_s g_{hik} \mathbf{e}_{2hik}(\bar{\boldsymbol{\theta}}_G) \{w_{hik(gj)} - w_{hik}\} \\ &= \hat{\mathbf{S}}_{G2}(\bar{\boldsymbol{\theta}}_G) + \frac{1}{n_g - 1} (\bar{\mathbf{e}}_{2gj}(\bar{\boldsymbol{\theta}}_G) - \bar{\bar{\mathbf{e}}}_{2g}(\bar{\boldsymbol{\theta}}_G)). \end{aligned} \quad (3.4.24)$$

Putting (3.4.24) into (3.4.23) we finally have

$$\check{\mathbf{S}}_{G2(gj)}(\check{\boldsymbol{\theta}}_{G(gj)}) - \hat{\mathbf{S}}_{G2}(\bar{\boldsymbol{\theta}}_G) \approx \frac{1}{n_g - 1} (\bar{\mathbf{e}}_{2.1gj}(\bar{\boldsymbol{\theta}}_G) - \bar{\bar{\mathbf{e}}}_{2.1g}(\bar{\boldsymbol{\theta}}_G)) \quad (3.4.25)$$

and an EF jackknife estimator of $V(\tilde{\mathbf{S}}_{G2})$ is given by

$$\begin{aligned} \widehat{\mathbf{V}}_{EFJ}(\tilde{\mathbf{S}}_{G2}) &= \sum_g \frac{n_g - 1}{n_g} \sum_j (\check{\mathbf{S}}_{G2(gj)}(\check{\boldsymbol{\theta}}_{G(gj)}) - \tilde{\mathbf{S}}_{G2}) (\check{\mathbf{S}}_{G2(gj)}(\check{\boldsymbol{\theta}}_{G(gj)}) - \tilde{\mathbf{S}}_{G2})^T \\ &\approx v(\bar{\mathbf{e}}_{2.1gj}(\bar{\boldsymbol{\theta}}_G)) \end{aligned}$$

which is approximately equal to $\widehat{\mathbf{V}}_L(\tilde{\mathbf{S}}_{G2})$ and $\widehat{\mathbf{V}}_{JL}(\tilde{\mathbf{S}}_{G2})$.

Chapter 4

Variance Estimation Under Unit Nonresponse

4.1 Introduction

The results in Chapters 2 and 3 assumed that all sampled elements responded and fully completed the survey questionnaires. This situation is commonly referred to as the full response case. In practice though, the full response case is rarely achieved as most surveys suffer from both unit and item nonresponse. Unit nonresponse occurs when no information is obtained from the sampled element, while item nonresponse occurs when partial information is obtained.

In this chapter we consider variance estimation in the presence of unit nonresponse with multiple weighting classes. Section 4.2 presents standard jackknife and EF jackknife variance estimators for the calibration estimator of a total. These methods are then extended in Section 4.3 to cover parameters obtained as solutions of estimating

equations.

4.2 Calibration Estimator of a Total

Compensation for unit nonresponse can be handled through the use of weighting adjustments where the weights of the respondents are increased so that they represent the nonrespondents. Weighting classes are usually defined based on characteristics known for all sample units and weighting adjustments are performed within each class. It is hoped that respondents and nonrespondents in the same weighting class are similar.

In the presence of unit nonresponse, we divide the sample from a stratified multistage sampling design into respondents, s_r , and nonrespondents, s_m . Let a_{hik} be a response indicator variable defined as

$$a_{hik} = \begin{cases} 1 & \text{if element } (hik) \text{ responds} \\ 0 & \text{otherwise.} \end{cases}$$

We assume that we can form $M \geq 2$ weighting classes based on characteristics known for all units in the selected sample. The probability of response, $p_m > 0$ for $m = 1, \dots, M$, is assumed to be uniform within each weighting class but different between the weighting classes. Then the nonresponse adjustment factor in the m^{th} weighting class is given by

$$d_m = \frac{\sum_s w_{hik} \eta_{hik}^m}{\sum_s w_{hik} a_{hik} \eta_{hik}^m} = \frac{\widehat{W}_m}{\widehat{R}_m}, \quad (4.2.1)$$

where $\widehat{W}_m = \sum_s w_{hik} \eta_{hik}^m$, $\widehat{R}_m = \sum_s w_{hik} a_{hik} \eta_{hik}^m$ and η_{hik}^m is the weighting class indi-

cator variable defined by

$$\eta_{hik}^m = \begin{cases} 1 & \text{if element } (hik) \text{ is in weighting class } m \\ 0 & \text{otherwise.} \end{cases}$$

A basic estimator of the population total Y , in the presence of unit nonresponse with M weighting classes, is

$$\hat{Y}_M = \sum_{m=1}^M \sum_s d_m w_{hik} a_{hik} \eta_{hik}^m y_{hik}. \quad (4.2.2)$$

Using the projection estimator (Sarndal et. al., 1989), we define the calibration adjustment factor by

$$g_{hik} = \mathbf{X}^T \tilde{\mathbf{A}}_M^{-1} \mathbf{x}_{hik} \quad (4.2.3)$$

where $\mathbf{x}_{hik}$ is a P -dimensional vector of indicator auxiliary variables, $\mathbf{X}$ is the vector of known population totals and $\tilde{\mathbf{A}}_M = \sum_{m=1}^M \sum_s d_m w_{hik} a_{hik} \eta_{hik}^m \mathbf{x}_{hik} \mathbf{x}_{hik}^T$. Then the calibration estimator of Y , in the presence of unit nonresponse with M weighting classes, is

$$\hat{Y}_{MG} = \mathbf{X}^T \tilde{\mathbf{B}}_M = \sum_{m=1}^M \sum_s d_m \tilde{w}_{hik} a_{hik} \eta_{hik}^m y_{hik}, \quad (4.2.4)$$

where $\tilde{\mathbf{B}}_M = \tilde{\mathbf{A}}_M^{-1} \tilde{\mathbf{U}}_M$, $\tilde{\mathbf{U}}_M = \sum_{m=1}^M \sum_s d_m w_{hik} a_{hik} \eta_{hik}^m \mathbf{x}_{hik} y_{hik}$.

We note that the calibration estimator in (4.2.4) is based on the responding units only and it will benchmark to the known population totals since $\widehat{\mathbf{X}}_{MG} = \mathbf{X}$. We are assuming here that the nonresponse adjustment is performed first followed by the calibration adjustment. This is necessary because when the calibration adjustment is performed first, the resulting estimator will not benchmark to the known population totals. Under a uniform response mechanism within each weighting class, we can

show that $\hat{Y}_{MG}$ is approximately unbiased for Y since

$$E(\hat{Y}_{MG}) \approx \sum_{m=1}^M \sum_s \frac{1}{p_m} p_m E(\tilde{w}_{hik} y_{hik}) = E(\hat{Y}_G) \approx Y.$$

4.2.1 Linearized Jackknife

To construct a jackknife variance estimator of $\hat{Y}_{MG}$, in the presence of unit nonresponse with M weighting classes, we first define the nonresponse adjustment within the m^{th} weighting class, when the $(gj)^{\text{th}}$ cluster is deleted, by

$$d_{m(gj)} = \frac{\sum_s w_{hik(gj)} \eta_{hik}^m}{\sum_s w_{hik(gj)} a_{hik} \eta_{hik}^m} = \frac{\widehat{W}_{m(gj)}}{\widehat{R}_{m(gj)}}. \quad (4.2.5)$$

A jackknife calibration estimator of the population total, Y , when the $(gj)^{\text{th}}$ cluster is deleted, is defined by

$$\hat{Y}_{MG(gj)} = \mathbf{X}^T \bar{\mathbf{B}}_{M(gj)} = \sum_{m=1}^M \sum_s d_{m(gj)} \tilde{w}_{hik(gj)} a_{hik} \eta_{hik}^m y_{hik}, \quad (4.2.6)$$

where

$$\begin{aligned} \bar{\mathbf{B}}_{M(gj)} &= \bar{\mathbf{A}}_{M(gj)}^{-1} \bar{\mathbf{U}}_{M(gj)}, \\ \bar{\mathbf{A}}_{M(gj)} &= \sum_{m=1}^M \sum_s d_{m(gj)} w_{hik(gj)} a_{hik} \eta_{hik}^m \mathbf{x}_{hik} \mathbf{x}_{hik}^T, \\ \bar{\mathbf{U}}_{M(gj)} &= \sum_{m=1}^M \sum_s d_{m(gj)} w_{hik(gj)} a_{hik} \eta_{hik}^m \mathbf{x}_{hik} y_{hik}, \end{aligned}$$

$\tilde{w}_{hik(gj)} = w_{hik(gj)} g_{hik(gj)}$ and $g_{hik(gj)} = \mathbf{X}^T \bar{\mathbf{A}}_{M(gj)}^{-1} \mathbf{x}_{hik}$. Then a customary jackknife variance estimator of $\hat{Y}_{MG}$ is given by

$$v_J(\hat{Y}_{MG}) = \sum_{g=1}^L \frac{n_g - 1}{n_g} \sum_{j=1}^{n_g} (\hat{Y}_{MG(gj)} - \hat{Y}_{MG})^2, \quad (4.2.7)$$

which can be linearized (Yung and Rao, 2000) to obtain a linearized jackknife variance estimator

$$v_{JL}(\hat{Y}_{MG}) = \sum_{g=1}^L n_g^{-1} (n_g - 1)^{-1} \sum_{j=1}^{n_g} (\bar{\tilde{r}}_g - \bar{r}_{gj})^2 = v(\bar{r}_{gj}) \quad (4.2.8)$$

where

$$\begin{aligned} \bar{r}_{gj} &= \sum_{m=1}^M d_m \sum_k (n_g \tilde{w}_{gjk}) a_{gjk} \eta_{gjk}^m e_{gjk} \\ &\quad + \sum_{m=1}^M \hat{\tau}_m \hat{R}_m^{-1} \sum_k (1 - d_m a_{gjk}) (n_g w_{gjk}) \eta_{gjk}^m, \end{aligned} \quad (4.2.9)$$

$\hat{\tau}_m = \sum_s \tilde{w}_{gjk} a_{gjk} \eta_{gjk}^m e_{gjk}$ and $e_{gjk} = y_{gjk} - \mathbf{x}_{gjk}^T \tilde{\mathbf{B}}_M$. We note that $v_{JL}(\hat{Y}_{MG})$ depends on the cluster totals of the estimated residuals and an additional term for the nonresponse adjustment.

4.2.2 Estimating Function Jackknife

In the case of unit nonresponse with M weighting classes, we note that $\tilde{\mathbf{B}}_M$ is the solution of $\hat{\mathbf{S}}_M(\mathbf{B}) = \mathbf{0}$, where

$$\hat{\mathbf{S}}_M(\mathbf{B}) = \sum_{m=1}^M \sum_s d_m w_{hik} a_{hik} \eta_{hik}^m \mathbf{x}_{hik} (y_{hik} - \mathbf{x}_{hik}^T \mathbf{B}) \quad (4.2.10)$$

is the sample estimating function of $\mathbf{S}(\mathbf{B})$. To apply the EF jackknife method, define

$$\begin{aligned} \hat{\mathbf{S}}_{M(gj)}(\tilde{\mathbf{B}}_M) &= \sum_{m=1}^M \sum_s d_{m(gj)} w_{hik(gj)} a_{hik} \eta_{hik}^m \mathbf{x}_{hik} (y_{hik} - \mathbf{x}_{hik}^T \tilde{\mathbf{B}}_M) \\ &= \sum_{m=1}^M \sum_s d_{m(gj)} w_{hik(gj)} a_{hik} \eta_{hik}^m \mathbf{x}_{hik} e_{hik} \end{aligned} \quad (4.2.11)$$

with residuals $e_{hik} = y_{hik} - \mathbf{x}_{hik}^T \tilde{\mathbf{B}}_M$. Let $d_{m(gj)} = f(\hat{W}_{m(gj)}, \hat{R}_{m(gj)}) = \hat{W}_{m(gj)} / \hat{R}_{m(gj)}$

be the nonresponse adjustment in the m^{th} weighting class when the $(gj)^{\text{th}}$ cluster

is deleted. Expanding $d_{m(gj)}$ around the full sample estimates $(\widehat{W}_m, \widehat{R}_m)$, Valliant (1993) obtained the following approximation

$$\begin{aligned} \frac{\widehat{W}_{m(gj)}}{\widehat{R}_{m(gj)}} &\approx \frac{\widehat{W}_m}{\widehat{R}_m} + \frac{1}{\widehat{R}_m} (\widehat{W}_{m(gj)} - \widehat{W}_m) - \frac{\widehat{W}_m}{\widehat{R}_m^2} (\widehat{R}_{m(gj)} - \widehat{R}_m) \\ &= d_m + \widehat{R}_m^{-1} \{ (\widehat{W}_{m(gj)} - \widehat{W}_m) - d_m (\widehat{R}_{m(gj)} - \widehat{R}_m) \} \\ &= d_m + \Delta_{gjm}, \end{aligned} \quad (4.2.12)$$

where $\Delta_{gjm} = \widehat{R}_m^{-1} \{ (\widehat{W}_{m(gj)} - \widehat{W}_m) - d_m (\widehat{R}_{m(gj)} - \widehat{R}_m) \}$ depends on the weighting class m . An EF jackknife estimator of $\mathbf{B}$, obtained by solving $\widehat{\mathbf{S}}_M(\mathbf{B}) = \widehat{\mathbf{S}}_{M(gj)}(\tilde{\mathbf{B}}_M)$, is

$$\begin{aligned} \ddot{\mathbf{B}}_{M(gj)} &= \tilde{\mathbf{B}}_M - \tilde{\mathbf{A}}_M^{-1} \widehat{\mathbf{S}}_{M(gj)}(\tilde{\mathbf{B}}_M) \\ &\approx \tilde{\mathbf{B}}_M - \tilde{\mathbf{A}}_M^{-1} \sum_{m=1}^M \sum_s (d_m + \Delta_{gjm}) w_{hik(gj)} a_{hik} \eta_{hik}^m \mathbf{x}_{hik} e_{hik}. \end{aligned} \quad (4.2.13)$$

Hence, a unit nonresponse adjusted EF jackknife calibration estimator of the population total Y is defined as

$$\ddot{Y}_{MG(gj)} = \mathbf{X}^T \ddot{\mathbf{B}}_{M(gj)} = \sum_{m=1}^M \sum_s \ddot{w}_{hik(gj)}^m a_{hik} \eta_{hik}^m y_{hik}, \quad (4.2.14)$$

where the EF jackknife calibration weights are defined by

$$\begin{aligned} \ddot{w}_{hik(gj)}^m &= d_m \bar{w}_{hik} - d_{m(gj)} w_{hik(gj)} g_{hik} \\ &+ d_m w_{hik} \left(\sum_{m=1}^M \sum_s d_{m(gj)} w_{hik(gj)} g_{hik} a_{hik} \eta_{hik}^m \mathbf{x}_{hik}^T \right) \tilde{\mathbf{A}}_M^{-1} \mathbf{x}_{hik}. \end{aligned} \quad (4.2.15)$$

Putting the approximation in equation (4.2.13) into (4.2.14) we have

$$\begin{aligned} \ddot{Y}_{MG(gj)} - \widehat{Y}_{MG} &\approx -\mathbf{X}^T \tilde{\mathbf{A}}_M^{-1} \sum_{m=1}^M \sum_s (d_m + \Delta_{gjm}) w_{hik(gj)} a_{hik} \eta_{hik}^m \mathbf{x}_{hik} e_{hik} \\ &= - \left\{ \sum_{m=1}^M \sum_s d_m w_{hik(gj)} g_{hik} a_{hik} \eta_{hik}^m e_{hik} \right. \\ &\quad \left. + \sum_{m=1}^M \sum_s \Delta_{gjm} w_{hik(gj)} g_{hik} a_{hik} \eta_{hik}^m e_{hik} \right\} \end{aligned} \quad (4.2.16)$$

and an EF jackknife variance estimator of $\hat{Y}_{MG}$, in the presence of unit nonresponse with M weighting classes, is

$$v_{EFJ}(\hat{Y}_{MG}) = \sum_{g=1}^L \frac{n_g - 1}{n_g} \sum_{j=1}^{n_g} \left(\ddot{Y}_{MG(gj)} - \hat{Y}_{MG} \right)^2. \quad (4.2.17)$$

Working with the first term of (4.2.16) we can show, using the steps in (2.4.8), that

$$\begin{aligned} & - \sum_{m=1}^M \sum_s d_m w_{hik(gj)} g_{hik} a_{hik} \eta_{hik}^m e_{hik} \\ &= - \sum_{m=1}^M \sum_s d_m g_{hik} a_{hik} \eta_{hik}^m e_{hik} (w_{hik(gj)} - w_{hik}) \\ &= \frac{1}{n_g - 1} (\bar{e}_{gj} - \bar{e}_g), \end{aligned} \quad (4.2.18)$$

where $\bar{e}_{gj} = \sum_{m=1}^M d_m \sum_k (n_g \bar{w}_{gjk}) a_{gjk} \eta_{gjk}^m e_{gjk}$ and $\bar{e}_g = n_g^{-1} \sum_i \bar{e}_{gim}$, using the fact that $\sum_{m=1}^M \sum_s d_m w_{hik} a_{hik} \eta_{hik}^m x_{hik} e_{hik} = 0$. Ignoring the lower order terms, the second term of (4.2.16) can be written as

$$\begin{aligned} - \sum_{m=1}^M \sum_s \Delta_{gjm} w_{hik(gj)} g_{hik} a_{hik} \eta_{hik}^m e_{hik} &\approx - \sum_{m=1}^M \sum_s \Delta_{gjm} w_{hik} g_{hik} a_{hik} \eta_{hik}^m e_{hik} \\ &= - \sum_{m=1}^M \Delta_{gjm} \left(\sum_s \bar{w}_{hik} a_{hik} \eta_{hik}^m e_{hik} \right) \\ &= - \sum_{m=1}^M \Delta_{gjm} \hat{\tau}_m \end{aligned} \quad (4.2.19)$$

where $\hat{\tau}_m = \sum_s \bar{w}_{hik} a_{hik} \eta_{hik}^m e_{hik}$. Now

$$\begin{aligned} \Delta_{gjm} &= \hat{R}_m^{-1} \left\{ (\hat{W}_{m(gj)} - \hat{W}_m) - d_m (\hat{R}_{m(gj)} - \hat{R}_m) \right\} \\ &= \sum_s \hat{R}_m^{-1} \eta_{hik}^m (1 - d_m a_{hik}) (w_{hik(gj)} - w_{hik}) \\ &= - \frac{1}{n_g - 1} \left\{ \sum_k \hat{R}_m^{-1} (n_g w_{gjk}) (1 - d_m a_{gjk}) \eta_{gjk}^m \right. \\ &\quad \left. - n_g^{-1} \sum_i \sum_k \hat{R}_m^{-1} (n_g w_{gik}) (1 - d_m a_{gik}) \eta_{gik}^m \right\} \\ &= - \frac{1}{n_g - 1} (\gamma_{gjm} - \bar{\gamma}_{gm}) \end{aligned} \quad (4.2.20)$$

where $\gamma_{gjm} = \sum_k \hat{R}_m^{-1} (n_g w_{gjk}) (1 - d_m a_{gjk}) \eta_{gjk}^m$ and $\bar{\gamma}_{gm} = n_g^{-1} \sum_i \gamma_{gim}$. Substituting equations (4.2.18) - (4.2.20) into (4.2.17), we then have

$$v_{EFJ}(\hat{Y}_{MG}) \approx \sum_g n_g^{-1} (n_g - 1)^{-1} \sum_j (\bar{r}_{gj} - \bar{\bar{r}}_g)^2 = v(\bar{r}_{gj}) \quad (4.2.21)$$

which is approximately equal to $v_{JL}(\hat{Y}_{MG})$. If poststrata are used as weighting classes then $\hat{\tau}_m = 0$ and $\bar{r}_{gj}$ in (4.2.9) reduces to

$$\bar{e}_{gj} = \sum_k d (n_g \tilde{w}_{gjk}) a_{gjk} e_{gjk}, \quad (4.2.22)$$

where

$$\begin{aligned} d &= \frac{\sum_s w_{hik}}{\sum_s w_{hik} a_{hik}}, \\ e_{gjk} &= y_{gjk} - \mathbf{x}_{gjk}^T \tilde{\mathbf{B}}, \\ \tilde{\mathbf{B}} &= \bar{\mathbf{A}}^{-1} \tilde{\mathbf{U}}, \\ \bar{\mathbf{A}} &= \sum_s d w_{hik} a_{hik} \mathbf{x}_{hik} \mathbf{x}_{hik}^T, \\ \text{and } \tilde{\mathbf{U}} &= \sum_s d w_{hik} a_{hik} \mathbf{x}_{hik} y_{hik}. \end{aligned}$$

4.3 Calibration Estimator of θ

We will now extend the results in Section 4.2 to cover parameters defined as solutions of census estimating equations, that are discussed in Chapter 3 for the full response case. We will consider the case of multiple weighting classes.

Suppose we have $M \geq 2$ weighting classes available from characteristics known for all units in the selected sample. Define the calibration estimator of the census

estimating equation $\mathbf{S}(\boldsymbol{\theta}) = \sum_U \mathbf{u}_{hik}(\boldsymbol{\theta})$, in the presence of unit nonresponse, as

$$\hat{\mathbf{S}}_{MG}(\boldsymbol{\theta}) = \sum_{m=1}^M \sum_s d_m \tilde{w}_{hik} a_{hik} \eta_{hik}^m \mathbf{u}_{hik}(\boldsymbol{\theta}) \quad (4.3.1)$$

where d_m is the nonresponse adjustment factor in the m^{th} weighting class, which is identified by the indicator variable η_{hik}^m . Let $\hat{\boldsymbol{\theta}}_{MG}$ be the solution of $\hat{\mathbf{S}}_{MG}(\boldsymbol{\theta}) = \mathbf{0}$. We will now present two variance estimators based on the jackknife method.

4.3.1 Linearized Jackknife

When the $(gj)^{\text{th}}$ first-stage cluster is deleted, let

$$\hat{\mathbf{S}}_{MG(gj)}(\boldsymbol{\theta}) = \sum_{m=1}^M \sum_s d_{m(gj)} \tilde{w}_{hik(gj)} a_{hik} \eta_{hik}^m \mathbf{u}_{hik}(\boldsymbol{\theta}) \quad (4.3.2)$$

where $\tilde{w}_{hik(gj)}$ are the unit nonresponse adjusted jackknife calibration weights defined in Subsection 4.2.1. Let $\hat{\boldsymbol{\theta}}_{MG(gj)}$ be the solution of $\hat{\mathbf{S}}_{MG(gj)}(\boldsymbol{\theta}) = \mathbf{0}$ and the usual jackknife variance estimator of $\hat{\boldsymbol{\theta}}_{MG}$ can be expressed as

$$\hat{\mathbf{V}}_J(\hat{\boldsymbol{\theta}}_{MG}) = \sum_g \frac{n_g - 1}{n_g} \sum_j \left(\hat{\boldsymbol{\theta}}_{MG(gj)} - \hat{\boldsymbol{\theta}}_{MG} \right) \left(\hat{\boldsymbol{\theta}}_{MG(gj)} - \hat{\boldsymbol{\theta}}_{MG} \right)^T. \quad (4.3.3)$$

From Yung, Hidioglou and Rao (2001), one can obtain a linearized jackknife variance estimator as

$$\hat{\mathbf{V}}_{JL}(\hat{\boldsymbol{\theta}}_{MG}) = v \left[\hat{\mathbf{J}}_{MG}^{-1}(\hat{\boldsymbol{\theta}}_{MG}) \left\{ \tilde{\mathbf{e}}_{gj}(\hat{\boldsymbol{\theta}}_{MG}) + \sum_{m=1}^M \gamma_{gjm} \hat{\boldsymbol{\tau}}_m(\hat{\boldsymbol{\theta}}_{MG}) \right\} \right], \quad (4.3.4)$$

where

$$\hat{\mathbf{J}}_{MG}(\hat{\boldsymbol{\theta}}_{MG}) = - \left. \frac{\partial \hat{\mathbf{S}}_{MG}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T} \right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_{MG}} = \sum_{m=1}^M \sum_{(hik) \in s} d_m \tilde{w}_{hik} a_{hik} \eta_{hik}^m \left. \frac{-\partial \mathbf{u}_{hik}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T} \right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_{MG}} \quad (4.3.5)$$

and $\bar{\mathbf{e}}_{gj}(\hat{\boldsymbol{\theta}}_{MG}) = \sum_{m=1}^M \sum_k d_m (n_g \tilde{w}_{gjk}) a_{gjk} \eta_{gjk}^m \mathbf{e}_{gjk}(\hat{\boldsymbol{\theta}}_{MG})$. The q^{th} element of $\mathbf{e}_{hik}(\hat{\boldsymbol{\theta}}_{MG})$ is given by

$$e_{hikq}(\hat{\boldsymbol{\theta}}_{MG}) = u_{hikq}(\hat{\boldsymbol{\theta}}_{MG}) - \mathbf{x}_{hik}^T \bar{\mathbf{B}}_q(\hat{\boldsymbol{\theta}}_{MG}), \quad q = 1, \dots, Q,$$

where

$$\begin{aligned} \bar{\mathbf{B}}_q(\hat{\boldsymbol{\theta}}_{MG}) &= \left[\sum_{m=1}^M \sum_s d_m w_{hik} a_{hik} \eta_{hik}^m \mathbf{x}_{hik} \mathbf{x}_{hik}^T \right]^{-1} \left\{ \sum_{m=1}^M \sum_s d_m w_{hik} a_{hik} \eta_{hik}^m \mathbf{x}_{hik} u_{hikq}(\hat{\boldsymbol{\theta}}_{MG}) \right\} \\ &= \tilde{\mathbf{A}}_M^{-1} \tilde{\mathbf{U}}_{qM}(\hat{\boldsymbol{\theta}}_{MG}) \end{aligned} \quad (4.3.6)$$

and

$$\begin{aligned} \hat{\tau}_m(\hat{\boldsymbol{\theta}}_{MG}) &= \sum_s \tilde{w}_{gjk} a_{gjk} \eta_{gjk}^m \mathbf{e}_{gjk}(\hat{\boldsymbol{\theta}}_{MG}), \\ \gamma_{gjm} &= \sum_k \hat{R}_m^{-1}(n_g w_{gjk}) (1 - d_m a_{gjk}) \eta_{gjk}^m. \end{aligned}$$

4.3.2 Estimating Function Jackknife

Using the EF jackknife calibration weights in (4.2.15) we define the EF jackknife calibration estimator of $\mathbf{S}(\boldsymbol{\theta})$ by

$$\ddot{\mathbf{S}}_{MG(gj)}(\boldsymbol{\theta}) = \sum_{m=1}^M \sum_s \ddot{w}_{hik(gj)}^m a_{hik} \eta_{hik}^m \mathbf{u}_{hik}(\boldsymbol{\theta}) \quad (4.3.7)$$

and denote the solution of $\ddot{\mathbf{S}}_{MG(gj)}(\boldsymbol{\theta}) = \mathbf{0}$ by $\ddot{\boldsymbol{\theta}}_{MG(gj)}$. Following the one-step jackknife method of Lipsitz et. al. (1994), we have

$$\ddot{\boldsymbol{\theta}}_{MG(gj)} = \hat{\boldsymbol{\theta}}_{MG} + \ddot{\mathbf{J}}_{MG(gj)}^{-1}(\hat{\boldsymbol{\theta}}_{MG}) \ddot{\mathbf{S}}_{MG(gj)}(\hat{\boldsymbol{\theta}}_{MG}), \quad (4.3.8)$$

where

$$\ddot{\mathbf{J}}_{MG(gj)}(\hat{\boldsymbol{\theta}}_{MG}) = - \left. \frac{\partial \ddot{\mathbf{S}}_{MG(gj)}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T} \right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_{MG}} = - \sum_{m=1}^M \sum_s \ddot{w}_{hik(gj)}^m a_{hik} \eta_{hik}^m \left. \frac{\partial \mathbf{u}_{hik}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T} \right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_{MG}}$$

and $\ddot{\mathbf{S}}_{MG(gj)}(\hat{\boldsymbol{\theta}}_{MG})$ is $\ddot{\mathbf{S}}_{MG(gj)}(\boldsymbol{\theta})$ evaluated at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_{MG}$. Noting that $\ddot{\mathbf{J}}_{MG(gj)}^{-1}(\hat{\boldsymbol{\theta}}_{MG}) \approx \hat{\mathbf{J}}_{MG}^{-1}(\hat{\boldsymbol{\theta}}_{MG})$ and $\hat{\mathbf{S}}_{MG}(\hat{\boldsymbol{\theta}}_{MG}) = \mathbf{0}$ then

$$\ddot{\boldsymbol{\theta}}_{MG(gj)} \approx \hat{\boldsymbol{\theta}}_{MG} + \hat{\mathbf{J}}_{MG}^{-1}(\hat{\boldsymbol{\theta}}_{MG}) \left\{ \ddot{\mathbf{S}}_{MG(gj)}(\hat{\boldsymbol{\theta}}_{MG}) - \hat{\mathbf{S}}_{MG}(\hat{\boldsymbol{\theta}}_{MG}) \right\}. \quad (4.3.9)$$

An EF jackknife variance estimator of $\hat{\boldsymbol{\theta}}_{MG}$ is

$$\widehat{\mathbf{V}}_{EFJ}(\hat{\boldsymbol{\theta}}_{MG}) = \sum_g \frac{n_g - 1}{n_g} \sum_j \left(\ddot{\boldsymbol{\theta}}_{MG(gj)} - \hat{\boldsymbol{\theta}}_{MG} \right) \left(\ddot{\boldsymbol{\theta}}_{MG(gj)} - \hat{\boldsymbol{\theta}}_{MG} \right)^T \quad (4.3.10)$$

which will be shown to be approximately equal to $\widehat{\mathbf{V}}_{JL}(\hat{\boldsymbol{\theta}}_{MG})$. Using the approximation in (4.3.9) we can write (4.3.10) as

$$\begin{aligned} \widehat{\mathbf{V}}_{EFJ}(\hat{\boldsymbol{\theta}}_{MG}) &\approx \hat{\mathbf{J}}_{MG}^{-1}(\hat{\boldsymbol{\theta}}_{MG}) \sum_g \frac{n_g - 1}{n_g} \sum_j \left(\ddot{\mathbf{S}}_{MG(gj)}(\hat{\boldsymbol{\theta}}_{MG}) - \hat{\mathbf{S}}_{MG}(\hat{\boldsymbol{\theta}}_{MG}) \right) \\ &\quad \left(\ddot{\mathbf{S}}_{MG(gj)}(\hat{\boldsymbol{\theta}}_{MG}) - \hat{\mathbf{S}}_{MG}(\hat{\boldsymbol{\theta}}_{MG}) \right)^T \hat{\mathbf{J}}_{MG}^{-1}(\hat{\boldsymbol{\theta}}_{MG}) \\ &= \hat{\mathbf{J}}_{MG}^{-1}(\hat{\boldsymbol{\theta}}_{MG}) \widehat{\mathbf{V}}_{EFJ}(\hat{\mathbf{S}}_{MG}(\hat{\boldsymbol{\theta}}_{MG})) \hat{\mathbf{J}}_{MG}^{-1}(\hat{\boldsymbol{\theta}}_{MG}). \end{aligned} \quad (4.3.11)$$

To calculate $\widehat{\mathbf{V}}_{EFJ}(\hat{\mathbf{S}}_{MG}(\hat{\boldsymbol{\theta}}_{MG}))$ we replace y_{hik} in (4.2.16) by a Q -dimensional vector $\mathbf{u}_{hik}(\hat{\boldsymbol{\theta}}_{MG})$ and get

$$\begin{aligned} &\ddot{\mathbf{S}}_{MG(gj)}(\hat{\boldsymbol{\theta}}_{MG}) - \hat{\mathbf{S}}_{MG}(\hat{\boldsymbol{\theta}}_{MG}) \\ &\approx - \left\{ \sum_{m=1}^M \sum_s d_m w_{hik(gj)} g_{hik} a_{hik} \eta_{hik}^m \mathbf{e}_{hik}(\hat{\boldsymbol{\theta}}_{MG}) \right. \\ &\quad \left. + \sum_{m=1}^M \sum_s \Delta_{gjm} w_{hik(gj)} g_{hik} a_{hik} \eta_{hik}^m \mathbf{e}_{hik}(\hat{\boldsymbol{\theta}}_{MG}) \right\} \\ &\approx \frac{1}{n_g - 1} \left(\bar{\mathbf{e}}_{gj}(\hat{\boldsymbol{\theta}}_{MG}) - \bar{\mathbf{e}}_g(\hat{\boldsymbol{\theta}}_{MG}) \right) - \sum_{m=1}^M \Delta_{gjm} \hat{\boldsymbol{\tau}}_m(\hat{\boldsymbol{\theta}}_{MG}) \end{aligned} \quad (4.3.12)$$

where Δ_{gjm} is defined in (4.2.20). Substituting (4.3.12) into (4.3.11) gives

$$\widehat{\mathbf{V}}_{EFJ}(\hat{\boldsymbol{\theta}}_{MG}) \approx v \left[\hat{\mathbf{J}}_{MG}^{-1}(\hat{\boldsymbol{\theta}}_{MG}) \left\{ \bar{\mathbf{e}}_{gj}(\hat{\boldsymbol{\theta}}_{MG}) + \sum_{m=1}^M \gamma_{gjm} \hat{\boldsymbol{\tau}}_m(\hat{\boldsymbol{\theta}}_{MG}) \right\} \right]. \quad (4.3.13)$$

Hence the EF jackknife variance estimator $\widehat{\mathbf{V}}_{EFJ}(\widehat{\boldsymbol{\theta}}_{MG})$ is approximately equal to the linearized jackknife variance estimator in (4.3.4). In the case of a single weighting class the second term in (4.3.13) vanishes because $\widehat{\boldsymbol{\tau}}_m(\widehat{\boldsymbol{\theta}}_{MG}) = \mathbf{0}$.

Chapter 5

Variance Estimation Under Mean Imputation

5.1 Introduction

Item nonresponse occurs frequently in sample surveys with many items and is usually handled by some form of imputation, where values are inserted for the missing items. These imputed values are obtained from the respondents under a given imputation rule. Imputation procedures are used not only to reduce the nonresponse bias but to create a complete data set, from which survey estimates can be computed. An advantage of imputation is that it permits the use of the same survey weight for all items, unlike weighting adjustment which is typically used for unit nonresponse. However, imputation in general, does not increase the statistical precision of survey estimates.

Imputation procedures include deterministic imputation; such as mean imputa-

tion, ratio imputation and regression imputation, and stochastic imputation; such as random imputation and hot deck imputation. For deterministic imputation, the imputed values are deterministic given the sample of respondents and any auxiliary information on the nonrespondents. Stochastic imputation substitutes a missing item by one of the respondents under a given imputation rule. Deterministic imputation eliminates imputation variance of the estimator of a mean or total, but the distribution of item values is not preserved. For example, mean imputation leads to a spike at the mean of the respondent values. On the other hand, stochastic imputation preserves the distribution of item values but leads to imputation variance which can be a significant component of the total variance if item nonresponse rate is high.

We will confine our discussion in this chapter to mean imputation where the missing items are substituted by the mean of the respondents, assuming that only item nonresponse occurs. Section 5.2 presents jackknife variance estimators for the calibration estimator of total. We then extend these results to cover parameters obtained as solutions of estimating equations in Section 5.3.

5.2 Calibration Estimator of a Total

We consider the stratified multistage sampling design in Section 2.2. For a sample, s , of size n , suppose r elements, s_r , respond and m elements, s_m , do not respond to an item y . Let y_{hik}^* , for all $(hik) \in s_m$ be the imputed values for the missing data, based on the set of respondents (donors) $\{y_{hik}, (hik) \in s_r\}$. The imputed estimator

of the population total $Y = \sum_U y_{hik}$ is then given by

$$\hat{Y}_I = \sum_{s_r} w_{hik} y_{hik} + \sum_{s_m} w_{hik} y_{hik}^*, \quad (5.2.1)$$

where the same basic design weight, w_{hik} , is used in (5.2.1) for all items y attached to the $(hik)^{\text{th}}$ sample element.

5.2.1 One Imputation Class

Let the response indicator variable, a_{hik} , be defined by

$$a_{hik} = \begin{cases} 1 & \text{if } (hik) \in s_r \\ 0 & \text{if } (hik) \in s_m \end{cases}$$

under a uniform response mechanism, where $E(a_{hik}) = p > 0$ for all $(hik) \in s$ and the a_{hik} 's are independent of each other and also independent of y_{hik} .

Mean imputation uses the mean of the respondents as the imputed values for the missing data. In other words, $y_{hik}^* = \hat{T}/\hat{R}$ for all $(hik) \in s_m$, where $\hat{T} = \sum_s w_{hik} a_{hik} y_{hik}$ and $\hat{R} = \sum_s w_{hik} a_{hik}$. In this case the imputed estimator of Y in (5.2.1) reduces to

$$\hat{Y}_I = \frac{\hat{T}}{\hat{R}} \hat{W} \quad (5.2.2)$$

where $\hat{W} = \sum_s w_{hik}$. This ratio estimator has no imputation variance because y_{hik}^* is deterministic given s_r . Noting that

$$E(\hat{T}) = E\left(\sum_s w_{hik} a_{hik} y_{hik}\right) = p E(\hat{Y}) = p Y,$$

$$E(\hat{R}) = E\left(\sum_s w_{hik} a_{hik}\right) = p E\left(\sum_s w_{hik}\right) = p M,$$

$$\text{and } E(\hat{W}) = E\left(\sum_s w_{hik}\right) = M,$$

where M is the total number of elements in the population, it follows that $\hat{Y}_I$ is approximately unbiased under uniform response. The sample variance under mean imputation seriously underestimates the true variance when the proportion of missing values is appreciable (see Chen, Rao and Sitter, 2000). That is, the variability of item values is understated due to a spike at the point $\hat{T}/\hat{R}$.

For the calibration estimator of total under imputation we have two options. We can perform the imputation procedure first and then apply calibration adjustments. Alternatively, we can calculate the calibration adjusted weights and use these weights in the imputation procedure to calculate the mean of the respondents. We follow the treatment in Yung and Rao (2000) where the calibration adjustment is performed first then followed by the imputation procedure. This approach should lead to more efficient imputation procedures because the calibration estimator of total is more efficient than the basic estimator of total. In this case, mean imputation uses

$$\tilde{y}_{hik}^* = \frac{\tilde{T}}{\tilde{R}} \text{ for all } (hik) \in s_m, \quad (5.2.3)$$

where $\tilde{T} = \sum_s \tilde{w}_{hik} a_{hik} y_{hik}$ and $\tilde{R} = \sum_s \tilde{w}_{hik} a_{hik}$, as the imputed values for the missing data. We note that the calibration weights, $\tilde{w}_{hik} = w_{hik} g_{hik}$ where $g_{hik} = \mathbf{X}^T \tilde{\mathbf{A}}^{-1} \mathbf{x}_{hik}$ and $\tilde{\mathbf{A}} = \sum_s w_{hik} a_{hik} \mathbf{x}_{hik} \mathbf{x}_{hik}^T$, are used in the imputation procedure. The calibration estimator of total under mean imputation is then given by

$$\begin{aligned} \hat{Y}_{IG} &= \sum_s \tilde{w}_{hik} a_{hik} y_{hik} + \sum_s \tilde{w}_{hik} (1 - a_{hik}) \tilde{y}_{hik}^* \\ &= \frac{\tilde{T}}{\tilde{R}} \tilde{W} \end{aligned} \quad (5.2.4)$$

where $\tilde{W} = \sum_s \tilde{w}_{hik}$. Since

$$\begin{aligned} E(\tilde{T}) &= E\left(\sum_s \tilde{w}_{hik} a_{hik} y_{hik}\right) = p E(\hat{Y}_G) = p Y, \\ E(\tilde{R}) &= E\left(\sum_s \tilde{w}_{hik} a_{hik}\right) = p E\left(\sum_s \tilde{w}_{hik}\right) = p M, \\ \text{and } E(\tilde{W}) &= E\left(\sum_s \tilde{w}_{hik}\right) = M, \end{aligned}$$

it follows that $\hat{Y}_{IG}$ is an approximately unbiased estimator of the population total Y .

If we replace y_{hik} and $\tilde{y}_{hik}^*$ in (5.2.4) by $\mathbf{x}_{hik}$ then we can see that $\hat{\mathbf{X}}_{IG}$ will benchmark to the known totals $\mathbf{X}$ of indicator variables $\mathbf{x}_{hik}$. We will now present jackknife variance estimators of $\hat{Y}_{IG}$.

A standard jackknife variance estimator of $\hat{Y}_{IG}$ is

$$v_{JN}(\hat{Y}_{IG}) = \sum_{g=1}^L \frac{n_g - 1}{n_g} \sum_{j=1}^{n_g} (\hat{Y}_{IG(gj)} - \hat{Y}_{IG})^2 \quad (5.2.5)$$

where

$$\hat{Y}_{IG(gj)} = \sum_s \tilde{w}_{hik(gj)} a_{hik} y_{hik} + \sum_s \tilde{w}_{hik(gj)} (1 - a_{hik}) \tilde{y}_{hik}^*.$$

We note that the jackknife variance estimator in (5.2.5) treats the imputed values as true values. Rao and Shao (1992) have shown that (5.2.5) leads to serious underestimation if the response rate ($\hat{p} = r/n$) is small.

Linearized jackknife

In view of the problem of underestimation associated with (5.2.5), Rao and Shao (1992) proposed a jackknife variance estimator that correctly estimates the variance of an estimator under imputation. Their method is based on making an adjustment to the imputed value whenever a respondent is to be deleted. The resulting jackknife

variance estimator is design-consistent under uniform response. Yung and Rao (2000) extended the Rao-Shao adjusted jackknife to calibration (GREG) estimators that use poststratification information. To illustrate the Rao-Shao adjusted jackknife method under mean imputation, let $\tilde{y}_{hik(gj)}^* = \tilde{T}_{(gj)}/\tilde{R}_{(gj)}$ for all $(hik) \in s_m$ when the $(gj)^{\text{th}}$ cluster is deleted, where $\tilde{T}_{(gj)} = \sum_s \tilde{w}_{hik(gj)} a_{hik} y_{hik}$ and $\tilde{R}_{(gj)} = \sum_s \tilde{w}_{hik(gj)} a_{hik}$. Then the adjusted imputed values, defined by

$$\tilde{z}_{hik(gj)}^* = \tilde{y}_{hik} + \tilde{y}_{hik(gj)}^* - \tilde{y}_{hik} = \frac{\tilde{T}_{(gj)}}{\tilde{R}_{(gj)}}, \quad (5.2.6)$$

make adjustments to $\tilde{y}_{hik}^*$ to reflect the fact that the donor set is changed when a sample cluster is deleted. The imputed values $\tilde{y}_{hik}^*$ remain unchanged whenever cluster $(gj) \in s_m$. Note that identification flags are needed when applying this method in order to identify observed and imputed values in the data file. Using the adjusted imputed values in (5.2.6), we define the adjusted imputed calibration estimator, when the $(gj)^{\text{th}}$ cluster is deleted, as

$$\hat{Y}_{IG(gj)}^a = \sum_s \tilde{w}_{hik(gj)} a_{hik} y_{hik} + \sum_s \tilde{w}_{hik(gj)} (1 - a_{hik}) \tilde{z}_{hik}^*, \quad (5.2.7)$$

and a jackknife variance estimator is given by

$$v_J(\hat{Y}_{IG}) = \sum_{g=1}^L \frac{n_g - 1}{n_g} \sum_{j=1}^{n_g} (\hat{Y}_{IG(gj)}^a - \hat{Y}_{IG})^2. \quad (5.2.8)$$

By linearizing (5.2.8) Yung and Rao (2000) derived a jackknife linearization variance estimator

$$v_{JL}(\hat{Y}_{IG}) = \sum_{g=1}^L \frac{1}{n_g(n_g - 1)} \sum_{j=1}^{n_g} (\tilde{r}_{gj} - \bar{\tilde{r}}_g)^2 = v(\tilde{r}_{gj}), \quad (5.2.9)$$

where

$$\tilde{r}_{gj} = \frac{\tilde{W}}{\tilde{R}} \sum_k (n_g \tilde{w}_{gjk}) \left(e_{1,gjk} - \frac{\tilde{T}}{\tilde{R}} e_{2,gjk} \right) + \frac{\tilde{T}}{\tilde{R}} \sum_k (n_g \tilde{w}_{gjk}) e_{3,gjk}, \quad (5.2.10)$$

with

$$\begin{aligned}
e_{1,gjk} &= (y_{gjk} - \mathbf{x}_{gjk}^T \tilde{\mathbf{B}}_1) a_{gjk}, \\
e_{2,gjk} &= (1 - \mathbf{x}_{gjk}^T \tilde{\mathbf{B}}_2) a_{gjk}, \\
e_{3,gjk} &= (1 - a_{gjk} \mathbf{x}_{gjk}^T \tilde{\mathbf{B}}_3), \\
\tilde{\mathbf{B}}_1 &= \tilde{\mathbf{A}}^{-1} \sum_s w_{hik} a_{hik} \mathbf{x}_{hik} y_{hik}, \\
\tilde{\mathbf{B}}_2 &= \tilde{\mathbf{A}}^{-1} \sum_s w_{hik} a_{hik} \mathbf{x}_{hik}, \\
\tilde{\mathbf{B}}_3 &= \tilde{\mathbf{A}}^{-1} \sum_s w_{hik} \mathbf{x}_{hik}, \\
\text{and } \tilde{\mathbf{A}} &= \sum_s w_{hik} a_{hik} \mathbf{x}_{hik} \mathbf{x}_{hik}^T.
\end{aligned}$$

Estimating function jackknife

In the presence of item nonresponse, define the sample estimating function by

$$\hat{\mathbf{S}}_I(\mathbf{B}) = \sum_s w_{hik} a_{hik} \mathbf{x}_{hik} (y_{hik} - \mathbf{x}_{hik}^T \mathbf{B}) \quad (5.2.11)$$

and denote the solution of $\hat{\mathbf{S}}_I(\mathbf{B}) = \mathbf{0}$ by $\tilde{\mathbf{B}}_I = \tilde{\mathbf{A}}^{-1} \tilde{\mathbf{U}}$, where $\tilde{\mathbf{A}} = \sum_s w_{hik} a_{hik} \mathbf{x}_{hik} \mathbf{x}_{hik}^T$ and $\tilde{\mathbf{U}} = \sum_s w_{hik} a_{hik} \mathbf{x}_{hik} y_{hik}$. We note that $\sum_s w_{hik} a_{hik} \mathbf{x}_{hik} e_{hik} = \mathbf{0}$ with $e_{hik} = y_{hik} - \mathbf{x}_{hik}^T \tilde{\mathbf{B}}_I$. A calibration estimator of the population total Y from the respondents is

$$\hat{Y}_{G_r} = \mathbf{X}^T \tilde{\mathbf{B}}_I = \sum_s \tilde{w}_{hik} a_{hik} y_{hik} \quad (5.2.12)$$

with $\tilde{w}_{hik}$ denoting the calibration weights. Applying the EF method, we replace w_{hik} and $\mathbf{B}$ in (5.2.11) by $w_{hik(gj)}$ and $\tilde{\mathbf{B}}_I$ respectively and define

$$\begin{aligned}
\hat{\mathbf{S}}_{I(gj)}(\tilde{\mathbf{B}}_I) &= \sum_s w_{hik(gj)} a_{hik} \mathbf{x}_{hik} (y_{hik} - \mathbf{x}_{hik}^T \tilde{\mathbf{B}}_I) \\
&= \sum_s w_{hik(gj)} a_{hik} \mathbf{x}_{hik} e_{hik}.
\end{aligned} \quad (5.2.13)$$

An EF jackknife estimator of $\mathbf{B}$, denoted by $\ddot{\mathbf{B}}_{I(gj)}$, is derived by solving $\hat{\mathbf{S}}_I(\mathbf{B}) = \hat{\mathbf{S}}_{I(gj)}(\bar{\mathbf{B}}_I)$ for $\mathbf{B}$. The solution is given by

$$\begin{aligned}\ddot{\mathbf{B}}_{I(gj)} &= \bar{\mathbf{B}}_I - \bar{\mathbf{A}}^{-1} \hat{\mathbf{S}}_{I(gj)}(\bar{\mathbf{B}}_I) \\ &= \bar{\mathbf{B}}_I - \bar{\mathbf{A}}^{-1} \sum_s w_{hik(gj)} a_{hik} \mathbf{x}_{hik} e_{hik}.\end{aligned}\quad (5.2.14)$$

An EF jackknife calibration estimator of the population total Y from the respondents is then given by

$$\ddot{Y}_{G_r(gj)} = \mathbf{X}^T \ddot{\mathbf{B}}_{I(gj)} = \sum_s \ddot{w}_{hik(gj)} a_{hik} y_{hik}, \quad (5.2.15)$$

where the EF jackknife calibration weights, $\ddot{w}_{hik(gj)}$, are defined by

$$\begin{aligned}\ddot{w}_{hik(gj)} &= \bar{w}_{hik} - w_{hik(gj)} g_{hik} \\ &\quad + w_{hik} \left(\sum_s w_{hik(gj)} g_{hik} a_{hik} \mathbf{x}_{hik}^T \right) \bar{\mathbf{A}}^{-1} \mathbf{x}_{hik}.\end{aligned}\quad (5.2.16)$$

Using these EF jackknife calibration weights, let

$$\ddot{y}_{hik(gj)}^* = \frac{\ddot{T}_{(gj)}}{\ddot{R}_{(gj)}}$$

be the imputed values for all $(hik) \in s_m$ when the $(gj)^{\text{th}}$ cluster is deleted, where $\ddot{T}_{(gj)} = \sum_s \ddot{w}_{hik(gj)} a_{hik} y_{hik}$ and $\ddot{R}_{(gj)} = \sum_s \ddot{w}_{hik(gj)} a_{hik}$. Define the EF jackknife adjusted imputed values by

$$\ddot{z}_{hik(gj)}^* = \bar{y}_{hik} + \ddot{y}_{hik(gj)}^* - \bar{y}_{hik} = \frac{\ddot{T}_{(gj)}}{\ddot{R}_{(gj)}}. \quad (5.2.17)$$

Then the EF jackknife adjusted calibration estimator of Y , under mean imputation, is

$$\begin{aligned}\ddot{Y}_{IG(gj)}^a &= \sum_s \ddot{w}_{hik(gj)} a_{hik} y_{hik} + \sum_s \ddot{w}_{hik(gj)} (1 - a_{hik}) \ddot{z}_{hik(gj)}^* \\ &= \frac{\ddot{T}_{(gj)}}{\ddot{R}_{(gj)}} \ddot{W}_{(gj)}\end{aligned}\quad (5.2.18)$$

where $\ddot{W}_{(gj)} = \sum_s \ddot{w}_{hik(gj)}$, and an EF jackknife variance estimator is given by

$$v_{EFJ}(\hat{Y}_{IG}) = \sum_{g=1}^L \frac{n_g - 1}{n_g} \sum_{j=1}^{n_g} (\ddot{Y}_{IG(gj)}^a - \hat{Y}_{IG})^2. \quad (5.2.19)$$

We will now show that $v_{EFJ}(\hat{Y}_{IG})$ is approximately equal to $v_{JL}(\hat{Y}_{IG})$. From (5.2.16) we note the following:

$$\begin{aligned} \ddot{T}_{(gj)} &= \bar{T} - \sum_s w_{hik(gj)} g_{hik} e_{1,hik}, \\ \ddot{R}_{(gj)} &= \bar{R} - \sum_s w_{hik(gj)} g_{hik} e_{2,hik}, \\ \text{and } \ddot{W}_{(gj)} &= \bar{W} - \sum_s w_{hik(gj)} g_{hik} e_{3,hik}. \end{aligned} \quad (5.2.20)$$

where

$$\begin{aligned} e_{1,hik} &= (y_{hik} - \mathbf{x}_{hik}^T \bar{\mathbf{B}}_1) a_{hik}, \\ e_{2,hik} &= (1 - \mathbf{x}_{hik}^T \bar{\mathbf{B}}_2) a_{hik}, \\ \text{and } e_{3,hik} &= (1 - a_{hik} \mathbf{x}_{hik}^T \bar{\mathbf{B}}_3). \end{aligned} \quad (5.2.21)$$

Using the following standard result for $|x| < 1$,

$$\begin{aligned} \frac{1}{1-x} &= \sum_{i=1}^{\infty} x^i \\ &\approx 1 + x \end{aligned} \quad (5.2.22)$$

by ignoring the lower order terms, we can write

$$\begin{aligned} \frac{1}{\ddot{R}_{(gj)}} &= \frac{1}{\bar{R} \left\{ 1 - \bar{R}^{-1} \sum_s w_{hik(gj)} g_{hik} e_{2,hik} \right\}} \\ &\approx \bar{R}^{-1} \left\{ 1 + \bar{R}^{-1} \sum_s w_{hik(gj)} g_{hik} e_{2,hik} \right\} \end{aligned} \quad (5.2.23)$$

and consequently

$$\ddot{Y}_{IG(gj)}^a \approx \frac{\bar{T}}{\bar{R}} \bar{W} - \frac{\bar{T}}{\bar{R}} \sum_s w_{hik(gj)} g_{hik} e_{3,hik} - \frac{\bar{W}}{\bar{R}} \sum_s w_{hik(gj)} g_{hik} e_{1,hik}$$

$$\begin{aligned}
& + \frac{\tilde{T}}{\bar{R}} \frac{\tilde{W}}{\bar{R}} \sum_s w_{hik(gj)} g_{hik} e_{2,hik} \\
= & \hat{Y}_{IG} - \left\{ \frac{\tilde{W}}{\bar{R}} \sum_s w_{hik(gj)} g_{hik} \left(e_{1,hik} - \frac{\tilde{T}}{\bar{R}} e_{2,hik} \right) \right. \\
& \left. + \frac{\tilde{T}}{\bar{R}} \sum_s w_{hik(gj)} g_{hik} e_{3,hik} \right\}. \tag{5.2.24}
\end{aligned}$$

Using the fact that $\sum_s w_{hik} a_{hik} \mathbf{x}_{hik} (y_{hik} - \mathbf{x}_{hik}^T \tilde{\mathbf{B}}_I) = 0$ we have

$$\begin{aligned}
\sum_s w_{hik} g_{hik} e_{1,hik} &= 0, \\
\sum_s w_{hik} g_{hik} e_{2,hik} &= 0, \\
\text{and } \sum_s w_{hik} g_{hik} e_{3,hik} &= 0,
\end{aligned} \tag{5.2.25}$$

and following the steps in (2.4.8) we can show that

$$\ddot{Y}_{IG(gj)}^a - \hat{Y}_{IG} \approx \frac{1}{n_g - 1} (\bar{r}_{gj} - \bar{r}_g). \tag{5.2.26}$$

Putting (5.2.26) into (5.2.19) shows that $v_{EFJ}(\hat{Y}_{IG}) \approx v_{JL}(\hat{Y}_{IG})$.

5.2.2 Multiple Imputation Classes

The assumption of uniform response is often unrealistic in practice. Suppose we can form $M \geq 2$ imputation classes and assume uniform response probabilities, p_m , within each of the imputation classes. Imputation is then performed independently within each imputation class, m , using the respondents from that class. Define the imputation class indicator variable by

$$\eta_{hik}^m = \begin{cases} 1 & \text{if element } (hik) \text{ is in imputation class } m \\ 0 & \text{otherwise} \end{cases}.$$

Under mean imputation, we use the mean of the respondents within the m^{th} imputation class as the imputed value for the missing data. That is, let $\tilde{y}_{hik}^* = \tilde{T}_m / \tilde{R}_m$, where $\tilde{T}_m = \sum_s \tilde{w}_{hik} a_{hik} \eta_{hik}^m y_{hik}$ and $\tilde{R}_m = \sum_s \tilde{w}_{hik} a_{hik} \eta_{hik}^m$. Then the calibration estimator of Y under mean imputation is given by

$$\begin{aligned} \hat{Y}_{GM} &= \sum_{m=1}^M \sum_s \tilde{w}_{hik} a_{hik} \eta_{hik}^m y_{hik} + \sum_{m=1}^M \sum_s \tilde{w}_{hik} (1 - a_{hik}) \eta_{hik}^m \tilde{y}_{hik}^* \\ &= \sum_{m=1}^M \frac{\tilde{T}_m}{\tilde{R}_m} \tilde{W}_m \end{aligned} \quad (5.2.27)$$

where $\tilde{W}_m = \sum_s \tilde{w}_{hik} \eta_{hik}^m$. Assuming uniform response within each imputation class, we can show that

$$\begin{aligned} E(\hat{Y}_{GM}) &\approx \sum_{m=1}^M \frac{E(\tilde{T}_m)}{E(\tilde{R}_m)} E(\tilde{W}_m) \\ &= E\left(\sum_{m=1}^M \sum_s \tilde{w}_{hik} \eta_{hik}^m y_{hik}\right) \\ &= E(\hat{Y}_G) \\ &= Y. \end{aligned}$$

Thus, the calibration imputed estimator, $\hat{Y}_{GM}$, is approximately unbiased.

Linearized jackknife

Using the Rao-Shao adjusted jackknife method, we define the value that one would impute for the $(hik)^{\text{th}}$ nonrespondent, in the m^{th} imputation class, if cluster (gj) is deleted as $\tilde{y}_{hik(gj)}^* = \tilde{T}_{m(gj)} / \tilde{R}_{m(gj)}$, where $\tilde{T}_{m(gj)} = \sum_s \tilde{w}_{hik(gj)} a_{hik} \eta_{hik}^m y_{hik}$ and $\tilde{R}_{m(gj)} = \sum_s \tilde{w}_{hik(gj)} a_{hik} \eta_{hik}^m$ with $\tilde{w}_{hik(gj)} = w_{hik(gj)} g_{hik(gj)}$. The adjusted imputed values are given by

$$\tilde{z}_{hik(gj)}^* = \tilde{y}_{hik}^* + \tilde{y}_{hik(gj)}^* - \tilde{y}_{hik}^* = \frac{\tilde{T}_{m(gj)}}{\tilde{R}_{m(gj)}} \quad (5.2.28)$$

and the adjusted imputed calibration estimator is

$$\hat{Y}_{GM(gj)}^a = \sum_{m=1}^M \sum_s \tilde{w}_{hik(gj)} a_{hik} \eta_{hik}^m y_{hik} + \sum_{m=1}^M \sum_s \tilde{w}_{hik(gj)} (1 - a_{hik}) \eta_{hik}^m \tilde{z}_{hik(gj)}^*. \quad (5.2.29)$$

Then a jackknife variance estimator is given by

$$v_J(\hat{Y}_{GM}) = \sum_{g=1}^L \frac{n_g - 1}{n_g} \sum_{j=1}^{n_g} (\hat{Y}_{GM(gj)}^a - \hat{Y}_{GM})^2 \quad (5.2.30)$$

which can be approximated by a jackknife linearization variance estimator (Yung and Rao, 2000)

$$v_{JL}(\hat{Y}_{GM}) = \sum_{g=1}^L \frac{1}{n_g(n_g - 1)} \sum_{j=1}^{n_g} (\tilde{r}_{gj} - \bar{\tilde{r}}_g)^2 = v(\tilde{r}_{gj}). \quad (5.2.31)$$

where

$$\tilde{r}_{gj} = \sum_{m=1}^M \left[\frac{\tilde{W}_m}{\tilde{R}_m} \sum_k (n_g \tilde{w}_{gjk}) \left(e_{1,gjk}^m - \frac{\tilde{T}_m}{\tilde{R}_m} e_{2,gjk}^m \right) + \frac{\tilde{T}_m}{\tilde{R}_m} \sum_k (n_g \tilde{w}_{gjk}) e_{3,gjk}^m \right] \quad (5.2.32)$$

with

$$e_{1,gjk}^m = (y_{gjk} - \mathbf{x}_{gjk}^T \tilde{\mathbf{B}}_1) a_{gjk} \eta_{gjk}^m,$$

$$e_{2,gjk}^m = (1 - \mathbf{x}_{gjk}^T \tilde{\mathbf{B}}_2) a_{gjk} \eta_{gjk}^m,$$

$$e_{3,gjk}^m = (1 - a_{gjk} \mathbf{x}_{gjk}^T \tilde{\mathbf{B}}_3) \eta_{gjk}^m,$$

$$\tilde{\mathbf{B}}_1 = \tilde{\mathbf{A}}_M^{-1} \sum_{m=1}^M \sum_s w_{hik} a_{hik} \eta_{hik}^m \mathbf{x}_{hik} y_{hik},$$

$$\tilde{\mathbf{B}}_2 = \tilde{\mathbf{A}}_M^{-1} \sum_{m=1}^M \sum_s w_{hik} a_{hik} \eta_{hik}^m \mathbf{x}_{hik},$$

$$\tilde{\mathbf{B}}_3 = \tilde{\mathbf{A}}_M^{-1} \sum_{m=1}^M \sum_s w_{hik} \eta_{hik}^m \mathbf{x}_{hik},$$

$$\text{and } \tilde{\mathbf{A}}_M = \sum_{m=1}^M \sum_s w_{hik} a_{hik} \eta_{hik}^m \mathbf{x}_{hik} \mathbf{x}_{hik}^T.$$

Estimating function jackknife

To apply the EF jackknife method under M imputation classes, define the sample estimating function by

$$\hat{\mathbf{S}}_{IM}(\mathbf{B}) = \sum_{m=1}^M \sum_s w_{hik} a_{hik} \eta_{hik}^m \mathbf{x}_{hik} (y_{hik} - \mathbf{x}_{hik}^T \mathbf{B}) \quad (5.2.33)$$

and solve $\hat{\mathbf{S}}_{IM}(\mathbf{B}) = \mathbf{0}$. Denote the solution by $\tilde{\mathbf{B}}_{IM} = \tilde{\mathbf{A}}_M^{-1} \tilde{\mathbf{U}}_M$, where $\tilde{\mathbf{A}}_M = \sum_{m=1}^M \sum_s w_{hik} a_{hik} \eta_{hik}^m \mathbf{x}_{hik} \mathbf{x}_{hik}^T$ and $\tilde{\mathbf{U}}_M = \sum_{m=1}^M \sum_s w_{hik} a_{hik} \eta_{hik}^m \mathbf{x}_{hik} y_{hik}$. We also note that $\sum_{m=1}^M \sum_s w_{hik} a_{hik} \eta_{hik}^m \mathbf{x}_{hik} e_{hik} = \mathbf{0}$ where $e_{hik} = y_{hik} - \mathbf{x}_{hik}^T \tilde{\mathbf{B}}_{IM}$. Then a calibration estimator of the population total Y from the respondents is

$$\hat{Y}_{GM_r} = \mathbf{X}^T \tilde{\mathbf{B}}_{IM} = \sum_{m=1}^M \sum_s \tilde{w}_{hik} a_{hik} \eta_{hik}^m y_{hik} \quad (5.2.34)$$

where $\tilde{w}_{hik} = w_{hik} g_{hik}$ with $g_{hik} = \mathbf{X}^T \tilde{\mathbf{A}}_M^{-1} \mathbf{x}_{hik}$. Define

$$\begin{aligned} \hat{\mathbf{S}}_{IM(gj)}(\tilde{\mathbf{B}}_{IM}) &= \sum_{m=1}^M \sum_s w_{hik(gj)} a_{hik} \eta_{hik}^m \mathbf{x}_{hik} (y_{hik} - \mathbf{x}_{hik}^T \tilde{\mathbf{B}}_{IM}) \\ &= \sum_{m=1}^M \sum_s w_{hik(gj)} a_{hik} \eta_{hik}^m \mathbf{x}_{hik} e_{hik} \end{aligned} \quad (5.2.35)$$

when w_{hik} and $\mathbf{B}$ in (5.2.33) are replaced by $w_{hik(gj)}$ and $\tilde{\mathbf{B}}_{IM}$ respectively. Denote the solution of $\hat{\mathbf{S}}_{IM}(\mathbf{B}) = \hat{\mathbf{S}}_{IM(gj)}(\tilde{\mathbf{B}}_{IM})$ by $\ddot{\mathbf{B}}_{IM(gj)}$, which can be expressed as

$$\begin{aligned} \ddot{\mathbf{B}}_{IM(gj)} &= \tilde{\mathbf{B}}_{IM} - \tilde{\mathbf{A}}_M^{-1} \hat{\mathbf{S}}_{IM(gj)}(\tilde{\mathbf{B}}_{IM}) \\ &= \tilde{\mathbf{B}}_{IM} - \tilde{\mathbf{A}}_M^{-1} \sum_{m=1}^M \sum_s w_{hik(gj)} a_{hik} \eta_{hik}^m \mathbf{x}_{hik} e_{hik}, \end{aligned} \quad (5.2.36)$$

and derive an EF jackknife calibration estimator of Y from the respondents as

$$\ddot{Y}_{GM_r(gj)} = \mathbf{X}^T \ddot{\mathbf{B}}_{IM(gj)} = \sum_{m=1}^M \sum_s \ddot{w}_{hik(gj)} a_{hik} \eta_{hik}^m y_{hik}, \quad (5.2.37)$$

where the EF jackknife calibration weights are defined by

$$\begin{aligned}\ddot{w}_{hik(gj)} &= \tilde{w}_{hik} - w_{hik(gj)} g_{hik} \\ &+ w_{hik} \left(\sum_{m=1}^M \sum_s w_{hik(gj)} g_{hik} a_{hik} \eta_{hik}^m \mathbf{x}_{hik}^T \right) \tilde{\mathbf{A}}_M^{-1} \mathbf{x}_{hik}.\end{aligned}\quad (5.2.38)$$

When the $(gj)^{\text{th}}$ cluster is deleted, we can define the imputed values for all nonrespondents in the m^{th} imputation class, using the weights $\ddot{w}_{hik(gj)}$, by

$$\ddot{y}_{hik(gj)}^* = \frac{\ddot{T}_{m(gj)}}{\ddot{R}_{m(gj)}}$$

where $\ddot{T}_{m(gj)} = \sum_s \ddot{w}_{hik(gj)} a_{hik} \eta_{hik}^m y_{hik}$ and $\ddot{R}_{m(gj)} = \sum_s \ddot{w}_{hik(gj)} a_{hik} \eta_{hik}^m$. Following the Rao-Shao adjusted jackknife, we define the EF jackknife adjusted imputed values by

$$\ddot{z}_{hik(gj)}^* = \ddot{y}_{hik}^* + \ddot{y}_{hik(gj)}^* - \ddot{y}_{hik}^* = \frac{\ddot{T}_{m(gj)}}{\ddot{R}_{m(gj)}} \quad (5.2.39)$$

and an EF jackknife adjusted calibration estimator of Y , under M imputation classes, is

$$\begin{aligned}\ddot{Y}_{GM(gj)}^a &= \sum_{m=1}^M \sum_s \ddot{w}_{hik(gj)} a_{hik} \eta_{hik}^m y_{hik} + \sum_{m=1}^M \sum_s \ddot{w}_{hik(gj)} (1 - a_{hik}) \eta_{hik}^m \ddot{z}_{hik}^* \\ &= \sum_{m=1}^M \frac{\ddot{T}_{m(gj)}}{\ddot{R}_{m(gj)}} \ddot{W}_{m(gj)}\end{aligned}\quad (5.2.40)$$

where $\ddot{W}_{m(gj)} = \sum_s \ddot{w}_{hik(gj)} \eta_{hik}^m$. An EF jackknife variance estimator is then given by

$$v_{EFJ}(\hat{Y}_{GM}) = \sum_{g=1}^L \frac{n_g - 1}{n_g} \sum_{j=1}^{n_g} (\ddot{Y}_{GM(gj)}^a - \hat{Y}_{GM})^2. \quad (5.2.41)$$

From the definition of the EF jackknife calibration weights in (5.2.38) we note that

$$\begin{aligned}\ddot{T}_{m(gj)} &= \tilde{T}_m - \sum_s w_{hik(gj)} g_{hik} e_{1,hik}^m, \\ \ddot{R}_{m(gj)} &= \tilde{R}_m - \sum_s w_{hik(gj)} g_{hik} e_{2,hik}^m, \\ \text{and } \ddot{W}_{m(gj)} &= \tilde{W}_m - \sum_s w_{hik(gj)} g_{hik} e_{3,hik}^m.\end{aligned}\quad (5.2.42)$$

Using the approximation in (5.2.22), we have

$$\begin{aligned} \frac{1}{\ddot{R}_{m(gj)}} &= \frac{1}{\tilde{R}_m \left[1 - \tilde{R}_m^{-1} \sum_s w_{hik(gj)} g_{hik} e_{2,hik}^m \right]} \\ &\approx \tilde{R}_m^{-1} \left\{ 1 + \tilde{R}_m^{-1} \sum_s w_{hik(gj)} g_{hik} e_{2,hik}^m \right\} \end{aligned} \quad (5.2.43)$$

and therefore (5.2.40) becomes

$$\begin{aligned} \ddot{Y}_{GM(gj)}^a &\approx \sum_{m=1}^M \frac{\tilde{T}_m}{\tilde{R}_m} \tilde{W}_m - \sum_{m=1}^M \frac{\tilde{T}_m}{\tilde{R}_m} \sum_s w_{hik(gj)} g_{hik} e_{3,hik}^m - \sum_{m=1}^M \frac{\tilde{W}_m}{\tilde{R}_m} \sum_s w_{hik(gj)} g_{hik} e_{1,hik}^m \\ &\quad + \sum_{m=1}^M \frac{\tilde{T}_m}{\tilde{R}_m} \frac{\tilde{W}_m}{\tilde{R}_m} \sum_s w_{hik(gj)} g_{hik} e_{2,hik}^m \\ &= \hat{Y}_{GM} - \sum_{m=1}^M \left\{ \frac{\tilde{W}_m}{\tilde{R}_m} \sum_s w_{hik(gj)} g_{hik} \left(e_{1,hik}^m - \frac{\tilde{T}_m}{\tilde{R}_m} e_{2,hik}^m \right) \right. \\ &\quad \left. + \frac{\tilde{T}_m}{\tilde{R}_m} \sum_s w_{hik(gj)} g_{hik} e_{3,hik}^m \right\}. \end{aligned} \quad (5.2.44)$$

Using the fact that $\sum_{m=1}^M \sum_s w_{hik} a_{hik} \eta_{hik}^m \mathbf{x}_{hik} (y_{hik} - \mathbf{x}_{hik}^T \tilde{\mathbf{B}}_{IM}) = 0$, it follows that

$$\begin{aligned} \sum_{m=1}^M \sum_s w_{hik} g_{hik} e_{1,hik}^m &= 0, \\ \sum_{m=1}^M \sum_s w_{hik} g_{hik} e_{2,hik}^m &= 0, \\ \text{and } \sum_{m=1}^M \sum_s w_{hik} g_{hik} e_{3,hik}^m &= 0. \end{aligned}$$

Following the steps in (2.4.8) we can show that

$$\ddot{Y}_{GM(gj)}^a - \hat{Y}_{GM} \approx \frac{1}{n_g - 1} (\tilde{r}_{gj} - \bar{\tilde{r}}_g) \quad (5.2.45)$$

and thus $v_{EFJ}(\hat{Y}_{GM}) \approx v_{JL}(\hat{Y}_{GM})$.

5.3 Calibration Estimator of θ

The estimating equations approach, discussed in Chapter 3, provides a unified method for estimating both simple and complex parameters from survey data. In this section, we discuss jackknife variance estimators for parameter estimates obtained from suitable estimating equations, in the presence of item nonresponse. For simplicity and ease of notation, we consider only the case of one imputation class.

Under mean imputation, we substitute a missing value by the weighted mean of the respondents. That is, we impute missing values by $\bar{y}_{hik}^*$ defined in (5.2.3). Then under item nonresponse, the resulting estimating equations that involve calibration weights are given by

$$\hat{\mathbf{S}}_{IG}(\theta) = \sum_s \tilde{w}_{hik} a_{hik} \mathbf{u}_{hik}(\theta) + \sum_s \tilde{w}_{hik} (1 - a_{hik}) \mathbf{u}_{hik}^*(\theta), \quad (5.3.1)$$

where $\mathbf{u}_{hik}(\theta)$ is defined in (3.2.4). Define the q^{th} component of $\mathbf{u}_{hik}^*(\theta)$ for $(hik) \in s_m$ by

$$\begin{aligned} \mathbf{u}_{hik}^*(\theta) &= \left(\frac{\partial \mu_{hik}}{\partial \theta_q} \right) \left\{ \frac{\bar{y}_{hik}^* - \mu_{hik}}{V_{0hik}} \right\}, \quad q = 1, \dots, Q \\ &= \left(\frac{\partial \mu_{hik}}{\partial \theta_q} \right) \{ \bar{y}_{hik}^* - \mu_{hik} \} \end{aligned} \quad (5.3.2)$$

where we assume $V_{0hik} = 1$ for simplicity. Solving the sample estimating equations

$$\hat{\mathbf{S}}_{IG}(\theta) = \mathbf{0}$$

we obtain an imputed estimator of the census parameter θ_N , denoted by $\hat{\theta}_{IG}$. We note that the estimator $\hat{\theta}_{IG}$ is asymptotically biased under the uniform response mechanism or under the assumed model in (3.2.1) and (3.2.2). We will now present jackknife estimators of the covariance matrix of $\hat{\theta}_{IG}$.

5.3.1 Linearized Jackknife

Yung, Hidioglou and Rao (2001) applied the Rao-Shao adjusted jackknife variance estimator which uses adjusted imputed values to compensate for the imputation. In the case of mean imputation, we define the q^{th} element of the adjusted imputed value when the $(gj)^{\text{th}}$ cluster has been deleted by

$$\mathbf{u}_{hik}^*(\boldsymbol{\theta}) = \left(\frac{\partial \mu_{hik}}{\partial \theta_q} \right) \{ \tilde{y}_{hik(gj)}^* - \mu_{hik} \} \quad (5.3.3)$$

where

$$\tilde{y}_{hik(gj)}^* = \frac{\tilde{T}_{(gj)}}{\tilde{R}_{(gj)}}$$

with $\tilde{T}_{(gj)} = \sum_s \tilde{w}_{hik(gj)} a_{hik} y_{hik}$ and $\tilde{R}_{(gj)} = \sum_s \tilde{w}_{hik(gj)} a_{hik}$. Then the parameter estimate of $\boldsymbol{\theta}_N$ when the $(gj)^{\text{th}}$ cluster has been deleted is defined as the solution of

$$\hat{\mathbf{S}}_{IG(gj)}(\boldsymbol{\theta}) = \sum_s \tilde{w}_{hik(gj)} a_{hik} \mathbf{u}_{hik}(\boldsymbol{\theta}) + \sum_s \tilde{w}_{hik(gj)} (1 - a_{hik}) \mathbf{u}_{hik(gj)}^*(\boldsymbol{\theta}) = \mathbf{0}. \quad (5.3.4)$$

The solution is denote by $\hat{\boldsymbol{\theta}}_{IG(gj)}$ and a jackknife estimator of the covariance matrix is

$$\hat{\mathbf{V}}_J(\hat{\boldsymbol{\theta}}_{IG}) = \sum_g \frac{n_g - 1}{n_g} \sum_j \left(\hat{\boldsymbol{\theta}}_{IG(gj)} - \hat{\boldsymbol{\theta}}_{IG} \right) \left(\hat{\boldsymbol{\theta}}_{IG(gj)} - \hat{\boldsymbol{\theta}}_{IG} \right)^T. \quad (5.3.5)$$

Yung, Hidioglou and Rao (2001) linearized (5.3.5) and obtain a jackknife linearization variance estimator of $\hat{\boldsymbol{\theta}}_{IG}$ as

$$\hat{\mathbf{V}}_{JL}(\hat{\boldsymbol{\theta}}_{IG}) = v \left[\hat{\mathbf{J}}_{IG}^{-1}(\hat{\boldsymbol{\theta}}_{IG}) \left\{ \tilde{\mathbf{e}}_{gj}(\hat{\boldsymbol{\theta}}_{IG}) + \hat{\boldsymbol{\Gamma}} D_{gj} \right\} \right], \quad (5.3.6)$$

where the q^{th} element of $\tilde{\mathbf{e}}_{gj}(\hat{\boldsymbol{\theta}}_{IG})$ is

$$\tilde{e}_{gjq}(\hat{\boldsymbol{\theta}}_{IG}) = \sum_k (n_g \tilde{w}_{gjk}) \left(\tilde{u}_{gjkq}(\hat{\boldsymbol{\theta}}_{IG}) - \mathbf{x}_{gjk}^T \tilde{\mathbf{B}}_q(\hat{\boldsymbol{\theta}}_{IG}) \right),$$

$$\begin{aligned}
D_{gj} &= \sum_k (n_g \tilde{w}_{gjk}) \left[(a_{gjk} y_{gjk} - \mathbf{x}_{gjk}^T \tilde{\mathbf{B}}_1) - \frac{\tilde{T}}{\tilde{R}} (a_{gjk} - \mathbf{x}_{gjk}^T \tilde{\mathbf{B}}_2) \right], \\
\tilde{\mathbf{B}}_q(\hat{\boldsymbol{\theta}}_{IG}) &= \tilde{\mathbf{A}}^{-1} \sum_s w_{hik} \mathbf{x}_{hik} \tilde{u}_{hikq}(\hat{\boldsymbol{\theta}}_{IG}), \\
\tilde{\mathbf{B}}_1 &= \tilde{\mathbf{A}}^{-1} \sum_s w_{hik} a_{hik} \mathbf{x}_{hik} y_{hik}, \\
\tilde{\mathbf{B}}_2 &= \tilde{\mathbf{A}}^{-1} \sum_s w_{hik} a_{hik} \mathbf{x}_{hik}, \\
\tilde{\mathbf{A}} &= \sum_s w_{hik} a_{hik} \mathbf{x}_{hik} \mathbf{x}_{hik}^T, \\
\tilde{\mathbf{u}}_{gjk}(\hat{\boldsymbol{\theta}}_{IG}) &= \begin{cases} \mathbf{u}_{gjk}(\hat{\boldsymbol{\theta}}_{IG}) & \text{if the } (gjk)^{\text{th}} \text{ sampled element responds} \\ \mathbf{u}_{gjk}^*(\hat{\boldsymbol{\theta}}_{IG}) & \text{otherwise} \end{cases}, \\
\text{and } \hat{\boldsymbol{\Gamma}} &= \sum_s \tilde{w}_{hik} (1 - a_{hik}) \mathbf{c}_{hik},
\end{aligned}$$

with the q^{th} element of $\mathbf{c}_{hik}$ being

$$c_{hikq} = \frac{\partial \mu_{hik}}{\partial \theta_q}$$

evaluated at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_{IG}$. We also note that the matrix $\hat{\mathbf{J}}_{IG}(\hat{\boldsymbol{\theta}}_{IG})$ is defined by

$$\hat{\mathbf{J}}_{IG}(\hat{\boldsymbol{\theta}}_{IG}) = - \left. \frac{\partial \hat{\mathbf{S}}_{IG}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T} \right|_{\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_{IG}}.$$

5.3.2 Estimating Function Jackknife

To apply the EF jackknife method in the estimating equations framework under mean imputation, we replace the calibration weights, $\tilde{w}_{hik}$, in (5.3.1) by the EF jackknife calibration weights, $\ddot{w}_{hik(gj)}$, and obtain

$$\ddot{\mathbf{S}}_{IG(gj)}(\boldsymbol{\theta}) = \sum_s \ddot{w}_{hik(gj)} a_{hik} \mathbf{u}_{hik}(\boldsymbol{\theta}) + \sum_s \ddot{w}_{hik(gj)} (1 - a_{hik}) \ddot{\mathbf{u}}_{hik(gj)}^*(\boldsymbol{\theta}), \quad (5.3.7)$$

where $\ddot{w}_{hik(gj)}$ is given by (5.2.16). The q^{th} element of the EF adjusted imputed value when the $(gj)^{\text{th}}$ cluster has been deleted is

$$\ddot{u}_{hikq(gj)}^*(\boldsymbol{\theta}) = \left(\frac{\partial \mu_{hik}}{\partial \theta_q} \right) \{ \ddot{y}_{hik(gj)}^* - \mu_{hik} \} \quad (5.3.8)$$

where

$$\ddot{y}_{hik(gj)}^* = \frac{\ddot{T}_{(gj)}}{\ddot{R}_{(gj)}}$$

with $\ddot{T}_{(gj)} = \sum_s \ddot{w}_{hik(gj)} a_{hik} y_{hik}$ and $\ddot{R}_{(gj)} = \sum_s \ddot{w}_{hik(gj)} a_{hik}$. Solving $\ddot{\mathbf{S}}_{IG(gj)}(\boldsymbol{\theta}) = \mathbf{0}$ we obtain an EF jackknife imputed estimator, $\ddot{\boldsymbol{\theta}}_{IG}$, of the parameter $\boldsymbol{\theta}_N$. Using the one-step jackknife method of Lipsitz et. al. (1994), we obtain

$$\ddot{\boldsymbol{\theta}}_{IG(gj)} \approx \hat{\boldsymbol{\theta}}_{IG} + \ddot{\mathbf{J}}_{IG(gj)}^{-1}(\hat{\boldsymbol{\theta}}_{IG}) \ddot{\mathbf{S}}_{IG(gj)}(\hat{\boldsymbol{\theta}}_{IG}), \quad (5.3.9)$$

where

$$\ddot{\mathbf{J}}_{IG(gj)}(\hat{\boldsymbol{\theta}}_{IG}) = - \left. \frac{\partial \ddot{\mathbf{S}}_{IG(gj)}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T} \right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_{IG}}$$

and $\ddot{\mathbf{S}}_{IG(gj)}(\hat{\boldsymbol{\theta}}_{IG})$ is $\ddot{\mathbf{S}}_{IG(gj)}(\boldsymbol{\theta})$ evaluated at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_{IG}$. Noting that $\ddot{\mathbf{J}}_{IG(gj)}^{-1}(\hat{\boldsymbol{\theta}}_{IG}) \approx \hat{\mathbf{J}}_{IG}^{-1}(\hat{\boldsymbol{\theta}}_{IG})$ and $\hat{\mathbf{S}}_{IG}(\hat{\boldsymbol{\theta}}_{IG}) = \mathbf{0}$, we now have

$$\ddot{\boldsymbol{\theta}}_{IG(gj)} \approx \hat{\boldsymbol{\theta}}_{IG} + \hat{\mathbf{J}}_{IG}^{-1}(\hat{\boldsymbol{\theta}}_{IG}) \{ \ddot{\mathbf{S}}_{IG(gj)}(\hat{\boldsymbol{\theta}}_{IG}) - \hat{\mathbf{S}}_{IG}(\hat{\boldsymbol{\theta}}_{IG}) \}, \quad (5.3.10)$$

and an EF jackknife variance estimator of $\hat{\boldsymbol{\theta}}_{IG}$ is given by

$$\widehat{\mathbf{V}}_{EFJ}(\hat{\boldsymbol{\theta}}_{IG}) = \sum_g \frac{n_g - 1}{n_g} \sum_j (\ddot{\boldsymbol{\theta}}_{IG(gj)} - \hat{\boldsymbol{\theta}}_{IG}) (\ddot{\boldsymbol{\theta}}_{IG(gj)} - \hat{\boldsymbol{\theta}}_{IG})^T. \quad (5.3.11)$$

Theorem 5.1 *The EF jackknife variance estimator is approximately equal to the jackknife linearization variance estimator:*

$$\widehat{\mathbf{V}}_{EFJ}(\hat{\boldsymbol{\theta}}_{IG}) \approx v \{ \hat{\mathbf{J}}_{IG}^{-1}(\hat{\boldsymbol{\theta}}_{IG}) [\bar{\mathbf{e}}_{gj}(\hat{\boldsymbol{\theta}}_{IG}) + \hat{\mathbf{\Gamma}} D_{gj}] \} = \widehat{\mathbf{V}}_{JL}(\hat{\boldsymbol{\theta}}_{IG}).$$

Proof: Putting (5.3.10) into (5.3.11), we have

$$\widehat{\mathbf{V}}_{EFJ}(\widehat{\boldsymbol{\theta}}_{IG}) \approx \widehat{\mathbf{J}}_{IG}^{-1}(\widehat{\boldsymbol{\theta}}_{IG}) \widehat{\mathbf{V}}_{EFJ}(\widehat{\mathbf{S}}_{IG}(\widehat{\boldsymbol{\theta}}_{IG})) \widehat{\mathbf{J}}_{IG}^{-1}(\widehat{\boldsymbol{\theta}}_{IG}) \quad (5.3.12)$$

where

$$\begin{aligned} \widehat{\mathbf{V}}_{EFJ}(\widehat{\mathbf{S}}_{IG}(\widehat{\boldsymbol{\theta}}_{IG})) &= \sum_g \frac{n_g - 1}{n_g} \sum_j \left(\ddot{\mathbf{S}}_{IG(gj)}(\widehat{\boldsymbol{\theta}}_{IG}) - \widehat{\mathbf{S}}_{IG}(\widehat{\boldsymbol{\theta}}_{IG}) \right) \\ &\quad \left(\ddot{\mathbf{S}}_{IG(gj)}(\widehat{\boldsymbol{\theta}}_{IG}) - \widehat{\mathbf{S}}_{IG}(\widehat{\boldsymbol{\theta}}_{IG}) \right)^T. \end{aligned} \quad (5.3.13)$$

Define the q^{th} component of $\ddot{\mathbf{S}}_{IG(gj)}(\widehat{\boldsymbol{\theta}}_{IG})$ in (5.3.7) by

$$\ddot{S}_{IG(gj)q}(\widehat{\boldsymbol{\theta}}_{IG}) = \sum_s \ddot{w}_{hik(gj)} a_{hik} u_{hikq}(\widehat{\boldsymbol{\theta}}_{IG}) + \sum_s \ddot{w}_{hik(gj)} (1 - a_{hik}) \ddot{u}_{hikq}^*(\widehat{\boldsymbol{\theta}}_{IG}) \quad (5.3.14)$$

using the EF jackknife calibration weights in (5.2.16). Combining the definitions of $\ddot{T}_{(gj)}$ and $\ddot{R}_{(gj)}$ in (5.2.20) with (5.2.23), we have

$$\begin{aligned} \ddot{u}_{hikq(gj)}^*(\widehat{\boldsymbol{\theta}}_{IG}) &\approx c_{hikq} \left\{ \frac{\ddot{T}}{\ddot{R}} - \mu_{hik} - \frac{1}{\ddot{R}} \sum_s w_{hik(gj)} g_{hik} a_{hik} (y_{hik} - \mathbf{x}_{hik}^T \bar{\mathbf{B}}_1) \right. \\ &\quad \left. + \frac{\ddot{T}}{\ddot{R}^2} \sum_s w_{hik(gj)} g_{hik} a_{hik} (1 - \mathbf{x}_{hik}^T \bar{\mathbf{B}}_2) \right\} \\ &= u_{hikq}^*(\widehat{\boldsymbol{\theta}}_{IG}) - c_{hikq} \left\{ \frac{1}{\ddot{R}} \sum_s w_{hik(gj)} g_{hik} e_{1,hik} \right. \\ &\quad \left. - \frac{\ddot{T}}{\ddot{R}^2} \sum_s w_{hik(gj)} g_{hik} e_{2,hik} \right\} \end{aligned}$$

where $e_{1,hik}$ and $e_{2,hik}$ are defined in (5.2.21). From (5.2.25) and (2.4.8) we can show that

$$\ddot{u}_{hikq(gj)}^*(\widehat{\boldsymbol{\theta}}_{IG}) \approx u_{hikq}^*(\widehat{\boldsymbol{\theta}}_{IG}) + c_{hikq} \frac{1}{n_g - 1} (D_{gj} - \bar{D}_g) \quad (5.3.15)$$

where

$$D_{gj} = \frac{1}{\ddot{R}} \left\{ \sum_k (n_g \ddot{w}_{gjk}) \left(e_{1,gjk} - \frac{\ddot{T}}{\ddot{R}} e_{2,gjk} \right) \right\}.$$

Now from (5.2.16) and (2.4.8) we can write

$$\begin{aligned}
& \sum_s \ddot{w}_{hik(gj)} a_{hik} u_{hikq}(\hat{\boldsymbol{\theta}}_{IG}) \\
&= \sum_s \tilde{w}_{hik} a_{hik} u_{hikq}(\hat{\boldsymbol{\theta}}_{IG}) - \sum_s w_{hik(gj)} g_{hik} a_{hik} (u_{hikq}(\hat{\boldsymbol{\theta}}_{IG}) - \mathbf{x}_{hik}^T \tilde{\mathbf{B}}_q(\hat{\boldsymbol{\theta}}_{IG})) \\
&= \sum_s \tilde{w}_{hik} a_{hik} u_{hikq}(\hat{\boldsymbol{\theta}}_{IG}) + \frac{1}{n_g - 1} (e_{gjq}(\hat{\boldsymbol{\theta}}_{IG}) - \bar{e}_{gq}(\hat{\boldsymbol{\theta}}_{IG})) \tag{5.3.16}
\end{aligned}$$

where

$$e_{gjq}(\hat{\boldsymbol{\theta}}_{IG}) = \sum_k (n_g \tilde{w}_{gjk}) a_{gjk} (u_{gjkq}(\hat{\boldsymbol{\theta}}_{IG}) - \mathbf{x}_{gjk}^T \tilde{\mathbf{B}}_q(\hat{\boldsymbol{\theta}}_{IG})).$$

The second term in (5.3.14) can be written as

$$\begin{aligned}
& \sum_s \ddot{w}_{hik(gj)} (1 - a_{hik}) \ddot{u}_{hikq(gj)}^*(\hat{\boldsymbol{\theta}}_{IG}) \\
&= \sum_s \tilde{w}_{hik} (1 - a_{hik}) \ddot{u}_{hikq(gj)}^*(\hat{\boldsymbol{\theta}}_{IG}) \\
&\quad - \sum_s w_{hik(gj)} g_{hik} (1 - a_{hik}) (\ddot{u}_{hikq(gj)}^*(\hat{\boldsymbol{\theta}}_{IG}) - \mathbf{x}_{hik}^T \ddot{\mathbf{B}}_q^*(\hat{\boldsymbol{\theta}}_{IG})). \tag{5.3.17}
\end{aligned}$$

Putting (5.3.15) in the first term of (5.3.17) gives

$$\begin{aligned}
& \sum_s \tilde{w}_{hik} (1 - a_{hik}) \ddot{u}_{hikq(gj)}^*(\hat{\boldsymbol{\theta}}_{IG}) \\
&\approx \sum_s \tilde{w}_{hik} (1 - a_{hik}) u_{hikq}^*(\hat{\boldsymbol{\theta}}_{IG}) + \sum_s \tilde{w}_{hik} (1 - a_{hik}) c_{hikq} \frac{1}{n_g - 1} (D_{gj} - \bar{D}_g) \\
&= \sum_s \tilde{w}_{hik} (1 - a_{hik}) u_{hikq}^*(\hat{\boldsymbol{\theta}}_{IG}) + \hat{\Gamma}_q \frac{1}{n_g - 1} (D_{gj} - \bar{D}_g). \tag{5.3.18}
\end{aligned}$$

For the second term in (5.3.17), we first note that

$$\begin{aligned}
& \ddot{\mathbf{B}}_q^*(\hat{\boldsymbol{\theta}}_{IG}) \\
&= \bar{\mathbf{A}}^{-1} \sum_s w_{hik} a_{hik} \mathbf{x}_{hik} \ddot{u}_{hikq}^*(\hat{\boldsymbol{\theta}}_{IG}) \\
&\approx \bar{\mathbf{A}}^{-1} \sum_s w_{hik} a_{hik} \mathbf{x}_{hik} u_{hikq}^*(\hat{\boldsymbol{\theta}}_{IG}) + \bar{\mathbf{A}}^{-1} \sum_s w_{hik} a_{hik} \mathbf{x}_{hik} c_{hikq} \frac{1}{n_g - 1} (D_{gj} - \bar{D}_g)
\end{aligned}$$

and this leads to

$$\begin{aligned} & \ddot{u}_{hikq(gj)}^*(\hat{\boldsymbol{\theta}}_{IG}) - \mathbf{x}_{hik}^T \ddot{\mathbf{B}}_q^*(\hat{\boldsymbol{\theta}}_{IG}) \\ & \approx u_{hikq}^*(\hat{\boldsymbol{\theta}}_{IG}) - \mathbf{x}_{hik}^T \ddot{\mathbf{B}}_q^*(\hat{\boldsymbol{\theta}}_{IG}) + c_{hikq} \frac{1}{n_g - 1} (D_{gj} - \bar{D}_g) \{1 - \mathbf{x}_{hik}^T \ddot{\mathbf{B}}_2\} \end{aligned} \quad (5.3.19)$$

Therefore, by ignoring the lower order terms,

$$\begin{aligned} & - \sum_s w_{hik(gj)} g_{hik} (1 - a_{hik}) \left(\ddot{u}_{hikq(gj)}^*(\hat{\boldsymbol{\theta}}_{IG}) - \mathbf{x}_{hik}^T \ddot{\mathbf{B}}_q^*(\hat{\boldsymbol{\theta}}_{IG}) \right) \\ & \approx - \sum_s w_{hik(gj)} g_{hik} (1 - a_{hik}) \left(u_{hikq}^*(\hat{\boldsymbol{\theta}}_{IG}) - \mathbf{x}_{hik}^T \ddot{\mathbf{B}}_q^*(\hat{\boldsymbol{\theta}}_{IG}) \right) \\ & \quad - \frac{1}{n_g - 1} (D_{gj} - \bar{D}_g) \left\{ \sum_s w_{hik(gj)} g_{hik} (1 - a_{hik}) c_{hikq} (1 - \mathbf{x}_{hik}^T \ddot{\mathbf{B}}_2) \right\} \\ & \approx \frac{1}{n_g - 1} (e_{gjq}^* - \bar{e}_{gq}^*) \end{aligned} \quad (5.3.20)$$

where

$$e_{gjq}^* = \sum_k (n_g \tilde{w}_{gjk} (u_{hikq}^*(\hat{\boldsymbol{\theta}}_{IG}) - \mathbf{x}_{hik}^T \ddot{\mathbf{B}}_q^*(\hat{\boldsymbol{\theta}}_{IG}))).$$

Substituting (5.3.18) and (5.3.20) into (5.3.17), we have

$$\begin{aligned} & \sum_s \ddot{w}_{hik(gj)} (1 - a_{hik}) \ddot{u}_{hikq(gj)}^*(\hat{\boldsymbol{\theta}}_{IG}) \\ & \approx \sum_s \tilde{w}_{hik} (1 - a_{hik}) u_{hikq}^*(\hat{\boldsymbol{\theta}}_{IG}) \\ & \quad + \frac{1}{n_g - 1} \{ (e_{gjq}^* - \bar{e}_{gq}^*) + \hat{\Gamma}_q (D_{gj} - \bar{D}_g) \}. \end{aligned} \quad (5.3.21)$$

Putting (5.3.16) and (5.3.21) in (5.3.7) gives

$$\ddot{\mathbf{S}}_{IG(gj)}(\hat{\boldsymbol{\theta}}_{IG}) - \hat{\mathbf{S}}_{IG}(\hat{\boldsymbol{\theta}}_{IG}) \approx \frac{1}{n_g - 1} \{ (\bar{e}_{gjq}^* - \bar{e}_{gq}^*) + \hat{\Gamma} (D_{gj} - \bar{D}_g) \} \quad (5.3.22)$$

completing the proof when (5.3.22) is substituted back into (5.3.13).

■

Chapter 6

Efficiency of Balanced Repeated Replication Variance Estimator in the Linear Case

6.1 Introduction

The balanced repeated replication (BRR) method (McCarthy, 1969) is often used for estimating the variance of nonlinear statistics from stratified samples with two primary sampling units per stratum. If there are two or more units in each stratum, a common practice is to divide the units in the stratum into two equal or nearly equal groups and apply the BRR method to these groups. The resulting variance estimators have been shown to be quite inefficient (Krewski, 1978). In this chapter, we will evaluate the loss in efficiency of the BRR variance estimator in the linear case, when two or more groups are employed to obtain the balanced replicates.

In Section 6.2 the usual variance estimator of the stratified sample mean is evaluated under some regularity conditions. Section 6.3 summarizes the loss in efficiency of the BRR variance estimator for two groups, as obtained in Krewski (1978). An alternative BRR variance estimator, based on mixed orthogonal arrays, is discussed in Section 6.4. The loss in efficiency of the BRR variance estimator due to grouping into two or more groups is evaluated in Section 6.5.

6.2 Variance Estimation in the Linear Case

In stratified random sampling, the finite population of N units is partitioned into L non-overlapping strata of sizes $N_1, \dots, N_L$ where $N = \sum_{h=1}^L N_h$. A simple random sample of size $n_h \geq 2$ is drawn with replacement from the h^{th} stratum, independently for each h . An unbiased estimator of the population mean $\bar{Y} = \sum_{h=1}^L W_h \bar{Y}_h$ is given by

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

where $\bar{y}_h = n_h^{-1} \sum_{i=1}^{n_h} y_{hi}$ and $W_h = N_h/N$. Letting $s_h^2 = (n_h - 1)^{-1} \sum_{i=1}^{n_h} (y_{hi} - \bar{y}_h)^2$ and σ_h^2 denote the sample and population variance in stratum h respectively, the variance of $\bar{y}_{st}$ is

$$V(\bar{y}_{st}) = \sum_{h=1}^L W_h^2 \frac{\sigma_h^2}{n_h}, \quad (6.2.1)$$

which is estimated by

$$v_L(\bar{y}_{st}) = \sum_{h=1}^L W_h^2 \frac{s_h^2}{n_h}, \quad (6.2.2)$$

provided $n_h \geq 2$ for $h = 1, \dots, L$. Krewski (1978) obtained the variance of $v_L(\bar{y}_{st})$ as

$$V(v_L(\bar{y}_{st})) = \sum_{h=1}^L \frac{(W_h \sigma_h)^4}{n_h^3} \left[\beta_h - \frac{(n_h - 3)}{(n_h - 1)} \right], \quad (6.2.3)$$

where $\beta_h = \frac{\mu_{4,h}}{\sigma_h^4}$ and $\mu_{4,h}$ denote the kurtosis and fourth central population moment in stratum h respectively.

Under the following simplifying assumptions (McCarthy, 1966),

$$W_h \sigma_h = M \text{ (a constant),}$$

$$\beta_h = 3 \text{ (as for the normal distribution) and} \quad (6.2.4)$$

$$n_h = p \text{ (an even positive number)}$$

for $h = 1, \dots, L$, (6.2.3) reduces to

$$V(v_L(\bar{y}_{st})) = \frac{2LM^4}{p^2(p-1)}. \quad (6.2.5)$$

This result applies to stratified multistage sampling designs by changing y_{hi} to the cluster totals $y_{hi} = \sum_k (n_h w_{hik}) y_{hik}$.

6.3 BRR for Two Groups

When $n_h = 2$ for all h , McCarthy (1969) proposed a balanced half-samples variance estimator of $\bar{y}_{st}$, given by

$$v_{BRR}(\bar{y}_{st}) = R^{-1} \sum_{r=1}^R \left(\bar{y}_{st}^{(r)} - \bar{y}_{st} \right)^2 \quad (6.3.1)$$

where $\bar{y}_{st}^{(r)}$ is the mean of the r^{th} half-sample. Using a balanced set of R half-samples, one can show that

$$\bar{y}_{st}^{(r)} = \sum_{h=1}^L W_h \frac{1}{2} \sum_{i=1}^2 \delta_{hi}^{(r)} y_{hi}$$

$$= \bar{y}_{st} + \sum_{h=1}^L W_h \gamma_h^{(r)} d_h \quad (6.3.2)$$

where $\delta_{hi}^{(r)}$, $i = 1, 2$, and $\gamma_h^{(r)}$ are defined in Section 2.3.3 and $d_h = y_{h1} - \bar{y}_h = -(y_{h2} - \bar{y}_h)$. Consequently the balanced half-samples variance estimator of $\bar{y}_{st}$ can be expressed as

$$v_{BRR}(\bar{y}_{st}) = \sum_{h=1}^L W_h^2 d_h^2, \quad (6.3.3)$$

which is identical to (6.2.2) in the special case of $n_h = 2$ for all h .

When the number of units selected in each stratum is even, Kish and Frankel (1970) suggested that the n_h units in stratum h be randomly divided into two groups of size $m_h = n_h/2$ and then the balanced half-samples variance formula (6.3.3) is applied by treating the two groups as the two units. That is, d_h in (6.3.3) is now defined as

$$d_h = \bar{y}_h^{(1)} - \bar{y}_h = -(\bar{y}_h^{(2)} - \bar{y}_h) \quad (6.3.4)$$

where $\bar{y}_h^{(1)} = m_h^{-1} \sum_{i=1}^{m_h} y_{hi}$ and $\bar{y}_h^{(2)} = m_h^{-1} \sum_{i=m_h+1}^{n_h} y_{hi}$ denote the means of the two groups in stratum h . Letting $E|_n$ and $V|_n$ denote the expectation and variance conditional on the $n = \sum_{h=1}^L n_h$ units selected in the sample, and E_n and V_n denote the expectation and variance over all possible samples, one can show that

$$\begin{aligned} E(d_h^2) &= E_n E|_n (\bar{y}_h^{(1)} - \bar{y}_h)^2 \\ &= E_n V|_n (\bar{y}_h^{(1)}) \\ &= \frac{n_h - m_h}{m_h} \frac{E_n(s_h^2)}{n_h} \\ &= \frac{\sigma_h^2}{n_h} \end{aligned}$$

since $n_h = 2m_h$. It follows that $E(v_{BRR}(\bar{y}_{st})) = V(\bar{y}_{st})$. Krewski (1978) obtained an

expression for the variance of $v_{BRR}(\bar{y}_{st})$ as

$$V(v_{BRR}(\bar{y}_{st})) = V(v_L(\bar{y}_{st})) + \sum_{h=1}^L W_h^4 E \left[d_h^2 - \frac{s_h^2}{n_h} \right]^2. \quad (6.3.5)$$

Krewski (1978) also showed that, under $n_h = p = 2m_h$ and the conditions in (6.2.4), the efficiency due to grouping is

$$\frac{V(v_{BRR}(\bar{y}_{st}))}{V(v_L(\bar{y}_{st}))} = p - 1. \quad (6.3.6)$$

which increases markedly as the sample size within each stratum increases. When $p = 2$, $v_{BRR}(\bar{y}_{st})$ is fully efficient in the linear case.

6.4 BRR and Orthogonal Arrays

Gupta and Nigam (1987) developed a balanced repeated replications method for general $n_h \geq 2$, based on mixed orthogonal arrays of strength two to form a set of balanced subsamples. They showed that the set of balanced subsamples gives the usual variance estimator of $\bar{y}_{st}$ defined by (6.2.2).

To form a set of balanced subsamples we will need the following notation. Suppose we define a set of R subsamples with one unit per stratum. Let $y_{i_{jh}}$ be the value of the unit selected from stratum h in the j^{th} subsample, where i_{jh} is one of $1, \dots, n_h$. Using an indicator variable,

$$\delta(i, i_{jh}) = \begin{cases} 1 & \text{if } i = i_{jh}, \quad i = 1, \dots, n_h, \quad j = 1, \dots, R \\ 0 & \text{otherwise} \end{cases} \quad (6.4.1)$$

we can write

$$y_{i_{jh}} = \sum_{i=1}^{n_h} \delta(i, i_{jh}) y_{hi}. \quad (6.4.2)$$

We will now define the conditions for the selection of balanced subsamples.

Definition 6.1 (Conditions for Balanced Subsamples)

$$\begin{aligned}
 \text{(A)} \quad & \sum_{j=1}^R \delta(i, i_{jh}) = \mu_h \\
 \text{(B)} \quad & \sum_{j=1}^R \sum_{i=1}^{n_h} \delta(i, i_{jh}) = R = \mu_h n_h \\
 \text{(C)} \quad & \sum_{j=1}^R \delta(i, i_{jh}) \delta(i', i'_{jh'}) = \mu_{hh'} \\
 \text{(D)} \quad & \sum_{j=1}^R \sum_{i=1}^{n_h} \sum_{i'=1}^{n_{h'}} \delta(i, i_{jh}) \delta(i', i'_{jh'}) = R = \mu_{hh'} n_h n_{h'}
 \end{aligned}$$

Condition **(A)** assumes that the number of times unit i from stratum h occurs over all R subsamples is constant for all i within h . Condition **(B)** follows from Condition **(A)** and gives $\mu_h = R/n_h$. Condition **(C)** assumes that the number of times unit i from stratum h appears in the same subsample as unit i' from stratum h' is constant for all pairs (i, i') from strata (h, h') . Condition **(D)** implies that $\mu_{hh'} = R/n_h n_{h'}$.

If $n_h = p$ for all h , let $\bar{y}_{st,j} = \sum_{h=1}^L W_h y_{i_{jh}}$ be an estimate of the population mean $\bar{Y}$ from the j^{th} subsample. Then, using conditions **(A)** and **(B)**,

$$\begin{aligned}
 \bar{y}_{st} &= \frac{1}{R} \sum_{j=1}^R \bar{y}_{st,j} \\
 &= \frac{1}{R} \sum_{j=1}^R \sum_{h=1}^L \sum_{i=1}^p W_h \delta(i, i_{jh}) y_{hi} \\
 &= \sum_{h=1}^L W_h \bar{y}_h
 \end{aligned}$$

as required. We will now show that the BRR variance estimator matches the usual estimator $v_L(\bar{y}_{st})$ in the linear case. The BRR variance estimator is given by

$$v_{BRR}(\bar{y}_{st}) = \frac{1}{R(p-1)} \sum_{j=1}^R (\bar{y}_{st,j} - \bar{y}_{st})^2 \quad (6.4.3)$$

and we note that

$$\begin{aligned}\bar{y}_{st,j} - \bar{y}_{st} &= \sum_{h=1}^L \frac{W_h}{p} \sum_{i=1}^p \{p\delta(i, i_{jh}) - 1\} y_{hi} \\ &= \sum_{h=1}^L \frac{W_h}{p} \sum_{i=1}^p \Delta(i, i_{jh}) y_{hi}\end{aligned}\quad (6.4.4)$$

where

$$\Delta(i, i_{jh}) = \begin{cases} p-1 & \text{if } i = i_{jh}, \quad i = 1, \dots, p, \quad j = 1, \dots, R \\ -1 & \text{otherwise} \end{cases}.$$

From conditions **(A)** and **(B)** we also have the following identities.

$$\sum_{j=1}^R \Delta(i, i_{jh}) = p \sum_{j=1}^R \delta(i, i_{jh}) - R = 0 \quad (6.4.5)$$

$$\sum_{i=1}^p \Delta(i, i_{jh}) = p - p = 0 \quad (6.4.6)$$

Similarly, using conditions **(C)** and **(D)** we have

$$\begin{aligned}\sum_{j=1}^R \Delta(i, i_{jh}) \Delta(i', i'_{jh'}) &= p^2 \sum_{j=1}^R \delta(i, i_{jh}) \delta(i', i'_{jh'}) - p \sum_{j=1}^R \delta(i, i_{jh}) \\ &\quad - p \sum_{j=1}^R \delta(i', i'_{jh'}) + R \\ &= p^2 \frac{R}{p^2} - p \frac{R}{p} - p \frac{R}{p} + R \\ &= 0.\end{aligned}\quad (6.4.7)$$

Substituting (6.4.4) into (6.4.3) gives

$$\begin{aligned}v_{BRR}(\bar{y}_{st}) &= \frac{1}{R(p-1)} \sum_{j=1}^R \left(\sum_{h=1}^L \frac{W_h}{p} a_{jh} \right)^2 \\ &= \frac{1}{R(p-1)} \left\{ \sum_{h=1}^L \frac{W_h^2}{p^2} \sum_{j=1}^R a_{jh}^2 + \sum_{h \neq h'} \frac{W_h}{p} \frac{W_{h'}}{p} \sum_{j=1}^R a_{jh} a_{jh'} \right\}\end{aligned}\quad (6.4.8)$$

where $a_{jh} = \sum_{i=1}^p \Delta(i, i_{jh}) y_{hi}$. Now

$$\sum_{j=1}^R a_{jh}^2 = \sum_{i=1}^p y_{hi}^2 \sum_{j=1}^R \Delta^2(i, i_{jh}) + \sum_{i \neq i'} y_{hi} y_{hi'} \sum_{j=1}^R \Delta(i, i_{jh}) \Delta(i', i'_{jh}) \quad (6.4.9)$$

and from conditions (A)-(D) we get

$$\begin{aligned}\sum_{j=1}^R \Delta^2(i, i_{jh}) &= R(p-1), \\ \sum_{j=1}^R \Delta(i, i_{jh}) \Delta(i', i'_{jh}) &= -R,\end{aligned}$$

and therefore

$$\sum_{j=1}^R a_{jh}^2 = R(p-1) \sum_{i=1}^p y_{hi}^2 - R \sum_{i \neq i'}^p y_{hi} y_{hi'}. \quad (6.4.10)$$

The second term in (6.4.8) vanishes because of the BRR property in (6.4.7). By substituting (6.4.10) into (6.4.8) we then get

$$\begin{aligned}v_{BRR}(\bar{y}_{st}) &= \frac{1}{p^2} \sum_{h=1}^L W_h^2 \sum_{i=1}^p y_{hi}^2 - \frac{1}{p^2(p-1)} \sum_{h=1}^L W_h^2 \sum_{i \neq i'}^p y_{hi} y_{hi'} \\ &= \sum_{h=1}^L W_h^2 \left\{ \frac{1}{p^2} \sum_{i=1}^p y_{hi}^2 - \frac{1}{(p-1)} \left(\bar{y}_h^2 - \frac{1}{p^2} \sum_{i=1}^p y_{hi}^2 \right) \right\} \\ &= \sum_{h=1}^L W_h^2 \left\{ \frac{1}{(p-1)} \left(\frac{1}{p} \sum_{i=1}^p (y_{hi} - \bar{y}_h)^2 \right) \right\} \\ &= \sum_{h=1}^L W_h^2 \frac{s_h^2}{p} \\ &= v_L(\bar{y}_{st}).\end{aligned}$$

6.5 BRR for Two or More Groups

In this section, we will extend the results of Krewski (1978) to include more than two groups. If $n_h = p = tm$ units are randomly divided into t groups of size m , then we can use the results in Section 6.4 with some modifications of the notation. Let

$$\delta_G(i, i_{jh}) = \begin{cases} 1 & \text{if } i = i_{jh}, \quad i = 1, \dots, t, \quad j = 1, \dots, R \\ 0 & \text{otherwise} \end{cases},$$

be the indicator variable for membership in a group, where i_{jh} denotes the i^{th} group selected from stratum h in the j^{th} subsample ($i_{jh} = 1, \dots, t$). Define the mean value of the i^{th} group selected from stratum h in the j^{th} subsample by

$$\bar{y}_{i,jh} = \sum_{i=1}^t \delta_G(i, i_{jh}) \bar{y}_{hi}.$$

Replacing p , $\delta(i, i_{jh})$ and y_{hi} in Section 6.4 by t , $\delta_G(i, i_{jh})$ and $\bar{y}_{hi}$ respectively, we can show that the grouped BRR variance estimator is

$$v_{GB}(\bar{y}_{st}) = \sum_{h=1}^L W_h^2 \frac{1}{t} \frac{1}{(t-1)} \sum_{i=1}^t (\bar{y}_{hi} - \bar{y}_h)^2 \quad (6.5.1)$$

where $\bar{y}_{hi} = m_h^{-1} \sum_{k=1}^{m_h} y_{hik}$ is the mean of the i^{th} group and $\bar{y}_h = t^{-1} \sum_{i=1}^t \bar{y}_{hi}$. In the case of two groups, this variance estimator reduces to (6.3.3) with d_h defined by (6.3.4).

We will now examine the expected value and variance of the grouped BRR variance estimator in (6.5.1).

Theorem 6.1 *The grouped BRR variance estimator is unbiased. That is,*

$$E(v_{GB}(\bar{y}_{st})) = V(\bar{y}_{st}).$$

Proof: Let $E|_n$ and $V|_n$ denote the expectation and variance conditional on the $n = \sum_{h=1}^L n_h$ units selected in the sample, and E_n and V_n denote the expectation and variance over all possible samples. Then

$$\begin{aligned} E|_n(v_{GB}(\bar{y}_{st})) &= E|_n \left\{ \sum_{h=1}^L W_h^2 \frac{1}{t} \frac{1}{(t-1)} \sum_{i=1}^t (\bar{y}_{hi} - \bar{y}_h)^2 \right\} \\ &= \sum_{h=1}^L W_h^2 \frac{1}{t} \frac{1}{(t-1)} \sum_{i=1}^t E|_n (\bar{y}_{hi} - \bar{y}_h)^2 \end{aligned}$$

$$\begin{aligned}
&= \sum_{h=1}^L W_h^2 \frac{1}{t} \frac{1}{(t-1)} \sum_{i=1}^t V|_n(\bar{y}_{hi}) \\
&= \sum_{h=1}^L W_h^2 \frac{1}{t} \frac{1}{(t-1)} \sum_{i=1}^t \left(\frac{p-m}{m} \right) \frac{s_h^2}{p} \\
&= \sum_{h=1}^L W_h^2 \frac{s_h^2}{p} \\
&= v_L(\bar{y}_{st}).
\end{aligned}$$

Hence

$$\begin{aligned}
E(v_{GB}(\bar{y}_{st})) &= E_n E|_n(v_{GB}(\bar{y}_{st})) \\
&= E_n(v_L(\bar{y}_{st})) \\
&= V(\bar{y}_{st}).
\end{aligned}$$

■

Theorem 6.2 *Under the conditions in (6.2.4) the variance of the grouped BRR variance estimator is given by*

$$V(v_{GB}(\bar{y}_{st})) = \frac{2LM^4}{p^2(p-1)} \left\{ \frac{p-1}{t-1} \right\}. \quad (6.5.2)$$

Proof: The variance of $v_{GB}(\bar{y}_{st})$ is defined by

$$\begin{aligned}
V(v_{GB}(\bar{y}_{st})) &= E_n V|_n(v_{GB}(\bar{y}_{st})) + V_n E|_n(v_{GB}(\bar{y}_{st})) \\
&= E_n V|_n(v_{GB}(\bar{y}_{st})) + V(v_L(\bar{y}_{st}))
\end{aligned} \quad (6.5.3)$$

where $V(v_L(\bar{y}_{st}))$ is given in (6.2.5). We will now consider the first term on the right hand side of (6.5.3). Let

$$\begin{aligned}
v_{GB}(\bar{y}_{st}) - E|_n(v_{GB}(\bar{y}_{st})) &= \sum_{h=1}^L W_h^2 \left\{ \frac{1}{t} \sum_{i=1}^t \left(\frac{1}{(t-1)} (\bar{y}_{hi} - \bar{y}_h)^2 - \frac{s_h^2}{p} \right) \right\} \\
&= \sum_{h=1}^L W_h^2 \frac{1}{t} \sum_{i=1}^t A_{hi}
\end{aligned} \quad (6.5.4)$$

where

$$A_{hi} = \frac{1}{(t-1)}(\bar{y}_{hi} - \bar{y}_h)^2 - \frac{s_h^2}{p}, \quad i = 1, \dots, t.$$

Now

$$\begin{aligned} V|_n(v_{GB}(\bar{y}_{st})) &= E|_n [v_{GB}(\bar{y}_{st}) - E|_n(v_{GB}(\bar{y}_{st}))]^2 \\ &= \sum_{h=1}^L W_h^4 \left[\frac{1}{t^2} \sum_{i=1}^t E|_n(A_{hi}^2) + \frac{1}{t^2} \sum_{i \neq i'}^t E|_n(A_{hi}A_{hi'}) \right] \end{aligned} \quad (6.5.5)$$

using (6.5.4). Then

$$\begin{aligned} E_n V|_n(v_{GB}(\bar{y}_{st})) &= \sum_{h=1}^L W_h^4 \left\{ \frac{1}{t^2} \sum_{i=1}^t E_n E|_n(A_{hi}^2) + \frac{1}{t^2} \sum_{i \neq i'}^t E_n E|_n(A_{hi}A_{hi'}) \right\} \\ &= \sum_{h=1}^L W_h^4 \left\{ \frac{1}{t^2} \sum_{i=1}^t E(A_{hi}^2) + \frac{1}{t^2} \sum_{i \neq i'}^t E(A_{hi}A_{hi'}) \right\}. \end{aligned} \quad (6.5.6)$$

Because of equal group sizes, $E(A_{hi}^2) = E(A_{h1}^2)$ for all i and $E(A_{hi}A_{hi'}) = E(A_{h1}A_{h2})$ for all $i \neq i'$. We will evaluate $E(A_{h1}^2)$ and $E(A_{h1}A_{h2})$ for the first two groups selected in stratum h . For $E(A_{h1}^2)$ we use the following result from Krewski (1978).

Theorem 6.3 *For $h = 1, \dots, L$, let $\bar{y}_h$ and s_h^2 denote the mean and variance of a simple random sample of size $n_h = p = tm \geq 2$ selected with replacement from a population with variance σ_h^2 and kurtosis β_h and let $\bar{y}_{h1}$ denote the mean of a subsample formed by randomly selecting m of the p units in the sample. Then*

$$E \left[\frac{m}{(p-m)}(\bar{y}_{h1} - \bar{y}_h)^2 - \frac{s_h^2}{p} \right]^2 = \frac{(\beta_h - 3)}{pm(p-m)}\sigma_h^4 - \frac{4\beta_h}{p^3}\sigma_h^4 + \frac{2(p^2 + 4p - 6)}{p^3(p-1)}\sigma_h^4$$

provided $p \geq 4$ or $p = 2$.

Using Theorem 6.3 and the conditions in (6.2.4), we have

$$E(A_{h1}^2) = \left[\frac{2(p^2 + 4p - 6)}{p^3(p-1)} - \frac{12}{p^3} \right] \sigma_h^4$$

$$= \frac{2(p-2)}{p^2(p-1)}\sigma_h^4. \quad (6.5.7)$$

Now

$$E(A_{h1}A_{h2}) = E_n \left[E|_n \left\{ \left(\frac{1}{(t-1)}(\bar{y}_{h1} - \bar{y}_h)^2 - \frac{s_h^2}{p} \right) \left(\frac{1}{(t-1)}(\bar{y}_{h2} - \bar{y}_h)^2 - \frac{s_h^2}{p} \right) \right\} \right] \quad (6.5.8)$$

and

$$\begin{aligned} E|_n & \left\{ \left(\frac{1}{(t-1)}(\bar{y}_{h1} - \bar{y}_h)^2 - \frac{s_h^2}{p} \right) \left(\frac{1}{(t-1)}(\bar{y}_{h2} - \bar{y}_h)^2 - \frac{s_h^2}{p} \right) \right\} \\ &= \frac{1}{(t-1)^2} E|_n (\bar{y}_{h1} - \bar{y}_h)^2 (\bar{y}_{h2} - \bar{y}_h)^2 - \frac{s_h^2}{p} E|_n \frac{1}{(t-1)} (\bar{y}_{h1} - \bar{y}_h)^2 \\ & \quad - \frac{s_h^2}{p} E|_n \frac{1}{(t-1)} (\bar{y}_{h2} - \bar{y}_h)^2 + \frac{s_h^4}{p^2} \\ &= \frac{1}{(t-1)^2} E|_n [(\bar{y}_{h1} - \bar{y}_h)^2 (\bar{y}_{h2} - \bar{y}_h)^2] - \frac{s_h^4}{p^2}. \end{aligned} \quad (6.5.9)$$

Substituting (6.5.9) into (6.5.8) gives

$$E(A_{h1}A_{h2}) = \frac{1}{(t-1)^2} E_n E|_n [(\bar{y}_{h1} - \bar{y}_h)^2 (\bar{y}_{h2} - \bar{y}_h)^2] - E_n \left(\frac{s_h^4}{p^2} \right). \quad (6.5.10)$$

We note from Wilks (1962, p.199) that

$$E(s_h^4) = \frac{\sigma_h^4}{p} \left[\beta_h + \frac{(p^2 - 2p + 3)}{(p-1)} \right], \quad (6.5.11)$$

and applying the conditions in (6.2.4) gives

$$E_n \left(\frac{s_h^4}{p^2} \right) = \sigma_h^4 \frac{(p+1)}{p^2(p-1)}. \quad (6.5.12)$$

For the first term on the right hand side of (6.5.10) we note that the sample s is randomly divided into t nonoverlapping random groups $s_1, \dots, s_t$ of equal size $m = p/t$. That is, first draw s_1 as a simple random sample without replacement

(SRSWOR) of size m from s . Then, draw s_2 as a SRSWOR of size m from the remaining units $s_{1c} = s - s_1$ and so on. Let $E|_{\bar{n}}$ and $V|_{\bar{n}}$ denote the expectation and variance conditional on the $p - m$ units in the reduced sample $s_{1c} = s - s_1$. Now

$$\begin{aligned}
E_n E|_n (\bar{y}_{h1} - \bar{y}_h)^2 (\bar{y}_{h2} - \bar{y}_h)^2 &= E_n E|_n E|_{\bar{n}} (\bar{y}_{h1} - \bar{y}_h)^2 (\bar{y}_{h2} - \bar{y}_h)^2 \\
&= E_n E|_n (\bar{y}_{h1} - \bar{y}_h)^2 E|_{\bar{n}} (\bar{y}_{h2} - \bar{y}_h)^2 \\
&= E_n E|_n (\bar{y}_{h1} - \bar{y}_h)^2 E|_{\bar{n}} (\bar{y}_{h2} - \bar{y}_{h1c} + \bar{y}_{h1c} - \bar{y}_h)^2 \\
&\quad \text{where } \bar{y}_{h1c} = \bar{y}_h - \frac{1}{(t-1)} (\bar{y}_{h1} - \bar{y}_h) \\
&= E_n E|_n (\bar{y}_{h1} - \bar{y}_h)^2 E|_{\bar{n}} (\bar{y}_{h2} - \bar{y}_{h1c})^2 \\
&\quad + \frac{1}{(t-1)^2} E_n E|_n (\bar{y}_{h1} - \bar{y}_h)^4. \tag{6.5.13}
\end{aligned}$$

We know that

$$E|_{\bar{n}} (\bar{y}_{h2} - \bar{y}_{h1c})^2 = \frac{p-2m}{p-m} \frac{s_{h1c}^2}{m} \tag{6.5.14}$$

where

$$s_{h1c}^2 = \frac{1}{(p-m)-1} \sum_{i=m+1}^p (y_{hi} - \bar{y}_{h1c})^2 \tag{6.5.15}$$

and use the following identity,

$$\begin{aligned}
(p-1)s_h^2 &= \sum_{i=1}^p (y_{hi} - \bar{y}_h)^2 \\
&= \sum_{i=1}^m (y_{hi} - \bar{y}_h)^2 + \sum_{i=m+1}^p (y_{hi} - \bar{y}_h)^2 \\
&= \sum_{i=1}^m (y_{hi} - \bar{y}_{h1} + \bar{y}_{h1} - \bar{y}_h)^2 + \sum_{i=m+1}^p (y_{hi} - \bar{y}_{h1c} + \bar{y}_{h1c} - \bar{y}_h)^2 \\
&= \sum_{i=1}^m (y_{hi} - \bar{y}_{h1})^2 + m(\bar{y}_{h1} - \bar{y}_h)^2 + \sum_{i=m+1}^p (y_{hi} - \bar{y}_{h1c})^2 + (p-m)(\bar{y}_{h1c} - \bar{y}_h)^2 \\
&= (m-1)s_{h1}^2 + \sum_{i=m+1}^p (y_{hi} - \bar{y}_{h1c})^2 + m(\bar{y}_{h1} - \bar{y}_h)^2 + (p-m)\frac{1}{(t-1)^2} (\bar{y}_{h1} - \bar{y}_h)^2 \\
&= (m-1)s_{h1}^2 + \sum_{i=m+1}^p (y_{hi} - \bar{y}_{h1c})^2 + \frac{p}{(t-1)} (\bar{y}_{h1} - \bar{y}_h)^2 \tag{6.5.16}
\end{aligned}$$

where $s_{h1}^2 = (m-1)^{-1} \sum_{i=1}^m (y_{hi} - \bar{y}_{h1})^2$. Putting (6.5.16) into (6.5.15) gives

$$s_{h1c}^2 = \frac{1}{(p-m)-1} \left\{ (p-1)s_h^2 - (m-1)s_{h1}^2 - \frac{p}{(t-1)}(\bar{y}_{h1} - \bar{y}_h)^2 \right\}, \quad (6.5.17)$$

and substituting (6.5.17) and (6.5.14) into the first term of equation (6.5.13), we have

$$\begin{aligned} & E_n E|_n (\bar{y}_{h1} - \bar{y}_h)^2 E|_{\bar{n}} (\bar{y}_{h2} - \bar{y}_{h1c})^2 \\ &= E_n E|_n (\bar{y}_{h1} - \bar{y}_h)^2 \frac{p-2m}{m(p-m)} \frac{1}{p-m-1} \left\{ (p-1)s_h^2 \right. \\ &\quad \left. - (m-1)s_{h1}^2 - \frac{p}{(t-1)}(\bar{y}_{h1} - \bar{y}_h)^2 \right\} \\ &= E_n \left[\frac{p-2m}{m(p-m)} \frac{1}{p-m-1} \left\{ (p-1)s_h^2 E|_n (\bar{y}_{h1} - \bar{y}_h)^2 \right. \right. \\ &\quad \left. \left. - (m-1)E|_n (\bar{y}_{h1} - \bar{y}_h)^2 s_{h1}^2 - \frac{p}{(t-1)} E|_n (\bar{y}_{h1} - \bar{y}_h)^4 \right\} \right] \\ &= \frac{p-2m}{m(p-m)} \frac{1}{p-m-1} \left\{ \frac{(p-1)(p-m)}{p} E_n(s_h^4) \right. \\ &\quad \left. - (m-1)E_n E|_n (\bar{y}_{h1} - \bar{y}_h)^2 s_{h1}^2 - \frac{p}{(t-1)} E_n E|_n (\bar{y}_{h1} - \bar{y}_h)^4 \right\}. \quad (6.5.18) \end{aligned}$$

Stuart and Ord (1993, p.447) show that

$$\begin{aligned} E|_n (\bar{y}_{h1} - \bar{y}_h)^4 &= \left[\alpha_3 - \frac{4}{p}\alpha_2 + \frac{6}{p^2}\alpha_1 - \frac{3(p-1)}{p(p+1)}\alpha_1^2 \right] K_4 \\ &\quad + 3\alpha_1^2 \frac{(p-1)}{(p+1)} K_2^2, \end{aligned} \quad (6.5.19)$$

where $\alpha_r = m^{-r} - p^{-r}$ ($r = 1, 2, 3$) and $K_2 = s_h^2$ and K_4 are the second and fourth sample K statistics. Noting that

$$E_n(K_4) = (\beta_h - 3)\sigma_h^4$$

and $E_n(K_2^2)$ is given in (6.5.11), we then have

$$\begin{aligned} E_n E|_n (\bar{y}_{h1} - \bar{y}_h)^4 &= \left[\alpha_3 - \frac{4}{p}\alpha_2 + \frac{6}{p^2}\alpha_1 - \frac{3(p-1)}{p(p+1)}\alpha_1^2 \right] (\beta_h - 3)\sigma_h^4 \\ &\quad + 3\alpha_1^2 \frac{(p-1)}{(p+1)} \frac{\sigma_h^4}{p} \left(\beta_h + \frac{(p^2 - 2p + 3)}{(p-1)} \right). \quad (6.5.20) \end{aligned}$$

Furthermore,

$$E_n E|_n (\bar{y}_{h1} - \bar{y}_h)^2 s_{h1}^2 = (\beta_h - 3) \sigma_h^4 \left(\alpha_2 - \frac{2}{p} \alpha_1^2 \right) + \sigma_h^4 \alpha_1. \quad (6.5.21)$$

Under the conditions in (6.2.4), we have

$$E_n E|_n (\bar{y}_{h1} - \bar{y}_h)^4 = \frac{3(p-m)^2}{p^2 m^2} \sigma_h^4, \quad (6.5.22)$$

$$E_n (s_h^4) = \frac{(p+1)}{(p-1)} \sigma_h^4, \quad (6.5.23)$$

$$\text{and } E_n E|_n (\bar{y}_{h1} - \bar{y}_h)^2 s_{h1}^2 = \frac{(p-m)}{pm} \sigma_h^4. \quad (6.5.24)$$

Substituting equations (6.5.22) - (6.5.24) into (6.5.18) gives

$$\begin{aligned} & E_n E|_n (\bar{y}_{h1} - \bar{y}_h)^2 E|_{\bar{n}} (\bar{y}_{h2} - \bar{y}_{h1c})^2 \\ &= \frac{p-2m}{m(p-m)} \frac{\sigma_h^4}{p-m-1} \left\{ \frac{(p-m)}{m} \frac{p+1}{p} \right. \\ &\quad \left. - (m-1) \frac{(p-m)}{pm} - \frac{p}{(t-1)} \frac{3(p-m)^2}{p^2 m^2} \right\} \\ &= \frac{p-2m}{m(p-m)} \frac{(p-m)}{p-m-1} \sigma_h^4 \left\{ \frac{(p+1) - (m-1) - 3}{pm} \right\} \\ &= \frac{p-2m}{pm^2} \sigma_h^4 \\ &= \frac{t(t-2)}{p^2} \sigma_h^4. \end{aligned} \quad (6.5.25)$$

Putting (6.5.25) and (6.5.22) back into (6.5.13), we have

$$E_n E|_n (\bar{y}_{h1} - \bar{y}_h)^2 (\bar{y}_{h2} - \bar{y}_h)^2 = \frac{t(t-2)}{p^2} \sigma_h^4 + \frac{3}{p^2} \sigma_h^4, \quad (6.5.26)$$

and combining (6.5.26) with (6.5.12) in (6.5.10) leads to

$$E(A_{h1} A_{h2}) = \sigma_h^4 \left\{ \frac{t(t-2)}{p^2(t-1)^2} + \frac{3}{p^2(t-1)^2} - \frac{(p+1)}{p^2(p-1)} \right\}. \quad (6.5.27)$$

Hence, using (6.5.7) and (6.5.27), we get

$$E_n V|_n (v_{GB}(\bar{y}_{st})) = \sum_{h=1}^L (W_h \sigma_h)^4 \frac{1}{t} \left\{ \frac{2(p-2)}{p^2(p-1)} + \frac{t(t-2)}{p^2(t-1)} + \frac{3}{p^2(t-1)} - \frac{(p+1)(t-1)}{p^2(p-1)} \right\}$$

$$\begin{aligned}
&= \sum_{h=1}^L (W_h \sigma_h)^4 \frac{1}{t} \left\{ \frac{2t(p-t)}{p^2(p-1)(t-1)} \right\} \\
&= \frac{2LM^4}{p^2(p-1)} \frac{(p-t)}{(t-1)}.
\end{aligned} \tag{6.5.28}$$

Substituting (6.5.28) into (6.5.3) we finally have

$$\begin{aligned}
V(v_{GB}(\bar{y}_{st})) &= \frac{2LM^4}{p^2(p-1)} \left\{ \frac{p-t}{t-1} + 1 \right\} \\
&= \frac{2LM^4}{p^2(p-1)} \left\{ \frac{p-1}{t-1} \right\}.
\end{aligned}$$

■

From Theorem 6.2 we can see that, under the conditions in (6.2.4), the efficiency due to grouping is

$$\frac{V(v_{GB}(\bar{y}_{st}))}{V(v_L(\bar{y}_{st}))} = \frac{p-1}{t-1}. \tag{6.5.29}$$

We note that (6.5.29) reduces to Krewski's result when $t = 2$. The efficiency loss is drastically reduced as the number of groups increases. But in practice, this reduction in efficiency loss should be weighed against the increased cost of computations.

Part II

Coverage Accuracy

Chapter 7

Coverage Accuracy of Confidence Intervals for the Mean

7.1 Introduction

In this chapter, we are interested in obtaining confidence intervals for a parameter. These confidence intervals are often obtained by inverting the distribution of some approximate pivotal quantity of the form,

$$\left\{ \frac{(\text{estimate} - \text{parameter})}{(\text{estimated variance})^{1/2}} \right\}.$$

For example, to obtain confidence intervals for the population mean $\bar{Y}$, one may use the studentized t -statistic

$$S_t = \frac{\bar{y} - \bar{Y}}{s_{\bar{y}}}, \tag{7.1.1}$$

where $s_{\bar{y}} = n^{-1/2} s_y$ is the standard deviation of the sample mean $\bar{y}$ and $s_y^2 = n^{-1} \sum_{i=1}^n (y_i - \bar{y})^2$ is the sample variance. We can then express the studentized

t -statistic as

$$S_t = \frac{n^{\frac{1}{2}}(\bar{y} - \bar{Y})}{\sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2}}. \quad (7.1.2)$$

Godambe and Thompson (1999) proposed a more direct method of constructing an approximate pivotal statistic using estimating function theory. For the estimation of the population mean $\bar{Y}$, the Godambe/Thompson statistic is given by

$$\tilde{S}_t = \frac{n^{\frac{1}{2}}(\bar{y} - \bar{Y})}{\sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \bar{Y})^2}}. \quad (7.1.3)$$

which is closer to $N(0, 1)$ than the studentized t -statistic. We will compare the studentized t -statistic with this approximate pivotal statistic proposed by Godambe and Thompson (1999), by examining Edgeworth expansions of these pivotal statistics.

Edgeworth expansion for the studentized t -statistic is given in Section 7.2, while Section 7.3 derives the Edgeworth expansion for the alternative pivotal statistic of Godambe and Thompson (1999). Section 7.4 gives a comparison of the coverage probabilities of the two statistics.

7.2 Studentized t -statistic

Suppose $y_1, \dots, y_n$ are independent and identically distributed (i.i.d.) random variables having mean $\bar{Y}$ and variance $\beta^2 > 0$. Let $\bar{y} = \frac{1}{n} \sum_i y_i$ denote the mean and $\hat{\beta}^2 = \frac{1}{n} \sum_i (y_i - \bar{y})^2$ the variance of the sample $\{y_1, \dots, y_n\}$. Suppose we have specified functions $g_i = g_i(y_i, \theta)$ such that

$$E\{g_i(y_i, \theta)\} = 0 \quad \text{for all } i = 1, \dots, n.$$

For estimating the population mean $\bar{Y}$, the function g_i is defined by

$$g_i(y_i, \theta) = y_i - \bar{Y}.$$

Let $\mathbf{h}_i(y_i, \theta) = (g_i(y_i, \theta), g_i^2(y_i, \theta))^T$ and define

$$\mathbf{T}_n = (\bar{g}, \bar{g}_2)^T,$$

where

$$\begin{aligned} \bar{g} &= \frac{1}{n} \sum_{i=1}^n g_i(y_i, \theta) = \frac{1}{n} \sum_{i=1}^n (y_i - \bar{Y}) = \bar{y} - \bar{Y}, \\ \text{and } \bar{g}_2 &= \frac{1}{n} \sum_{i=1}^n g_i^2(y_i, \theta) = \frac{1}{n} \sum_{i=1}^n (y_i - \bar{Y})^2. \end{aligned}$$

The studentized t -statistic can be expressed as

$$\begin{aligned} S_t &= \frac{n^{\frac{1}{2}} \bar{g}}{(\bar{g}_2 - \bar{g}^2)^{\frac{1}{2}}} \\ &= n^{\frac{1}{2}} H(\mathbf{T}_n) \end{aligned} \tag{7.2.1}$$

where

$$H(\mathbf{x}) = \frac{x^{(1)}}{(x^{(2)} - (x^{(1)})^2)^{\frac{1}{2}}}$$

with $\mathbf{x} = (x^{(1)}, x^{(2)})^T$ and the i^{th} element of a d -vector $\mathbf{x}$ is denoted by $x^{(i)}$. Since $\hat{\beta}^2 = \bar{g}_2 - \bar{g}^2$, we note that $E(\mathbf{T}_n) = (0, \beta^2)^T$. Using the definition of $H(\mathbf{x})$ we can write $H(\mathbf{T}_n)$ in (7.2.1) as

$$H(\mathbf{T}_n) = \frac{\bar{g}}{(\bar{g}_2 - \bar{g}^2)^{\frac{1}{2}}}. \tag{7.2.2}$$

Using the notation in Hall (1992), let

$$\mathbf{Z} = n^{\frac{1}{2}} (\mathbf{T}_n - E(\mathbf{T}_n)) = (Z^{(1)}, Z^{(2)})^T$$

where $Z^{(1)} = n^{\frac{1}{2}} \bar{g}$ and $Z^{(2)} = n^{\frac{1}{2}} (\bar{g}_2 - \beta^2)$, and define

$$a_{i_1, \dots, i_j} = \frac{\partial^j}{\partial x^{(i_1)} \dots \partial x^{(i_j)}} H(\mathbf{x}) \Big|_{\mathbf{x}=\mathbf{E}(\mathbf{x})}. \quad (7.2.3)$$

Then

$$\begin{aligned} a_1 &= \frac{\partial}{\partial \bar{g}} H(\mathbf{T}_n) \Big|_{\mathbf{T}_n=\mathbf{E}(\mathbf{T}_n)} = \frac{1}{\beta}, \\ a_{12} &= \frac{\partial^2}{\partial \bar{g} \partial \beta^2} H(\mathbf{T}_n) \Big|_{\mathbf{T}_n=\mathbf{E}(\mathbf{T}_n)} = -\frac{1}{2\beta^3} = a_{21}, \\ \text{and } a_{11} &= a_2 = a_{22} = 0. \end{aligned}$$

Define

$$\mu_{i_1, \dots, i_j} = E \left\{ (\mathbf{T}_n - E(\mathbf{T}_n))^{(i_1)} \dots (\mathbf{T}_n - E(\mathbf{T}_n))^{(i_j)} \right\}, \quad j \geq 1 \quad (7.2.4)$$

and therefore $\mu_i = 0$ for each i ,

$$E(Z^{(i)} Z^{(j)}) = \mu_{ij},$$

$$E(Z^{(i)} Z^{(j)} Z^{(k)}) = n^{-\frac{1}{2}} \mu_{ijk},$$

$$\text{and } E(Z^{(i)} Z^{(j)} Z^{(k)} Z^{(l)}) = \mu_{ij} \mu_{kl} + \mu_{ik} \mu_{jl} + \mu_{il} \mu_{jk} + O(n^{-1}).$$

Then

$$\mu_{11} = E(Z^{(1)} Z^{(1)}) = E(n^{\frac{1}{2}} \bar{g} n^{\frac{1}{2}} \bar{g}) = n E(\bar{g}^2) = n \frac{\beta^2}{n} = \beta^2,$$

$$\mu_{12} = E(Z^{(1)} Z^{(2)})$$

$$= E(n^{\frac{1}{2}} \bar{g} n^{\frac{1}{2}} (\bar{g}_2 - \beta^2))$$

$$= E(n^{\frac{1}{2}} \bar{g} n^{\frac{1}{2}} \bar{g}_2) - n^{\frac{1}{2}} \beta^2 E(\bar{g})$$

$$= \frac{1}{n} \sum_{i=1}^n E(g_i^3(y_i, \theta)) \quad \text{since } E(\bar{g}) = 0$$

$$= \frac{1}{n} \sum_{i=1}^n \mu_3 \quad \text{where } \mu_3 = E(g_i^3(y_i, \theta)) = E(y_i - \bar{Y})^3$$

$$\begin{aligned}
&= \mu_3, \\
\mu_{21} &= E(Z^{(2)} Z^{(1)}) \\
&= \mu_3, \\
\text{and } \mu_{111} &= n^{\frac{1}{2}} E\left(Z^{(1)} Z^{(1)} Z^{(1)}\right) \\
&= n^{\frac{1}{2}} E(n^{\frac{1}{2}} \bar{g} n^{\frac{1}{2}} \bar{g} n^{\frac{1}{2}} \bar{g}) \\
&= E\left(\frac{1}{n} \sum_{i=1}^n g_i^3(y_i, \theta)\right) \\
&= \mu_3.
\end{aligned}$$

Using equations (2.31) - (2.33) of Hall (1992), we have

$$\sigma^2 = \sum_{i=1}^2 \sum_{j=1}^2 a_i a_j \mu_{ij} = a_1 a_1 \mu_{12} = 1, \quad (7.2.5)$$

$$A_1 = \frac{1}{2} \sum_{i=1}^2 \sum_{j=1}^2 a_{ij} \mu_{ij} = \frac{1}{2} \{a_{12} \mu_{12} + a_{21} \mu_{21}\} = -\frac{1}{2} \frac{\mu_3}{\beta^3}, \quad (7.2.6)$$

$$\begin{aligned}
\text{and } A_2 &= \sum_{i=1}^2 \sum_{j=1}^2 \sum_{k=1}^2 a_i a_j a_k \mu_{ijk} + 3 \sum_{i=1}^2 \sum_{j=1}^2 \sum_{k=1}^2 \sum_{l=1}^2 a_i a_j a_{kl} \mu_{ik} \mu_{jl} \\
&= a_1 a_1 a_1 \mu_{111} + 3[a_1 a_1 a_{12} \mu_{11} \mu_{12} + a_1 a_1 a_{21} \mu_{12} \mu_{11}] \\
&= \frac{\mu_3}{\beta^3} + 3 \left[\frac{1}{\beta^2} \left(-\frac{1}{2\beta^3} \right) \beta^2 \mu_3 + \frac{1}{\beta^2} \left(-\frac{1}{2\beta^3} \right) \beta^2 \mu_3 \right] \\
&= \frac{\mu_3}{\beta^3} - 3 \left(\frac{\mu_3}{\beta^3} \right) \\
&= -2 \frac{\mu_3}{\beta^3}. \quad (7.2.7)
\end{aligned}$$

An Edgeworth expansion of the studentized t -statistic S_t , defined in (7.2.1), is given by

$$\begin{aligned}
P(S_t \leq x) &= \Phi(x) + n^{-\frac{1}{2}} q_1(x) \phi(x) + n^{-1} q_2(x) \phi(x) + \dots \\
&\quad + n^{-\frac{1}{2}} q_j(x) \phi(x) + o(n^{-\frac{1}{2}}), \quad (7.2.8)
\end{aligned}$$

where q_j is a polynomial of degree at most $3j - 1$, odd for even j and even for odd j , $\Phi(\cdot)$ and $\phi(\cdot)$ are the probability distribution function and the probability density function of a $N(0, 1)$ variable respectively.

From (7.2.5) - (7.2.7), it follows that the first polynomial is given by

$$\begin{aligned} q_1(x) &= -\left\{A_1\sigma^{-1} + \frac{1}{6}A_2\sigma^{-3}(x^2 - 1)\right\} \\ &= \frac{1}{6}\frac{\mu_3}{\beta^3}(2x^2 + 1) \\ &= \frac{1}{6}\gamma(2x^2 + 1) \end{aligned} \quad (7.2.9)$$

where $\gamma = \mu_3/\beta^3$. The second polynomial is given in equation (2.57) of Hall (1992) as

$$q_2(x) = x \left\{ \frac{1}{12}\kappa(x^2 - 3) - \frac{1}{18}\gamma^2(x^4 + 2x^2 - 3) - \frac{1}{4}(x^2 + 3) \right\} \quad (7.2.10)$$

where

$$\kappa = \frac{E(g_i^4(y_i, \theta))}{\beta^4} - 3.$$

An Edgeworth expansion of the coverage probability of a two-sided interval can be obtained from (7.2.8) as

$$\begin{aligned} P(-x \leq S_t \leq x) &= P(S_t \leq x) - P(S_t \leq -x) \\ &= \Phi(x) + n^{-\frac{1}{2}}q_1(x)\phi(x) + n^{-1}q_2(x)\phi(x) + o(n^{-\frac{3}{2}}) \\ &\quad - \left\{ \Phi(-x) + n^{-\frac{1}{2}}q_1(-x)\phi(-x) + n^{-1}q_2(-x)\phi(-x) + o(n^{-\frac{3}{2}}) \right\} \\ &\approx \Phi(x) - \Phi(-x) + n^{-\frac{1}{2}}(q_1(x) - q_1(-x))\phi(x) + n^{-1}(q_2(x) - q_2(-x))\phi(x) \\ &\quad \text{by ignoring the lower terms} \\ &= \Phi(x) - \Phi(-x) + n^{-1} 2q_2(x) \phi(x), \end{aligned} \quad (7.2.11)$$

since

$$q_1(x) = q_1(-x),$$

$$q_2(-x) = -q_2(x),$$

$$\text{and } \phi(x) = \phi(-x).$$

Using (7.2.11), Rao, Jocelyn and Hidirolou (2001) obtained the following Edgeworth expansion for the coverage error of the $(1 - \alpha)$ -level normal interval on $\bar{Y}$, $I_{1-\alpha} = [\bar{y} - z_{\alpha/2}s_{\bar{y}}, \bar{y} + z_{\alpha/2}s_{\bar{y}}]$, as

$$\begin{aligned} \text{CE} &= P(\bar{Y} \in I_{1-\alpha}) - (1 - \alpha) \\ &\approx 2 q_2(z_{\alpha/2}) \phi(z_{\alpha/2}) \\ &= \frac{2z_{\alpha/2}}{n} \left\{ \frac{1}{12} \kappa(z_{\alpha/2}^2 - 3) - \frac{1}{18} \gamma^2(z_{\alpha/2}^4 + 2z_{\alpha/2}^2 - 3) \right. \\ &\quad \left. - \frac{1}{4}(z_{\alpha/2}^2 + 3) \right\} \phi(z_{\alpha/2}), \end{aligned} \tag{7.2.12}$$

where $z_{\alpha/2}$ and $\phi(\cdot)$ are the upper $\alpha/2$ -point and the probability density function of a $N(0, 1)$ variable respectively. We note that the coverage error CE is of the order n^{-1} . As noted in Rao et. al. (2001), for $\alpha = 0.05$ such that $z_{\alpha/2} \approx 2$, it follows from (7.2.12) that the actual coverage probability will be smaller than the nominal level if

$$\kappa < 7(2\gamma^2 + 3), \tag{7.2.13}$$

which means that a large skewness coefficient, γ , can lead to coverage probabilities substantially smaller than $1 - \alpha = 0.95$.

7.3 Alternative Approximate Pivotal Statistic

Godambe and Thompson (1999) suggested the following approximate pivotal statistic

$$\tilde{S}_t = n^{\frac{1}{2}} \tilde{H}(\mathbf{T}_n) \quad (7.3.1)$$

instead of the studentized t -statistic, where

$$\tilde{H}(\mathbf{T}_n) = \frac{\bar{g}}{\sqrt{\bar{g}_2}} = \frac{\bar{g}}{\sqrt{\hat{\beta}_a^2}} \quad (7.3.2)$$

with $\hat{\beta}_a^2 = \frac{1}{n} \sum_i g_i^2(y_i, \theta)$, since $E(\hat{\beta}_a^2) = E(\frac{1}{n} \sum_i g_i^2(y_i, \theta)) = \beta$. Defining

$$\tilde{a}_{i_1, \dots, i_j} = \frac{\partial^j}{\partial x^{(i_1)} \dots \partial x^{(i_j)}} \tilde{H}(\mathbf{x}) \Big|_{\mathbf{x}=\mathbf{E}(\mathbf{x})} \quad (7.3.3)$$

we have

$$\begin{aligned} \tilde{a}_1 &= \frac{\partial}{\partial \bar{g}} \tilde{H}(\mathbf{T}_n) \Big|_{\mathbf{T}_n=\mathbf{E}(\mathbf{T}_n)} = \frac{1}{\beta}, \\ \tilde{a}_{12} &= \frac{\partial^2}{\partial \bar{g} \partial \hat{\beta}_a^2} \tilde{H}(\mathbf{T}_n) \Big|_{\mathbf{T}_n=\mathbf{E}(\mathbf{T}_n)} = -\frac{1}{2\beta^3} = \tilde{a}_{21}, \\ \text{and } \tilde{a}_{11} &= \tilde{a}_2 = \tilde{a}_{22} = 0. \end{aligned}$$

The functions $\mu_{i_1, \dots, i_j}$ remain unchanged and thus

$$\tilde{q}_1(x) = \frac{1}{6} \gamma(2x^2 + 1), \quad (7.3.4)$$

which is identical to $q_1(x)$. To find the second polynomial we use the delta method, discussed in Chapter 2 of Hall (1992). First, we write

$$\begin{aligned} \tilde{S}_t &= n^{\frac{1}{2}} \frac{\bar{g}}{(\hat{\beta}_a^2)^{\frac{1}{2}}} \\ &\approx n^{\frac{1}{2}} \bar{g} \left\{ \frac{1}{\beta} - \left(\frac{1}{2} \right) (\hat{\beta}_a^2 - \beta^2) (\beta^2)^{-\frac{3}{2}} + \left(\frac{1}{4} \right) \left(\frac{3}{2} \right) (\hat{\beta}_a^2 - \beta^2)^2 (\beta^2)^{-\frac{5}{2}} + O_p(n^{-\frac{3}{2}}) \right\} \\ &= n^{\frac{1}{2}} \bar{g} \left\{ \frac{1}{\beta} - \frac{1}{2\beta^3} (\bar{g}_2 - \beta^2) + \frac{3}{8\beta^5} (\bar{g}_2 - \beta^2)^2 + O_p(n^{-\frac{3}{2}}) \right\}. \end{aligned} \quad (7.3.5)$$

Now

$$\begin{aligned}
E(\tilde{S}_t) &= -\frac{1}{2\beta^3} n^{-\frac{1}{2}} E\left(n^{\frac{1}{2}} \bar{g} n^{\frac{1}{2}} (\bar{g}_2 - \beta^2)\right) + O(n^{-1}) \\
&= -\frac{1}{2} n^{-\frac{1}{2}} \frac{\mu_3}{\beta^3} + O(n^{-1}) \\
&= -\frac{1}{2} n^{-\frac{1}{2}} \gamma + O(n^{-1}),
\end{aligned} \tag{7.3.6}$$

$$\begin{aligned}
E(\tilde{S}_t^2) &= E\left(n\bar{g}^2 \left\{ \frac{1}{\beta^2} - \frac{1}{\beta^4} (\bar{g}_2 - \beta^2) + \frac{1}{\beta^6} (\bar{g}_2 - \beta^2)^2 \right\}\right) + O(n^{-\frac{3}{2}}) \\
&= 1 - n^{-1}(\kappa + 2) + n^{-1}(2\gamma^2 + \kappa + 2) + O(n^{-\frac{3}{2}}) \\
&= 1 + 2n^{-1}\gamma^2 + O(n^{-\frac{3}{2}}),
\end{aligned} \tag{7.3.7}$$

$$\begin{aligned}
E(\tilde{S}_t^3) &= E\left(n^{\frac{3}{2}} \bar{g}^3 \left\{ \frac{1}{\beta^3} - \frac{3}{2\beta^5} (\bar{g}_2 - \beta^2) \right\}\right) + O(n^{-1}) \\
&= n^{-\frac{1}{2}} \gamma - \frac{9}{2} n^{-\frac{1}{2}} \gamma + O(n^{-1}) \\
&= -\frac{7}{2} n^{-\frac{1}{2}} \gamma + O(n^{-1}),
\end{aligned} \tag{7.3.8}$$

$$\begin{aligned}
E(\tilde{S}_t^4) &= E\left(n^2 \bar{g}^4 \left\{ \frac{1}{\beta^4} - \frac{2}{\beta^6} (\bar{g}_2 - \beta^2) + \frac{3}{\beta^8} (\bar{g}_2 - \beta^2)^2 \right\}\right) + O(n^{-\frac{3}{2}}) \\
&= (3 + n^{-1}\kappa) - 2n^{-1} \{6(\kappa + 2) + 4\gamma^2\} + 3n^{-1} \{3(\kappa + 2) + 12\gamma^2\} + O(n^{-\frac{3}{2}}) \\
&= 3 + n^{-1}(28\gamma^2 - 2\kappa - 6) + O(n^{-\frac{3}{2}}).
\end{aligned} \tag{7.3.9}$$

Using the notation in Hall (1992, p.73), we have

$$\kappa_{1,n} = E(\tilde{S}_t) = -\frac{1}{2} n^{-\frac{1}{2}} \gamma + O(n^{-\frac{3}{2}}),$$

$$\kappa_{2,n} = E(\tilde{S}_t^2) - (E(\tilde{S}_t))^2$$

$$\begin{aligned}
&= 1 + 2n^{-1}\gamma^2 - \frac{1}{4}n^{-1}\gamma^2 + O(n^{-2}) \\
&= 1 + \frac{1}{4}n^{-1}(7\gamma^2) + O(n^{-2}),
\end{aligned}$$

$$\begin{aligned}
\kappa_{3,n} &= E(\tilde{S}_t^3) - 3E(\tilde{S}_t^2)E(\tilde{S}_t) + 2(E(\tilde{S}_t))^3 \\
&= -\frac{7}{2}n^{-\frac{1}{2}}\gamma - 3(1 + 2n^{-1}\gamma^2) \left(-\frac{1}{2}n^{-\frac{1}{2}}\gamma\right) \\
&\quad + 2\left(-\frac{1}{2}n^{-\frac{1}{2}}\gamma\right)^3 + O(n^{-\frac{3}{2}}) \\
&= -\frac{7}{2}n^{-\frac{1}{2}}\gamma + \frac{3}{2}n^{-\frac{1}{2}}\gamma + O(n^{-\frac{3}{2}}) \\
&= -2n^{-\frac{1}{2}}\gamma + O(n^{-\frac{3}{2}}),
\end{aligned}$$

$$\begin{aligned}
\kappa_{4,n} &= E(\tilde{S}_t^4) - 4E(\tilde{S}_t^3)E(\tilde{S}_t) - 3(E(\tilde{S}_t))^2 \\
&\quad + 12E(\tilde{S}_t^2)(E(\tilde{S}_t))^2 - 6(E(\tilde{S}_t))^4 \\
&= 3 + n^{-1}(28\gamma^2 - 2\kappa - 6) - 4\left(-\frac{7}{2}n^{-\frac{1}{2}}\gamma\right)\left(-\frac{1}{2}n^{-\frac{1}{2}}\gamma\right) \\
&\quad - 3(1 + 2n^{-1}\gamma^2)^2 + 12(1 + 2n^{-1}\gamma^2)\left(-\frac{1}{2}n^{-\frac{1}{2}}\gamma\right)^2 \\
&\quad - 6\left(-\frac{1}{2}n^{-\frac{1}{2}}\gamma\right)^4 + O(n^{-2}) \\
&= n^{-1}(12\gamma^2 - 2\kappa - 6) + O(n^{-2}).
\end{aligned}$$

Then

$$k_{1,2} = -\frac{1}{2}\gamma, \quad k_{2,2} = \frac{7}{4}\gamma^2, \quad k_{3,1} = -2\gamma, \quad k_{4,1} = 12\gamma^2 - 2\kappa - 6,$$

and hence

$$\begin{aligned}
\tilde{q}_2(x) &= -x \left\{ \frac{1}{2}(k_{2,2} + k_{1,2}^2) + \frac{1}{24}(k_{4,1} + 4k_{1,2}k_{3,1})(x^2 - 3) \right. \\
&\quad \left. + \frac{1}{72}k_{3,1}^2(x^4 - 10x^2 + 15) \right\}
\end{aligned}$$

$$\begin{aligned}
&= -x \left\{ \frac{1}{2} \left(\frac{7}{4} \gamma^2 + \frac{1}{4} \gamma^2 \right) + \frac{1}{24} \left(12\gamma^2 - 2\kappa - 6 + 4 \left(-\frac{1}{2} \gamma \right) (-2\gamma) \right) (x^2 - 3) \right. \\
&\quad \left. + \frac{1}{72} (-2\gamma)^2 (x^4 - 10x^2 + 15) \right\} \\
&= x \left\{ \frac{1}{12} \kappa (x^2 - 3) - \frac{1}{18} \gamma^2 (x^4 + 2x^2 - 3) + \frac{1}{4} (x^2 - 3) \right\}. \tag{7.3.10}
\end{aligned}$$

Therefore, the Edgeworth expansion for the coverage error of the $(1 - \alpha)$ -level normal interval on $\bar{Y}$, $I_{1-\alpha} = [\bar{y} - z_{\alpha/2} \hat{\beta}_a, \bar{y} + z_{\alpha/2} \hat{\beta}_a]$, is given by

$$\begin{aligned}
CE_{GT} &= P(-z_{\alpha/2} \leq \tilde{S}_t \leq z_{\alpha/2}) - [\Phi(z_{\alpha/2}) - \Phi(-z_{\alpha/2})] \\
&= P(\bar{Y} \in I_{1-\alpha}) - (1 - \alpha) \\
&\approx 2 \tilde{q}_2(z_{\alpha/2}) \phi(z_{\alpha/2}) \\
&= \frac{2z_{\alpha/2}}{n} \left\{ \frac{1}{12} \kappa (z_{\alpha/2}^2 - 3) - \frac{1}{18} \gamma^2 (z_{\alpha/2}^4 + 2z_{\alpha/2}^2 - 3) \right. \\
&\quad \left. + \frac{1}{4} (z_{\alpha/2}^2 - 3) \right\} \phi(z_{\alpha/2}) \tag{7.3.11}
\end{aligned}$$

using equation (7.2.11) with S_t , x and q_2 replaced by $\tilde{S}_t$, x and $\tilde{q}_2$ respectively.

We note from (7.3.11) that the actual coverage probability, for $\alpha = 0.05$ such that $z_{\alpha/2} \approx 2$, will be smaller than the nominal level if

$$\kappa < 14\gamma^2 - 3. \tag{7.3.12}$$

We also note that if $\kappa < -3$, then (7.3.12) is satisfied for any γ . We can see from (7.2.13) and (7.3.12) that the Godambe/Thompson statistic has smaller undercoverage area than the studentized t -statistic.

7.4 Comparison of Coverage Probabilities

The coverage probabilities for the studentized t -statistic and the Godambe/Thompson statistic are given in Table 7.1. We notice that when both κ and γ equal zero, the studentized t -statistic will underestimate the nominal level of $1 - \alpha = 0.95$ while the Godambe/Thompson statistic will overestimate the nominal level. When only γ equals zero the Godambe/Thompson statistic is closer to the nominal level than the studentized t -statistic. As the skewness term $|\gamma|$ becomes large, both statistics will give coverage probabilities substantially smaller than $1 - \alpha = 0.95$. However, the Godambe/Thompson statistic gives smaller undercoverage than the studentized t -statistic.

Table 7.1: Coverage Probabilities (%) of the confidence intervals for the mean: nominal level = 0.95

Skewness	Kurtosis	Sample size					
$ \gamma $	κ	$n = 10$		$n = 30$		$n = 50$	
		ST	GT	ST	GT	ST	GT
0	-3	90.60	95.00	93.53	95.00	94.12	95.00
	0	91.08	95.48	93.69	95.16	94.22	95.10
	3	91.56	95.96	93.85	95.32	94.31	95.19
2	-3	80.70	85.10	90.23	91.70	92.14	93.02
	0	81.18	85.58	90.39	91.86	92.24	93.12
	3	81.67	86.07	90.56	92.02	92.33	93.21
4	-3	51.01	55.41	80.34	81.80	86.20	87.08
	0	51.49	55.89	80.50	81.96	86.30	87.18
	3	51.97	56.38	80.66	82.13	86.39	87.28

ST = Studentized t -statistic

GT = Godambe/Thompson statistic

Part III

Combined Classification

Chapter 8

A Review of Classification

8.1 Introduction

In this chapter we give a review of statistical classification problems where one is interested in predicting the unknown nature of an observation, based on the available data. Section 8.2 describes the statistical classification setup where classifiers and misclassification error rates are defined. Some commonly used classification rules are given in Section 8.3. Section 8.4 outlines error estimation methods for data-based classifiers while empirical risk minimization is discussed in Section 8.5. We then introduce the theory of Vapnik-Chervonenkis inequalities in Section 8.6 and Section 8.7. Finally, Section 8.8 gives an overview of random partitioning of the data and discusses the idea of splitting the data into a training sample and a testing set.

8.2 Statistical Classification Setup

In pattern recognition or classification, we are interested in guessing or predicting the unknown nature of an observation, usually a discrete quantity such as one or zero, good or bad, healthy or sick, a group of diseases, etc. This observation is a d -dimensional vector $\mathbf{X}$, also known as the feature or predictor vector, of observable measurements. The unknown nature of $\mathbf{X}$ is represented by Y , the class membership with values in $\{1, \dots, K\}$, assuming that a class structure already exists and there are K classes. More formally, we consider the random pair $(\mathbf{X}, Y)$, where $\mathbf{X} \in \mathbb{R}^d$ is observable and $Y \in \{1, \dots, K\}$ is to be predicted.

Definition 8.1 (Classifier) *A classifier ψ is any map of the form $\psi : \mathbb{R}^d \rightarrow \{1, \dots, K\}$, which is used to predict Y given the feature vector $\mathbf{X}$ [i.e. $\psi(\mathbf{X}) \in \{1, \dots, K\}$ is the predicted value of Y given $\mathbf{X}$].*

An error occurs if $\psi(\mathbf{X}) \neq Y$ and the misclassification error rate of a classifier ψ , denoted by $\text{err}(\psi)$, is given by

$$\text{err}(\psi) = \mathbf{P}(\psi(\mathbf{X}) \neq Y). \quad (8.2.1)$$

A best possible classifier, denoted by ψ_B and often called the Bayes' classifier, is defined to satisfy

$$\text{err}(\psi_B) = \mathbf{P}(\psi_B(\mathbf{X}) \neq Y) = \inf_{\psi: \mathbb{R}^d \rightarrow \{1, \dots, K\}} \mathbf{P}(\psi(\mathbf{X}) \neq Y). \quad (8.2.2)$$

We note that ψ_B depends on the distribution of $(\mathbf{X}, Y)$ which, in many practical problems, is unknown and therefore ψ_B is also unknown. However, we do have access

to some data set $T_N = \{(\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_N, Y_N)\}$ where $(\mathbf{X}_i, Y_i)$, $i = 1, \dots, N$ are independent and identically distributed (i.i.d.) $\mathbb{R}^d \times \{1, \dots, K\}$ - valued random vectors, which may have been observed in the past through experiments or some other sources. We also assume that the random vectors in the data set $\{(\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_N, Y_N)\}$ have the same distribution as the new observation $(\mathbf{X}, Y)$. Suppose we have access to a training sample of size n and a testing set of size m , where $n + m = N$. In particular, we define $T_n = \{(\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_n, Y_n)\}$ and $T_m = \{(\mathbf{X}_{n+1}, Y_{n+1}), \dots, (\mathbf{X}_{n+m}, Y_{n+m})\}$ as the training sample and the testing set respectively, where $(\mathbf{X}_{n+j}, Y_{n+j})$, $j = 1, \dots, m$ are independent of T_n and $(\mathbf{X}, Y)$.

Given that the distribution of $(\mathbf{X}, Y)$ is unknown in many practical problems, we must find a sequence of data-based classifiers $\{\psi_n, n \geq 1\}$, that is, a classification rule based on the training sample of size n , in such a way that the conditional error rate

$$\text{err}_n(\psi_n) = \mathbf{P}(\psi_n(\mathbf{X}) \neq Y \mid T_n) \quad (8.2.3)$$

is, in some sense, as small as possible.

Definition 8.2 (Data-based Classifier) Let $T_n = \{(\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_n, Y_n)\}$ be a training sample of n i.i.d. $\mathbb{R}^d \times \{1, \dots, K\}$ - valued random vectors, then a data-based classifier ψ_n is a mapping of the form $\psi_n : \mathbb{R}^d \times \{\mathbb{R}^d \times \{1, \dots, K\}\}^n \rightarrow \{1, \dots, K\}$.

Definition 8.3 (Weak and Strong Consistency) A classification rule $\{\psi_n, n \geq 1\}$ is (weakly) consistent for a certain distribution of $(\mathbf{X}, Y)$ if

$$\text{err}_n(\psi_n) \xrightarrow{P} \text{err}(\psi_B) \text{ as } n \rightarrow \infty.$$

A classification rule $\{\psi_n\}$ is strongly consistent if

$$\text{err}_n(\psi_n) \xrightarrow{\text{a.s.}} \text{err}(\psi_B) \text{ as } n \rightarrow \infty.$$

Remarks:

- (i) A consistent classification rule guarantees that by increasing the amount of training data the probability that $\text{err}_n(\psi_n)$ is within a very small distance of $\text{err}(\psi_B)$ gets arbitrarily close to one. This would mean that the classification rule would eventually learn the optimal decision from a large amount of training data with high probability.
- (ii) Strong consistency means that by using more data $\text{err}_n(\psi_n)$ gets arbitrarily close to $\text{err}(\psi_B)$ for every training sample except for a set of sequences that has zero probability altogether.
- (iii) A classification rule is said to be universally consistent if it is consistent for all distributions of $(\mathbf{X}, Y)$.

8.3 Classification Rules

In this section we will define some commonly used classification rules for the case where there are only two classes, namely $Y \in \{0, 1\}$. The book by Devroye, Györfi and Lugosi (1996) provides an excellent review of many important classification rules.

8.3.1 Linear Discriminant Rules

For the feature vector $\mathbf{X} \in \Re^d$, the linear discriminant rules split the space by a hyperplane and assign a different class to each halfspace, based on the weights associated with each component. More formally, the linear discriminant rule is given by

$$\psi(\mathbf{x}) = \begin{cases} 1 & \text{if } a_0 + \mathbf{a}^T \mathbf{x} > 0 \\ 0 & \text{otherwise.} \end{cases} \quad (8.3.1)$$

where $\mathbf{x} = (x_1, \dots, x_d)^T$ and $a_0, \mathbf{a} = (a_1, \dots, a_d)^T$ are the weights. These weights determine the relative importance of the components. The procedure for implementing linear discriminant rules is proportional to the dimension d of $\mathbf{X}$ and it is quite simple to implement. However, in general, the linear discriminant rules cannot provide error probabilities that are close to the Bayes error probability. In the case of $\Re^2$, the linear discriminant rule ψ defines a separating line. Fisher's (1936) Linear Discriminant Function is one example of a linear discriminant rule.

It is worth noting that these linear discriminant rules may be arbitrarily bad for distributions in which the two classes are linearly separable and yet the error probabilities of these types of classifiers are close to one. The multivariate normal distribution, however, is an exception in which the Bayes rule is a linear discriminant rule. Define the general multivariate normal density of a random variable $\mathbf{X}$ by

$$f(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} e^{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu})}$$

where $\boldsymbol{\mu} = \mathbf{E}(\mathbf{X})$ is the mean ($\boldsymbol{\mu}$ is a d -dimensional column vector), and $\Sigma = \mathbf{E}\{(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^T\}$ is the $d \times d$ covariance matrix. Let $f_0(\mathbf{x})$ and $f_1(\mathbf{x})$ be the conditional

densities given $\{Y = 0\}$ and $\{Y = 1\}$ respectively. In this case, the Bayes rule is given by (8.3.1) where, by letting $p = \mathbf{P}\{Y = 1\}$ and $1 - p = \mathbf{P}\{Y = 0\}$,

$$a_0 = 2 \log \left(\frac{p}{1-p} \right) + \boldsymbol{\mu}_0^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_0 - \boldsymbol{\mu}_1^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_1$$

$$\text{and } \mathbf{a} = (a_1, \dots, a_d)^T = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^T \boldsymbol{\Sigma}^{-1},$$

assuming $\boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}$. In practice though, p , $\boldsymbol{\mu}_0$, $\boldsymbol{\mu}_1$, $\boldsymbol{\Sigma}_0$, $\boldsymbol{\Sigma}_1$ are unknown and have to be estimated from the data.

8.3.2 Nearest Neighbour Rules

Suppose $T_N = \{(\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_N, Y_N)\}$ is the data set of N i.i.d. $\mathbb{R}^d \times \{1, \dots, K\}$ -valued random vectors with distribution $F_{\mathbf{X}, Y}$. Let $(\mathbf{X}, Y)$ be a new observation from $F_{\mathbf{X}, Y}$ for which Y is to be predicted and define $\|\mathbf{X}_i - \mathbf{X}\|$ as the Euclidean distance between $\mathbf{X}_i$ and $\mathbf{X}$. Then $\mathbf{X}_i$ is said to be the k^{th} nearest neighbour of $\mathbf{X}$ if $\|\mathbf{X}_i - \mathbf{X}\|$ is the k^{th} smallest among $\|\mathbf{X}_1 - \mathbf{X}\|, \dots, \|\mathbf{X}_N - \mathbf{X}\|$. In the case of a distance tie, the candidate with the smaller index is said to be closer to $\mathbf{X}$. Let $T_{(N)} = \{(\mathbf{X}_{(1)}, Y_{(1)}), \dots, (\mathbf{X}_{(N)}, Y_{(N)})\}$ be the re-ordered data sequence, where $\mathbf{X}_{(i)} = \mathbf{X}_{(i)}(\mathbf{X})$, $i = 1, \dots, N$, is the i^{th} nearest neighbour of $\mathbf{X}$ among the points $\mathbf{X}_1, \dots, \mathbf{X}_N$ and $Y_{(i)} = Y_{(i)}(\mathbf{X})$, $i = 1, \dots, N$.

The k -NN (k nearest neighbour) rule, proposed by Fix and Hodges (1951), is then defined by

$$\psi_N(\mathbf{x}) = \begin{cases} 1 & \text{if } \sum_{i=1}^k \mathbf{I}_{\{Y_{(i)}(\mathbf{x})=1\}} > \sum_{i=1}^k \mathbf{I}_{\{Y_{(i)}(\mathbf{x})=0\}} \\ 0 & \text{otherwise.} \end{cases} \quad (8.3.2)$$

We will note here that the k -NN rule is based on a majority vote among the labels of the k nearest neighbours of $\mathbf{x}$. It is convenient to let k be odd, in order to avoid voting ties.

We can show, by an application of Stone's theorem (Stone, 1977), that the k -NN classification rule is weakly universally consistent if k is allowed to grow with N in such a way that $k/N \rightarrow 0$.

Theorem 8.1 (Stone, 1977) *If $k \rightarrow \infty$ and $k/N \rightarrow 0$, then for all distributions $\mathbf{E}\{\text{err}_N(\psi_N)\} \rightarrow \text{err}(\psi_B)$.*

Strong consistency of the k -NN classification rule is given in Devroye et. al. (1996, Theorem 11.1), under the assumption that $\mathbf{X}$ has a density. This assumption is needed to avoid the problems of breaking distance ties.

8.3.3 Kernel Rules

Kernel rules are derived from the kernel estimate in density estimation. For these methods, the density at a point is estimated from a cell located at that point, where the volume of the cell is fixed and we find the number of data points in this volume.

Let $\psi_n(\mathbf{x})$ be a classifier, based on the training sample T_n , that takes the data points within a certain distance of the point to be classified, and classify that point according to a majority vote. That is, let h be a positive number and define the moving window rule (also known as the naive kernel) as

$$\psi_n(\mathbf{x}) = \begin{cases} 1 & \text{if } \sum_{i=1}^n \mathbf{I}_{\{Y_i=1, \mathbf{X}_i \in S_{\mathbf{x},h}\}} > \sum_{i=1}^n \mathbf{I}_{\{Y_i=0, \mathbf{X}_i \in S_{\mathbf{x},h}\}} \\ 0 & \text{otherwise.} \end{cases} \quad (8.3.3)$$

where $S_{\mathbf{x},h}$ denote the closed ball of radius $h > 0$ centered at $\mathbf{x}$.

Let $\mathbf{K} : \mathbb{R}^d \rightarrow \mathbb{R}$ be a kernel function, where $\mathbf{K} \geq 0$ and $\mathbf{K}(\mathbf{t})$ is monotone decreasing in $\|\mathbf{t}\|$. Then, the kernel classification rule, which is a generalization of the moving window rule in (8.3.3), is given by

$$\psi_n(\mathbf{x}) = \begin{cases} 1 & \text{if } \sum_{i=1}^n \mathbf{I}_{\{Y_i=1\}} \mathbf{K}\left(\frac{\mathbf{x}-\mathbf{X}_i}{h}\right) > \sum_{i=1}^n \mathbf{I}_{\{Y_i=0\}} \mathbf{K}\left(\frac{\mathbf{x}-\mathbf{X}_i}{h}\right) \\ 0 & \text{otherwise.} \end{cases} \quad (8.3.4)$$

We will note that the kernel classification rule gives more weight to the points that are closer to the point that needs to be classified than to more distant ones. This form of distance weighting, denoted by h , is called the smoothing factor (parameter) or bandwidth. Some popular kernels include the one-dimensional normal kernel, $\mathbf{K}(x) = (1/\sqrt{2\pi})e^{-(x^2/2)}$; the Gaussian kernel, $\mathbf{K}(\mathbf{x}) = e^{-\|\mathbf{x}\|^2}$; and the Cauchy kernel, $\mathbf{K}(\mathbf{x}) = 1/(1 + \|\mathbf{x}\|^{d+1})$ where $\|\cdot\|$ denotes the Euclidean distance. The following definition of a regular kernel appears on page 149 of Devroye et. al. (1996).

Definition 8.4 (Regular Kernel) *The kernel $\mathbf{K}$ is called regular if $\mathbf{K} \geq 0$ and there is a ball $S_{0,r}$ and constant $b > 0$ such that $\mathbf{K}(\mathbf{x}) \geq b\mathbf{I}_{S_{0,r}}$ and $\int \sup_{\mathbf{y} \in \mathbf{x}+S_{0,r}} \mathbf{K}(\mathbf{y}) d\mathbf{x} < \infty$.*

We note that, in all cases, regular kernels are bounded and integrable. Moreover, the second condition in Definition 8.4 holds whenever $\mathbf{K}$ is integrable and uniformly continuous.

Theorem 8.2 (Devroye and Krzyzak (1989)) *Suppose $\mathbf{K}$ is a regular kernel. If $h \rightarrow 0$ and $nh^d \rightarrow \infty$ as $n \rightarrow \infty$, then for any distribution of $(\mathbf{X}, Y)$, and for every*

$\epsilon > 0$ there is an integer n_0 such that for $n > n_0$,

$$\mathbf{P} \{ \text{err}_n(\psi_n) - \text{err}(\psi_B) > \epsilon \} \leq 2 e^{-n\epsilon^2 / (32\rho^2)},$$

where the constant $\rho = \rho(\mathbf{K}, d)$. Thus, the kernel classification rule is strongly universally consistent.

Theorem 8.2 demonstrates the strong universal consistency of kernel-based classification rules for all regular kernels, under general conditions on h and $\mathbf{K}$.

Kernel rules and nearest neighbour rules make minimal assumptions about the form of the underlying distribution. However, the following limitations, as noted in Breiman et. al. (1984), are common to both methods:

- They are sensitive to the choice of the metric $\|\mathbf{X}\|$.
- There is no natural or simple way to handle categorical variables or missing data.
- They are computationally expensive to implement, as the data set T_N must be stored, $\|\mathbf{X}_i - \mathbf{X}\|$ and $\psi_N(\mathbf{X})$ recomputed for the new point $\mathbf{X}$.
- They give very little usable information regarding the structure of the data.

8.3.4 Classification Trees

Classification trees partition $\Re^d$ into regions, often hyperrectangles parallel to the axes. Among these, the most important are the binary classification trees which are constructed by repeated splits of subsets of $\Re^d$ into two descendant subsets, beginning

Figure 8.1: Binary tree

with $\mathfrak{R}^d$ itself. This process is pictured in Figure 8.1 where the top of the binary tree is called the root. Each node has either no child (in which case it is called a terminal node or leaf), a left child, a right child, or a left child and a right child. Each node is the root of a tree itself. The trees rooted at the children of a node are called the left and right subtrees of that node. The depth of a node is the length of the path from that node to the root. The height of a tree is the maximal depth of any node.

Each node in Figure 8.1 represents a set in the space $\mathfrak{R}^d$ and has exactly two or zero children. If a node u represents the set A and its children u', u'' represent A' and A'' respectively, then we require that $A = A' \cup A''$ and $A' \cap A'' = \emptyset$. The root represents $\mathfrak{R}^d$, and the leaves, taken together, form a partition of $\mathfrak{R}^d$.

The construction of a tree then revolves around the following three elements:

1. The selection of the splits.
2. The decisions when to declare a node terminal (that is, a leaf) or to continue splitting it.
3. The assignment of each leaf to a class.

The tree structure then uses the data set T_N to determine the splits, the leaves, and their assignments to the classes. For the two class problem, if a leaf represents region A , then we say that the classifier ψ_N is natural if

$$\psi_N(\mathbf{x}) = \begin{cases} 1 & \text{if } \sum_{i: \mathbf{X}_i \in A} Y_i > \sum_{i: \mathbf{X}_i \in A} (1 - Y_i), \quad \mathbf{x} \in A \\ 0 & \text{otherwise.} \end{cases}$$

That is, in every leaf region, a majority vote is taken over all $(\mathbf{X}_i, Y_i)$'s with $\mathbf{X}_i$ in the same region. Ties are broken, as usual, in favour of class 0. We note that natural tree classifiers are special cases of data-dependant partitioning rules discussed in Section 8.8.

Classification trees have the advantage of being fully dynamic because as new data are added, the tree may be updated slightly, say, by splitting a leaf or two, to obtain an updated classifier. Moreover, binary tree classifiers are used to explain complicated data (where some components may be nonnumerical in nature, others may represent integers, and still others may be real numbers) and have a classifier that is easy to analyze and understand.

8.4 Error Estimation

Given a data-based classifier ψ_n and the fact that the distribution of $(\mathbf{X}, Y)$ is unknown, we would like to know the performance of this classifier in terms of error estimation. In this regard, we would like to search for error estimation methods that have distribution-free performance bounds, whenever possible.

Suppose we can split our data into a training sample T_n and a testing set T_m defined in Section 8.2. Then we can estimate $\text{err}_n(\psi_n)$ by

$$\widehat{\text{err}}_N(\psi_n) = \frac{1}{m} \sum_{j=1}^m \mathbf{I}\{\psi_n(\mathbf{X}_{n+j}) \neq Y_{n+j}\}, \quad (8.4.1)$$

which represents the average number of errors committed on the testing set T_m . For this estimator of $\text{err}_n(\psi_n)$ we note the following:

- (i) $\widehat{\text{err}}_N(\psi_n)$ is unbiased in the sense that $\mathbf{E}(\widehat{\text{err}}_N(\psi_n) \mid T_n) = \text{err}_n(\psi_n)$ and
- (ii) The conditional distribution of $\mathbf{I}\{\psi_n(\mathbf{X}_{n+j}) \neq Y_{n+j}\}$, $j = 1, \dots, m$, given the training sample T_n , is Bernoulli ($\text{err}_n(\psi_n)$). So the conditional distribution of $m \cdot \widehat{\text{err}}_N(\psi_n)$, given T_n , is Binomial (m , $\text{err}_n(\psi_n)$).

Hence we can obtain a performance bound for $\widehat{\text{err}}_N(\psi_n)$ by noting that

$$\mathbf{E} \left\{ (\widehat{\text{err}}_N(\psi_n) - \text{err}_n(\psi_n))^2 \mid T_n \right\} = \frac{\text{err}_n(\psi_n) (1 - \text{err}_n(\psi_n))}{m} \leq \frac{1}{4m}$$

which does not depend on ψ_n . It is worth noting that one would usually want to incorporate all available data $\{(\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_N, Y_N)\}$ in the classification rule and an additional independent testing set is rarely available, so that equation (8.4.1) can be used as an estimator of the misclassification error rate. In order to estimate

$\text{err}_n(\psi_n)$ in such cases, we will have to rely on the training sample T_n only and assume that we do not have access to a testing set. However, we may be able to limit the classifiers to a certain class Ψ and select the best classifier from this class. This approach will be discussed in Section 8.5.

8.5 Empirical Risk Minimization

Suppose we want to pick a classifier ψ , from a class of classifiers Ψ , that minimizes an estimate of the error probability over the class. We also assume that the training sample is used to pick this classifier. If we can estimate the error probability for the classifiers in Ψ uniformly well, then the classifier that minimizes the estimated error probability may likely have an error probability that is close to the best classifier in the class.

For the two class problem, let $\psi : \mathbb{R}^d \rightarrow \{1, 0\}$ be a classifier from a class Ψ . We assume that the empirical error probability

$$\widehat{\text{err}}_n(\psi) = \frac{1}{n} \sum_{i=1}^n \mathbf{I}\{\psi(\mathbf{X}_i) \neq Y_i\}, \quad (8.5.1)$$

is used to estimate the error probability $\text{err}(\psi) = \mathbf{P}(\psi(\mathbf{X}) \neq Y)$ of each classifier $\psi \in \Psi$. Let ψ_n^* be the classifier picked from the class Ψ that minimizes the empirical error probability in (8.5.1). In other words, ψ_n^* is the classifier where

$$\widehat{\text{err}}_n(\psi_n^*) \leq \widehat{\text{err}}_n(\psi) \quad \text{for all } \psi \in \Psi.$$

This selected classifier, ψ_n^* , should be good in the sense that its error probability,

defined by

$$\text{err}_n(\psi_n^*) = \mathbf{P}(\psi_n^*(\mathbf{X}) \neq Y \mid T_n),$$

is expected to be close to the optimal error probability within the class Ψ .

Lemma 8.1 (Vapnik and Chervonenkis, 1974) *For the selected classifier ψ_n^* we have the following inequalities,*

$$\begin{aligned} \text{err}_n(\psi_n^*) - \inf_{\psi \in \Psi} \text{err}(\psi) &\leq 2 \sup_{\psi \in \Psi} | \widehat{\text{err}}_n(\psi) - \text{err}(\psi) | \\ \text{and } | \widehat{\text{err}}_n(\psi_n^*) - \text{err}_n(\psi_n^*) | &\leq \sup_{\psi \in \Psi} | \widehat{\text{err}}_n(\psi) - \text{err}(\psi) |. \end{aligned}$$

Lemma 8.2 (Hoeffding's Inequality; Hoeffding, 1963) *Let $X_1, \dots, X_n$ be independent bounded random variables such that $X_i \in [a_i, b_i]$ with probability one and let $S_n = \sum_{i=1}^n X_i$. Then for any $\epsilon > 0$*

$$\mathbf{P} (| S_n - \mathbf{E}S_n | \geq \epsilon) \leq 2 e^{-2\epsilon^2 / \sum_{i=1}^n (b_i - a_i)^2}. \quad (8.5.2)$$

Theorem 8.3 (Finite Cardinality of the Class) *Suppose that the cardinality of the class of classifiers Ψ is bounded by N . Then for every $\epsilon > 0$,*

$$\mathbf{P} \left\{ \text{err}_n(\psi_n^*) - \inf_{\psi \in \Psi} \text{err}(\psi) > \epsilon \right\} \leq 2Ne^{-n\epsilon^2 / 2}. \quad (8.5.3)$$

Proof

$$\begin{aligned} &\mathbf{P} \left\{ \text{err}_n(\psi_n^*) - \inf_{\psi \in \Psi} \text{err}(\psi) > \epsilon \right\} \\ &\leq \mathbf{P} \left\{ \sup_{\psi \in \Psi} | \widehat{\text{err}}_n(\psi) - \text{err}(\psi) | > \frac{\epsilon}{2} \right\} \text{ by Lemma 8.1.} \\ &\leq \sum_{\psi \in \Psi} \mathbf{P} \left\{ | \widehat{\text{err}}_n(\psi) - \text{err}(\psi) | > \frac{\epsilon}{2} \right\} \end{aligned}$$

$$\begin{aligned}
&= \sum_{\psi \in \Psi} \mathbf{P} \left\{ \left| \frac{1}{n} \sum_{i=1}^n \mathbf{I}\{\psi(\mathbf{X}_i) \neq Y_i\} - \mathbf{P}\{\psi(\mathbf{X}) \neq Y\} \right| > \frac{\epsilon}{2} \right\} \\
&= \sum_{\psi \in \Psi} \mathbf{P} \left\{ \left| \sum_{i=1}^n W_i - n\mathbf{E}(W_i) \right| > \frac{n\epsilon}{2} \right\} \\
&\quad \text{where } W_i = \mathbf{I}\{\psi(\mathbf{X}_i) \neq Y_i\} \\
&\leq \sum_{\psi \in \Psi} 2e^{-2(\frac{n\epsilon}{2})^2 / n} \quad \text{by Lemma 8.2} \\
&= 2Ne^{-n\epsilon^2 / 2}.
\end{aligned}$$

■

We note here that the right hand side of equation (8.5.3) goes to 0 at an exponential rate.

Lemma 8.1 provides upper bounds for two things; an upper bound for the suboptimality of our selected classifier ψ_n^* within the class Ψ , and an upper bound for the error committed when $\widehat{\text{err}}_n(\psi_n^*)$ is used to estimate the probability of error $\text{err}_n(\psi_n^*)$ of this selected classifier. Hoeffding's Inequality (Lemma 8.2) is used to obtain distribution-free performance bounds. Theorem 8.3 gives the distribution-free upper bounds when the class Ψ contains finitely many classifiers. That is, the actual distribution of the data does not play a role at all in the upper bounds.

Since

$$\sum_{n=1}^{\infty} \mathbf{P} \left\{ \text{err}_n(\psi_n^*) - \inf_{\psi \in \Psi} \text{err}(\psi) > \epsilon \right\} \leq 2N \sum_{n=1}^{\infty} e^{-n\epsilon^2 / 2} < \infty \quad \forall \epsilon > 0,$$

then by the first Borel-Cantelli lemma

$$\text{err}_n(\psi_n^*) \xrightarrow{\text{a.s.}} \inf_{\psi \in \Psi} \text{err}(\psi).$$

Suppose our class of classifiers Ψ is of the form $\psi_n = \psi_n(x, T_n)$ and the same training sample T_n is used to select a classification rule by minimizing some estimates

$\widehat{\text{err}}_n(\psi_n)$ of the error probabilities $\text{err}_n(\psi_n) = \mathbf{P}\{\psi_n(\mathbf{X}) \neq Y|T_n\}$. Then, if ψ_n^* denotes the selected classification rule, we replace ψ in Lemma 8.1 by ψ_n and obtain a result similar to Theorem 8.3.

Theorem 8.4 (Without Testing Data) *If the cardinality of class Ψ is bounded by N and $\psi_n = \psi_n(x, T_n)$ then for all $\epsilon > 0$,*

$$\mathbf{P} \left\{ \text{err}_n(\psi_n^*) - \inf_{\psi_n \in \Psi} \text{err}_n(\psi_n) > \epsilon \right\} \leq N \sup_{\psi_n \in \Psi} \mathbf{P} \left\{ |\widehat{\text{err}}_n(\psi_n) - \text{err}_n(\psi_n)| > \frac{\epsilon}{2} \right\}. \quad (8.5.4)$$

8.6 Vapnik - Chervonenkis Distribution - Free Performance Bound

In order to obtain distribution-free performance bounds in Theorems 8.3 and 8.4, we will need to look at the notion of uniform deviations of relative frequencies from their probabilities, discussed in Chapters 12 and 13 of Devroye et. al. (1996).

Let ν be the probability measure of $(\mathbf{X}, Y)$ on $\mathbb{R}^d \times \{0, 1\}$, and let ν_n be the empirical measure based on the training sample T_n . That is, for a fixed measurable set $A \subset \mathbb{R}^d \times \{0, 1\}$, $\nu(A) = \mathbf{P}\{(\mathbf{X}, Y) \in A\}$ and $\nu_n(A) = \frac{1}{n} \sum_{i=1}^n \mathbf{I}\{(\mathbf{X}_i, Y_i) \in A\}$. Then the misclassification error rate of the classifier ψ is given by

$$\text{err}(\psi) = \nu(\{(\mathbf{x}, y) : \psi(\mathbf{x}) \neq y\}),$$

which is the ν -measure of the set of pairs $(\mathbf{x}, y) \in \mathbb{R}^d \times \{0, 1\}$ where $\psi(\mathbf{x}) \neq y$. In other words, $\text{err}(\psi)$ is the ν -measure of the set

$$\{\{\mathbf{x} : \psi(\mathbf{x}) = 1\} \times \{0\}\} \cup \{\{\mathbf{x} : \psi(\mathbf{x}) = 0\} \times \{1\}\}$$

and similarly,

$$\widehat{\text{err}}_n(\psi) = \nu_n(\{(\mathbf{x}, y) : \psi(\mathbf{x}) \neq y\}).$$

Thus,

$$\sup_{\psi \in \Psi} | \widehat{\text{err}}_n(\psi) - \text{err}(\psi) | = \sup_{A \in \mathcal{A}} | \nu_n(A) - \nu(A) | ,$$

where $\mathcal{A}$ is the collection of all sets

$$\{ \{ \mathbf{x} : \psi(\mathbf{x}) = 1 \} \times \{0\} \} \cup \{ \{ \mathbf{x} : \psi(\mathbf{x}) = 0 \} \times \{1\} \} , \quad \psi \in \Psi. \quad (8.6.1)$$

If the class of sets $\mathcal{A}$ is finite then the bound is given by

$$\mathbf{P} \left\{ \sup_{A \in \mathcal{A}} | \nu_n(A) - \nu(A) | > \epsilon \right\} \leq 2 |\mathcal{A}| e^{-2n\epsilon^2} \quad (8.6.2)$$

where $|\mathcal{A}|$ is the cardinality of the class $\mathcal{A}$. However, if $\mathcal{A}$ contains infinitely many sets then we can use the Vapnik-Chervonenkis inequalities to provide distribution-free upper bounds for random variables of the type

$$\sup_{A \in \mathcal{A}} | \nu_n(A) - \nu(A) | ,$$

which yield performance bounds for any classifier selected by minimizing the empirical error probability defined in (8.5.1).

In the one-dimensional setting we will define the Glivenko-Cantelli Theorem for the uniform almost sure convergence of the empirical distribution function to the true distribution function.

Theorem 8.5 (Glivenko-Cantelli Theorem) *Let $Z_1, \dots, Z_n$ be i.i.d. real-valued random variables with distribution function $F(z) = \mathbf{P}(Z_1 \leq z)$ and empirical distri-*

bution function defined by

$$F_n(z) = \frac{1}{n} \sum_{i=1}^n \mathbf{I}\{Z_i \leq z\}.$$

Then

$$\mathbf{P} \left\{ \sup_{z \in \mathfrak{R}} |F(z) - F_n(z)| > \epsilon \right\} \leq 8(n+1)e^{-n\epsilon^2/32}$$

and by the Borel-Cantelli lemma,

$$\sup_{z \in \mathfrak{R}} |F(z) - F_n(z)| \xrightarrow{\text{a.s.}} 0 \text{ as } n \rightarrow \infty.$$

A proof of Theorem 8.5 is given in Section 12.3 of Devroye et. al. (1996) where $\nu(A) = \mathbf{P}\{Z_1 \in A\}$ and $\nu_n(A) = \frac{1}{n} \sum_{i=1}^n \mathbf{I}\{Z_i \in A\}$ for all measurable sets $A \subset \mathfrak{R}$. Denoting $\mathcal{A}$ as the class of sets of the form $(-\infty, z]$ for $z \in \mathfrak{R}$ we have

$$\sup_{z \in \mathfrak{R}} |F(z) - F_n(z)| = \sup_{A \in \mathcal{A}} |\nu_n(A) - \nu(A)|.$$

To apply the Vapnik-Chervonenkis inequality, which is a generalization of Theorem 8.5, we will need the following definitions of shatter coefficient and Vapnik-Chervonenkis (VC) dimension. In the general setting, let $\mathbf{Z}_1, \dots, \mathbf{Z}_n$ be i.i.d. random vectors taking values in $\mathfrak{R}^d$ and denote $\nu(A) = \mathbf{P}(\mathbf{Z}_1 \in A)$ and $\nu_n(A) = (1/n) \sum_{i=1}^n \mathbf{I}\{\mathbf{Z}_i \in A\}$ for all measurable sets $A \subset \mathfrak{R}^d$.

Definition 8.5 (Shatter Coefficient) Let $\mathcal{A}$ be a collection of all measurable sets.

For $(\mathbf{z}_1, \dots, \mathbf{z}_n) \in \{\mathfrak{R}^d\}^n$, let

$$N_{\mathcal{A}}(\mathbf{z}_1, \dots, \mathbf{z}_n) = \text{number of different sets in } \{\{\mathbf{z}_1, \dots, \mathbf{z}_n\} \cap A; A \in \mathcal{A}\}.$$

Then the n^{th} shatter coefficient of $\mathcal{A}$ is

$$s(\mathcal{A}, n) = \max_{(\mathbf{z}_1, \dots, \mathbf{z}_n) \in \{\mathfrak{R}^d\}^n} N_{\mathcal{A}}(\mathbf{z}_1, \dots, \mathbf{z}_n). \quad (8.6.3)$$

We note that the shatter coefficient is the combinatorial quantity with the maximal number of different subsets of n points that can be picked out by the class of sets $\mathcal{A}$. The shatter coefficient is a measure of the richness of the class $\mathcal{A}$ and we also note the following:

1. $s(\mathcal{A}, n) \leq 2^n$ because there are 2^n subsets of a set with n elements.
2. If $N_{\mathcal{A}}(\mathbf{z}_1, \dots, \mathbf{z}_n) = 2^n$ for some $(\mathbf{z}_1, \dots, \mathbf{z}_n)$ then we say $\mathcal{A}$ shatters $(\mathbf{z}_1, \dots, \mathbf{z}_n)$,
3. If $s(\mathcal{A}, n) < 2^n$, then any set of n points has a subset such that there is no set in $\mathcal{A}$ that contains exactly that subset of the n points,
4. If $s(\mathcal{A}, k) < 2^k$ for some integer k , then $s(\mathcal{A}, n) < 2^n$ for all $n > k$.

Definition 8.6 (Vapnik-Chervonenkis Dimension) *Suppose the collection of sets $\mathcal{A}$ is such that $|\mathcal{A}| \geq 2$. Then the largest integer $k \geq 1$ for which $s(\mathcal{A}, k) = 2^k$, denoted by $V_{\mathcal{A}}$, is called the Vapnik-Chervonenkis (VC) dimension of the class $\mathcal{A}$. If $s(\mathcal{A}, n) = 2^n$ for all n , then by definition $V_{\mathcal{A}} = \infty$. Intuitively, $V_{\mathcal{A}}$ can be considered as the complexity of the class of sets $\mathcal{A}$.*

Theorem 8.6 (Vapnik - Chervonenkis Inequality) *For any probability measure ν and a class of sets $\mathcal{A}$, and for any n and $\epsilon > 0$,*

$$\mathbf{P} \left\{ \sup_{A \in \mathcal{A}} | \nu_n(A) - \nu(A) | > \epsilon \right\} \leq 8 s(\mathcal{A}, n) e^{-n\epsilon^2/32}. \quad (8.6.4)$$

The bound in Theorem 8.6 is useful when the shatter coefficient $s(\mathcal{A}, n)$ does not increase too quickly with n . In order to apply Theorem 8.6 we will need to compute

shatter coefficients for certain families of sets. Moreover, measurability of the supremum in (8.6.4) must be verified for every family $\mathcal{A}$. A proof of Theorem 8.6 can be found in Chapter 12 of Devroye et. al. (1996).

We will now use Theorem 8.6 to select any classifier ψ from a class of classifiers Ψ in such a way that the empirical error probability is minimized.

Theorem 8.7 (Selecting a Classifier) *Let Ψ be a class of classifiers of the form $\psi : \mathbb{R}^d \rightarrow \{0, 1\}$. Suppose ψ_n^* is the classifier, selected from the class of classifiers Ψ , which has minimal empirical error probability. Define the n^{th} shatter coefficient and the VC dimension of Ψ as $s(\Psi, n) = s(\mathcal{A}, n)$ and $V_\Psi = V_{\mathcal{A}}$ respectively, where $\mathcal{A}$ is defined by (8.6.1). Then*

$$\mathbf{P} \left\{ \sup_{\psi \in \Psi} | \widehat{\text{err}}_n(\psi) - \text{err}(\psi) | > \epsilon \right\} \leq 8 s(\Psi, n) e^{-n\epsilon^2/32} \quad (8.6.5)$$

and therefore by Lemma 8.1

$$\mathbf{P} \left\{ \text{err}_n(\psi_n^*) - \inf_{\psi \in \Psi} \text{err}(\psi) > \epsilon \right\} \leq 8 s(\Psi, n) e^{-n\epsilon^2/128}. \quad (8.6.6)$$

We note that Theorem 8.7 provides a bound for the behaviour of the empirically optimal classifier which, in practice, may be computationally very expensive. So we will have to look for suboptimal classifiers.

8.7 Properties of Shatter Coefficient and VC Dimension

In this section we will mention some properties of shatter coefficients and VC dimensions that can be used in other parts of the thesis. (For more on the properties of

shatter coefficients see, for example, Chapter 13 of Devroye et. al. (1996) or Chapter II of Pollard (1984)).

Suppose $\mathcal{A}$ is a class of sets with shatter coefficient $s(\mathcal{A}, n)$ and VC dimension $V_{\mathcal{A}}$.

Let $\bar{\mathcal{A}}$ define classes of sets of the form

$$\bar{\mathcal{A}} = \{ A \times \{0\} \cup A^c \times \{1\}; A \in \mathcal{A} \},$$

where A is a set of the form $\{\mathbf{x} : \psi(\mathbf{x}) = 1\}$, and the sets in $\bar{\mathcal{A}}$ are sets of pairs $(\mathbf{x}, y)$ for which $\psi(\mathbf{x}) \neq y$. Then:

1. For every n , $s(\bar{\mathcal{A}}, n) = s(\mathcal{A}, n)$ and $V_{\bar{\mathcal{A}}} = V_{\mathcal{A}}$
2. $s(\mathcal{A}, n) \leq \sum_{i=1}^{V_{\mathcal{A}}} \binom{n}{i} \leq (n+1)^{V_{\mathcal{A}}}$ for every n
3. If $\mathcal{A} = \mathcal{A}_1 \cup \mathcal{A}_2$, then $s(\mathcal{A}, n) \leq s(\mathcal{A}_1, n) + s(\mathcal{A}_2, n)$
4. If $\mathcal{A}_c = \{A^c; A \in \mathcal{A}\}$, then $s(\mathcal{A}_c, n) = s(\mathcal{A}, n)$
5. For $\mathcal{A} = \{\hat{A} \cap \tilde{A}; \hat{A} \in \hat{\mathcal{A}}, \tilde{A} \in \tilde{\mathcal{A}}\}$ then $s(\mathcal{A}, n) \leq s(\hat{\mathcal{A}}, n)s(\tilde{\mathcal{A}}, n)$
6. For $\mathcal{A} = \{\hat{A} \cup \tilde{A}; \hat{A} \in \hat{\mathcal{A}}, \tilde{A} \in \tilde{\mathcal{A}}\}$ then $s(\mathcal{A}, n) \leq s(\hat{\mathcal{A}}, n)s(\tilde{\mathcal{A}}, n)$
7. For $\mathcal{A} = \{\hat{A} \times \tilde{A}; \hat{A} \in \hat{\mathcal{A}}, \tilde{A} \in \tilde{\mathcal{A}}\}$ then $s(\mathcal{A}, n) \leq s(\hat{\mathcal{A}}, n)s(\tilde{\mathcal{A}}, n)$
8. If $|\mathcal{A}| < \infty$ then $V_{\mathcal{A}} \leq \log_2 |\mathcal{A}|$ and $s(\mathcal{A}, n) \leq |\mathcal{A}|$ for every n .

8.8 Partitioning and Splitting the Data

In this section we give an overview of the consistency of the classification rules $\{\psi_n, n \geq 1\}$ for a large class of random (data-based) partitions of $\mathfrak{R}^d$ induced by

the data. The results are based on the work of Lugosi and Nobel (1996). We will also discuss the notion of splitting the data set into a training sample and a testing sequence to obtain an optimal classifier within a class of classifiers. These results will then be used to establish the consistency properties of some new procedures for combined classification in Chapter 9.

Definition 8.7 (Partition) *A partition of $\mathbb{R}^d$ is a finite collection $\mathcal{P} = \{A_1, \dots, A_r\}$ of subsets of $\mathbb{R}^d$ such that $\cup_{i=1}^r A_i = \mathbb{R}^d$ and $A_i \cap A_j = \emptyset$ for $i \neq j$ where each set $A_i, i = 1, \dots, r$ is called a cell of the partition $\mathcal{P}$.*

Given the training sample T_n , let $\mathcal{P}_n = \pi_n(T_n)$ be a partition of $\mathbb{R}^d$ produced according to a prescribed n -sample partitioning rule π_n . Here, π_n is a function that associates every n -tuple of pairs $(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n) \in \mathbb{R}^d \times \{0, 1\}$ with a measurable partition of $\mathbb{R}^d$. Also, observe that $\mathcal{P}_n$ is a random partition of $\mathbb{R}^d$. We will then use the partition $\mathcal{P}_n$ in conjunction with the training sample T_n to produce a classification rule based on a majority vote within the cells of the partition. Associated with every partitioning rule π_n there is a fixed, nonrandom family of partitions

$$\Omega_n = \{\pi_n(\mathbf{x}_1, y_1, \dots, \mathbf{x}_n, y_n) : (\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n) \in \mathbb{R}^d \times \{0, 1\}\}.$$

We note that Ω_n is the family of partitions produced by π_n for all possible realizations of the training sample T_n .

For any family of partitions Ω of $\mathbb{R}^d$, let $\Delta_n(\Omega, n)$ be the n^{th} shatter coefficient of Ω defined by

$$\Delta_n(\Omega, n) = \tag{8.8.1}$$

$$\max_B \{\text{number of different sets in } \{(A_1 \cap B, \dots, A_r \cap B) : (A_1, \dots, A_r) \in \Omega\}\}$$

where $B = (\mathbf{x}_1, \dots, \mathbf{x}_n)$, $\mathbf{x}_i \in \mathbb{R}^d$, $i = 1, \dots, n$ and $r = |\Omega|$ is the cardinality (number of cells) of Ω . The quantity Δ_n measures the richness (or complexity) of the family Ω .

The following lemma extends Vapnik-Chervonenkis inequality in Theorem 8.6 to families of partitions.

Lemma 8.3 *Let $\mathbf{X}_1, \dots, \mathbf{X}_n$ be i.i.d. random variables in $\mathbb{R}^d$ with measure μ and empirical measure μ_n . Let Ω be a collection of partitions of $\mathbb{R}^d$. For each $n \geq 1$ and $\epsilon > 0$,*

$$\mathbf{P} \left\{ \sup_{\mathcal{P} \in \Omega} \sum_{A \in \mathcal{P}} |\mu_n(A) - \mu(A)| > \epsilon \right\} \leq 4\Delta_{2n}(\Omega, n) 2^{m(\Omega)} e^{-n\epsilon^2/32} \quad (8.8.2)$$

where $m(\Omega)$ is the maximal cell count of Ω defined by

$$m(\Omega) = \sup_{\mathcal{P} \in \Omega} |\mathcal{P}|.$$

Corollary: *Let $(\mathbf{X}_1, Y_1), (\mathbf{X}_2, Y_2), \dots$ be a sequence of i.i.d. random pairs in $\mathbb{R}^d \times \{0, 1\}$ and let $\Omega_1, \Omega_2, \dots$ be a sequence of families of partitions of $\mathbb{R}^d$. If, as $n \rightarrow \infty$, (a) $n^{-1}m(\Omega_n) \rightarrow 0$ and (b) $n^{-1} \log \Delta_n(\Omega_n, n) \rightarrow 0$, then*

$$\sup_{\mathcal{P} \in \Omega_n} \sum_{A \in \mathcal{P}} \left| \frac{1}{n} \sum_{i=1}^n \mathbf{I}\{Y_i = 1\} \mathbf{I}\{\mathbf{X}_i \in A\} - \mathbf{E}\{\mathbf{I}\{Y = 1\} \mathbf{I}\{\mathbf{X} \in A\}\} \right| \rightarrow 0 \text{ a.s.}$$

For the family of partitions Ω_n , let $A_n[\mathbf{x}]$ be the unique cell of $\mathcal{P}_n \in \Omega_n$ that contains the point $\mathbf{x}$ for every $\mathbf{x} \in \mathbb{R}^d$. Also, let $\{\pi_1, \pi_2, \dots\}$ be a fixed sequence of partitioning rules. Then the classification rule $\psi_n(\cdot) = \psi_n(\cdot, T_n)$ is defined by taking a majority vote among the classes in a given cell of $\mathcal{P}_n$, that is,

$$\psi_n(\mathbf{x}) = \begin{cases} 1 & \text{if } \sum_{i=1}^n \mathbf{I}\{\mathbf{X}_i \in A_n[\mathbf{x}], Y_i = 1\} > \sum_{i=1}^n \mathbf{I}\{\mathbf{X}_i \in A_n[\mathbf{x}], Y_i = 0\} \\ 0 & \text{otherwise.} \end{cases}$$

The following theorem is due to Lugosi and Nobel (1996), where they prove the strong consistency of the classification rule $\{\psi_n\}$ for a large class of partitioning rules.

Theorem 8.8 (Strong Consistency of the Classification Rules) *Let*

$\{\pi_1, \pi_2, \dots\}$ *be a fixed sequence of partitioning rules. Let* $\text{diam}(A) = \sup_{\mathbf{x}, \mathbf{y} \in A} \|\mathbf{x} - \mathbf{y}\|$ *denote the diameter of a set* A *and let* μ *be the distribution of* $\mathbf{X}$. *For each* n *let* Ω_n *be the collection of partitions associated with the* n -*sample partitioning rule* π_n . *If, as* $n \rightarrow \infty$,

1. $\frac{\log(\Delta_n(\Omega_n, n))}{n} \rightarrow 0$ *where* $\Delta_n(\Omega_n, n)$ *is the* n^{th} *shatter coefficient of* Ω_n , *and*
2. *for every* $\gamma > 0$ *and* $\delta \in (0, 1)$,

$$\inf_{\mathbf{S} \subseteq \mathbb{R}^d: \mu(\mathbf{S}) \geq 1-\delta} \mu\{\mathbf{x} : \text{diam}(A_n[\mathbf{x}] \cap \mathbf{S}) > \gamma\} \rightarrow_{\text{a.s.}} 0,$$

then the classification rule $\{\psi_n\}$ *corresponding to* π_n *is strongly consistent in the sense that*

$$\text{err}_n(\psi_n) - \text{err}(\psi_B) \rightarrow_{\text{a.s.}} 0.$$

The first condition in Theorem 8.8 controls the richness or cardinality of the family of partitions Ω_n of $\mathbb{R}^d$, while the second condition is a shrinking cell condition which requires the cells of π_n to have small diameters when n is large enough.

The idea of splitting the data, to obtain an optimal classifier within a class of classifiers, was analyzed in Devroye (1988). Suppose the training sample T_n defines a class of classifiers Ψ_n whose members are denoted by $\psi_n(\cdot) = \psi_n(\cdot, T_n)$. The testing set T_m is then used to select a classifier from Ψ_n , denoted by ψ_N , that minimizes

(8.4.1) and satisfies

$$\widehat{\text{err}}_N(\psi_N) \leq \widehat{\text{err}}_N(\psi_n) \text{ for all } \psi_n \in \Psi_n.$$

We note that both ψ_N and $\widehat{\text{err}}_N$ depend on the entire data set T_N , as the rest of the data is used for testing the classifiers in Ψ_n . Now we are interested in the difference between the misclassification error rate of the selected classifier ψ_N , given by

$$\text{err}(\psi_N) = \mathbf{P}\{\psi_N(\mathbf{X}) \neq Y \mid T_N\},$$

and that of the best classifier in Ψ_n , defined by $\inf_{\psi_n \in \Psi_n} \text{err}(\psi_n)$, where $\text{err}(\psi_n) = \mathbf{P}\{\psi_n(\mathbf{X}) \neq Y \mid T_n\}$ denotes the error probability conditioned on the training sample T_n .

If the class Ψ_n contains finitely many classification rules, then from equation (8.6.2), we have

$$\mathbf{P} \left\{ \sup_{\psi_n \in \Psi_n} \mid \widehat{\text{err}}_N(\psi_n) - \text{err}_n(\psi_n) \mid > \epsilon \mid T_n \right\} \leq 2|\Psi_n|e^{-2m\epsilon^2}$$

and therefore by Lemma 8.1,

$$\mathbf{P} \left\{ \text{err}_n(\psi_N) - \inf_{\psi_n \in \Psi_n} \text{err}_n(\psi_n) > \epsilon \mid T_n \right\} \leq 2|\Psi_n|e^{-m\epsilon^2/2}. \quad (8.8.3)$$

However, if Ψ_n is of infinite cardinality, then we can use the Vapnik-Chervonenkis theory in Section 8.6 to get the following inequality.

Theorem 8.9 (Devroye, 1988) *For a fixed training sample $\{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)\}$, let Ψ_n^* be the class of sets of the form*

$$\{ \{ \mathbf{x} : \psi_n(\mathbf{x}) = 1 \} \times \{0\} \} \cup \{ \{ \mathbf{x} : \psi_n(\mathbf{x}) = 0 \} \times \{1\} \}, \quad \psi_n \in \Psi_n.$$

Define the r^{th} shatter coefficient of Ψ_n^* by

$$\Delta_n(\Psi_n^*, r) = \max_{B: B \in \{\mathbb{R}^d \times \{0,1\}\}^r} \{\text{number of different sets in } \{B \cap \psi_n^* : \psi_n^* \in \Psi_n^*\}\}.$$

Then

$$\mathbb{P} \left\{ \sup_{\psi_n \in \Psi_n} | \widehat{\text{err}}_N(\psi_n) - \text{err}_n(\psi_n) | > \epsilon \right\} \leq 4 \sup_{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)} \Delta_n(\Psi_n^*, 2m) e^{-(m\epsilon^2/8)} \quad (8.8.4)$$

where $m\epsilon^2 \geq 2$.

The shatter coefficients $\Delta_n(\Psi_n^*, 2m)$ may depend on the training sample T_n since the class of classifiers Ψ_n depends on T_n . However, quite often, it may be possible to find upper bounds on $\Delta_n(\Psi_n^*, 2m)$ that depend on n and m only, but not on the fixed training sample $\{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)\}$. In these cases, one can drop the supremum on the right side of (8.8.4) and therefore the upper bound becomes distribution-free.

Chapter 9

Combining Classifiers

9.1 Introduction

In this chapter we consider methods for combining a number of individual classifiers in order to produce more effective classification rules, in terms of smallest error rates. Suppose we have L individual classifiers available for a given problem. These individual classifiers could be, for example, Fisher's linear discriminant functions, nearest-neighbour classifiers, kernel rules, partitioning rules or tree classifiers. Often, the approach used is to find the single best classifier among the L individual classifiers. Some recent research has looked at the idea of combining these individual classifiers rather than selecting the best classifier.

In Section 9.2 we present methods for combining classifiers based on linear combinations of the estimated class conditional probabilities. Leblanc and Tibshirani (1996) considered a leave-one-out cross-validation procedure whereas Mojirsheibani (1997) used a data-splitting approach for combining classifiers. Section 9.3 presents a

discretization- or partitioning-based procedure for combining several classifiers, proposed by Mojirsheibani (1999). Kernel-based combined classification procedures (Mojirsheibani, 2000) are discussed in Section 9.4. Our contributions, which extend the results of Mojirsheibani (2000), involve kernel-based combined classification procedures with variable smoothing parameters.

9.2 Linear Combined Classifiers

In a recent article, Leblanc and Tibshirani (1996) considered a leave-one-out cross-validation procedure for combining classifiers. Their method is based on the “model-mix” proposal of Stone (1974) and works as follows.

Suppose a data set $T_N = \{(\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_N, Y_N)\}$ is available where $(\mathbf{X}_i, Y_i)$, $i = 1, \dots, N$ are independent and identically distributed $\mathfrak{R}^d \times \{1, \dots, K\}$ -valued random vectors with distribution $F_{\mathbf{X}, Y}$, and let $(\mathbf{X}, Y)$ be a new observation from $F_{\mathbf{X}, Y}$ for which Y is to be determined. We assume that L different classifiers, $\psi_{N,1}, \dots, \psi_{N,L}$, are also available from the data set T_N , each of which producing estimates of the class probabilities

$$\hat{\Pi}_{l,N}(\mathbf{x}) = \{\hat{\pi}_{l,N}^{(1)}(\mathbf{x}), \dots, \hat{\pi}_{l,N}^{(K)}(\mathbf{x})\}, \quad l = 1, \dots, L,$$

where $\hat{\pi}_{l,N}^{(k)}(\mathbf{x})$, $k = 1, \dots, K$ is the estimate of

$$\pi^{(k)}(\mathbf{x}) = \mathbf{P}(Y = k | \mathbf{X} = \mathbf{x})$$

produced by the l^{th} classifier $\psi_{N,l}$. We note that each classifier is a map defined by

$$\psi_{N,l}: \{\mathfrak{R}^d \times \{1, \dots, K\}\}^N \times \mathfrak{R}^d \rightarrow \{1, \dots, K\}, \quad l = 1, \dots, L.$$

The combined estimator is defined by a set of weighted sums

$$w_k(\mathbf{x}) = \beta_k^T \hat{\Pi}_N(\mathbf{x}), \quad k = 1, \dots, K$$

where each β_k is a LK -dimensional column vector of weights and $\hat{\Pi}_N(\mathbf{x}) = \{\hat{\Pi}_{1,N}(\mathbf{x}), \dots, \hat{\Pi}_{L,N}(\mathbf{x})\}^T$. Leblanc and Tibshirani then used a multivariate regression of the response vector $\mathbf{y}$ to estimate the β_k 's, where $\mathbf{y}$ has a 1 in the k^{th} position if the observation belongs to class k and 0 otherwise, on a leave-one-out cross-validated version of $\hat{\Pi}_N(\mathbf{x})$. In other words, one finds the multivariate least-squares estimates $\hat{\beta}_1, \dots, \hat{\beta}_K$ based on the data $(\mathbf{y}_i, \hat{\Pi}_{-i,N}(\mathbf{x}_i))$, $i = 1, \dots, N$, where $\hat{\Pi}_{-i,N}(\mathbf{x}_i)$ is the leave-one-out cross-validated version of $\hat{\Pi}_N(\mathbf{x})$ evaluated at $\mathbf{x}_i$. Finally, given the combined estimators

$$\hat{w}_k(\mathbf{x}) = \hat{\beta}_k^T \hat{\Pi}_N(\mathbf{x}), \quad k = 1, \dots, K$$

we classify $\mathbf{x}$ as belonging to class k if

$$\hat{w}_k(\mathbf{x}) = \max\{\hat{w}_1(\mathbf{x}), \dots, \hat{w}_K(\mathbf{x})\},$$

and the estimated class probabilities are taken to be

$$\hat{\pi}_{l,N}^{(k)}(\mathbf{x}) = \frac{\hat{w}_k(\mathbf{x})}{\sum_{k=1}^K \hat{w}_k(\mathbf{x})}, \quad l = 1, \dots, L.$$

Mojirsheibani(1997) proposed a data-splitting method for constructing combined classifiers. His method starts by splitting the data set T_N into a training sample $T_n = \{(\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_n, Y_n)\}$ and a testing set $T_m = \{(\mathbf{X}_{n+1}, Y_{n+1}), \dots, (\mathbf{X}_{n+m}, Y_{n+m})\}$, where $n + m = N$. We now assume that L different classifiers, $\psi_{n,1}, \dots, \psi_{n,L}$, are available from the training sample T_n , each of which producing estimates of the class

probabilities

$$\hat{\Pi}_{l,n}(\mathbf{x}) = \{\hat{\pi}_{l,n}^{(1)}(\mathbf{x}), \dots, \hat{\pi}_{l,n}^{(K)}(\mathbf{x})\}, \quad l = 1, \dots, L,$$

where $\hat{\pi}_{l,n}^{(k)}(\mathbf{x})$, $k = 1, \dots, K$ is the estimate of

$$\pi^{(k)}(\mathbf{x}) = \mathbf{P}(Y = k | \mathbf{X} = \mathbf{x})$$

produced by the l^{th} classifier $\psi_{n,l}$ obtained from the training sample T_n . Now we consider the class of combined classifiers

$$\tilde{\Psi}_n = \{\tilde{\psi}_n : \tilde{\psi}_n(\mathbf{x}) = \arg\max_{1 \leq k \leq K} \alpha_k^T \hat{\Pi}_n(\mathbf{x}); \alpha_k \in \mathbb{R}^{LK}, k = 1, \dots, K, \mathbf{x} \in \mathbb{R}^d\},$$

based on linear combinations of the estimated class conditional probabilities, where $\hat{\Pi}_n(\mathbf{x})$ is a LK -dimensional column vector $\{\hat{\Pi}_{1,n}(\mathbf{x}), \dots, \hat{\Pi}_{L,n}(\mathbf{x})\}^T$ and $\tilde{\Psi}_n$ depends on the training sample T_n through $\hat{\Pi}_n(\mathbf{x})$. Let $\tilde{\psi}_N$ be the classifier, selected from the class $\tilde{\Psi}_n$, that minimizes the average number of errors committed on the testing set T_m

$$\widehat{\text{err}}_N(\tilde{\psi}_n) = \frac{1}{m} \sum_{j=1}^m \mathbf{I}\{\tilde{\psi}_n(\mathbf{X}_{n+j}) \neq Y_{n+j}\}, \quad (9.2.1)$$

where $\mathbf{I}\{A\}$ denotes the indicator function of a set A . Note that both $\tilde{\psi}_N$ and $\widehat{\text{err}}_N(\tilde{\psi}_N)$ depend on the entire data set of size N . If at least one of the L classifiers is strongly (weakly) consistent then, Mojirsheibani (1997) showed that the combined classifier $\tilde{\psi}_N$, picked to minimize (9.2.1), is also strongly (weakly) consistent. In other words,

$$\text{err}_N(\tilde{\psi}_N(\mathbf{X})) \xrightarrow{\text{a.s.}} \text{err}(\psi_B) \text{ as } N \rightarrow \infty,$$

where $\text{err}_N(\tilde{\psi}_N(\mathbf{X})) = \mathbf{P}(\tilde{\psi}_N(\mathbf{X}) \neq Y | T_N)$.

Other related work on combining individual classifiers are those of Wolpert (1992), who presented “stacked generalization” in the neural network literature, and Breiman (1995) who applied Wolpert’s proposal in the regression setting, calling it “stacked regression”. These methods are closely related to the combining procedure of Leblanc and Tibshirani (1996).

9.3 Combined Classifiers Via Discretization

More recently, Mojirsheibani (1999) proposed a combined classification method, where the combined classifier $\tilde{\psi} : \{1, \dots, K\}^L \rightarrow \{1, \dots, K\}$ acts on the L -dimensional vector of data-based classifiers $\{\psi_{N,1}(\mathbf{X}), \dots, \psi_{N,L}(\mathbf{X})\}$ rather than the more traditional feature vector $\mathbf{X}$. In this combined classification setup, we would like to find a combined classifier $\tilde{\psi} = \tilde{\psi}(\psi_{N,1}(\mathbf{X}), \dots, \psi_{N,L}(\mathbf{X}))$ such that the error probability

$$\text{err}_N(\tilde{\psi}) = \mathbf{P}(\tilde{\psi}(\psi_{N,1}(\mathbf{X}), \dots, \psi_{N,L}(\mathbf{X})) \neq Y \mid T_N)$$

is, in some sense, as small as possible. Let $\tilde{\psi}_N$ be the combined classifier defined by

$$\begin{aligned} \tilde{\psi}_N &= \tilde{\psi}_N(\psi_{N,1}(\mathbf{x}), \dots, \psi_{N,L}(\mathbf{x})) \\ &= \operatorname{argmax}_{1 \leq k \leq K} \sum_{i=1}^N \prod_{l=1}^L \mathbf{I}\{\psi_{N,l}(\mathbf{X}_i) = \psi_{N,l}(\mathbf{x})\} \mathbf{I}\{Y_i = k\}. \end{aligned} \quad (9.3.1)$$

We note that the combined classifier in (9.3.1) is based on a class-majority vote in the sense that we select class k if the number of class k data points for which all L classifiers predict (correctly or incorrectly) the same class, at both the data point and the new observation, is larger than for any other classes. If there are ties we take $\tilde{\psi}_N$ to be the smallest k for which (9.3.1) holds. The combined classifier in (9.3.1) is

motivated by the fact that, in an unrealistic situation, if all the individual classifiers $\psi_{N,1}(\mathbf{X}), \dots, \psi_{N,L}(\mathbf{X})$ are nonrandom, then (9.3.1) is simply a partitioning-type classification rule based on the i.i.d. discretized “data” $\{(\psi_{N,1}(\mathbf{X}_i), \dots, \psi_{N,L}(\mathbf{X}_i)), Y_i\}$, for $i = 1, \dots, N$. The discretized “data” vectors $\{\psi_{N,1}(\mathbf{X}_i), \dots, \psi_{N,L}(\mathbf{X}_i)\}$, $i = 1, \dots, N$ in (9.3.1) are not i.i.d. and, in fact, depend on the data set T_N .

In the two class problem, the combined classifier in (9.3.1) reduces to

$$\tilde{\psi}_N = \begin{cases} 1 & \text{if } \sum_{i=1}^N \prod_{l=1}^L \mathbf{I}\{\psi_{N,l}(\mathbf{X}_i) = \psi_{N,l}(\mathbf{x})\} \mathbf{I}\{Y_i = 1\} \\ & > \sum_{i=1}^N \prod_{l=1}^L \mathbf{I}\{\psi_{N,l}(\mathbf{X}_i) = \psi_{N,l}(\mathbf{x})\} \mathbf{I}\{Y_i = 0\} \\ 0 & \text{otherwise} \end{cases} \quad (9.3.2)$$

9.3.1 Example

Before we present the large-sample properties of the combined classifier via discretization, we consider a simple example that will demonstrate how this method works. Suppose we have only two classes (0 and 1) and there are four classifiers available, namely $\psi_{N,1}$, $\psi_{N,2}$, $\psi_{N,3}$, $\psi_{N,4}$. Let $(\mathbf{x}_i, y_i)$, $i = 1, \dots, 12$, be the observed data values, and consider the predicted class memberships for the 12 data points given in Table 9.1.

Let $\psi_{N,1}(\mathbf{x}_0) = 0$, $\psi_{N,2}(\mathbf{x}_0) = 1$, $\psi_{N,3}(\mathbf{x}_0) = 0$, and $\psi_{N,4}(\mathbf{x}_0) = 1$, where $\psi_{N,l}(\mathbf{x}_0)$, $l = 1, 2, 3, 4$, is the predicted class membership of a future observation $(\mathbf{x}_0, y_0)$. Now, because

$$\sum_{i: y_i=0} \mathbf{I} \{ \psi_{N,1}(\mathbf{x}_i) = \psi_{N,1}(\mathbf{x}_0), \psi_{N,2}(\mathbf{x}_i) = \psi_{N,2}(\mathbf{x}_0), \psi_{N,3}(\mathbf{x}_i) = \psi_{N,3}(\mathbf{x}_0), \\ \psi_{N,4}(\mathbf{x}_i) = \psi_{N,4}(\mathbf{x}_0) \} = 2$$

Table 9.1: Predicted Class Memberships

Data point		Classifiers				
	i	$\psi_{N,1}$	$\psi_{N,2}$	$\psi_{N,3}$	$\psi_{N,4}$	y_i
	1	0	1	0	1	0
	2	0	0	1	1	1
	3	0	0	1	1	0
	4	0	1	0	1	0
	5	1	0	1	1	1
	6	1	1	0	0	1
	7	0	0	1	1	1
	8	1	0	0	1	0
	9	0	0	0	1	0
	10	1	1	1	0	0
	11	1	0	1	0	0
	12	0	1	0	1	1

and

$$\sum_{i:y_i=1} \mathbf{I} \{ \psi_{N,1}(\mathbf{x}_i) = \psi_{N,1}(\mathbf{x}_0), \psi_{N,2}(\mathbf{x}_i) = \psi_{N,2}(\mathbf{x}_0), \psi_{N,3}(\mathbf{x}_i) = \psi_{N,3}(\mathbf{x}_0), \\ \psi_{N,4}(\mathbf{x}_i) = \psi_{N,4}(\mathbf{x}_0) \} = 1$$

the combined classifier $\tilde{\psi}_N$ predicts y_0 to be in class 0.

9.3.2 Large-sample Results

We will now present the large-sample behaviour of the combined classifier given in equation (9.3.2) for the two-class problem (Mojirsheibani, 1999). Denote a data-based classifier by ψ_N , with the conditional error probability given by $\text{err}_N(\psi_N) = \mathbf{P}(\psi_N(\mathbf{X}) \neq Y \mid T_N)$. Let $\text{err}(\psi_B) = \mathbf{P}(\psi_B(\mathbf{X}) \neq Y)$ and define $\pi(\mathbf{x}) = \mathbf{P}(Y = 1 \mid \mathbf{X} = \mathbf{x})$. Now let $\hat{\pi}^{(1)}(\mathbf{x})$ and $\hat{\pi}^{(0)}(\mathbf{x})$ be the estimates of $\pi(\mathbf{x})$ and $(1 - \pi(\mathbf{x}))$, where $\hat{\pi}^{(0)}(\mathbf{x}) + \hat{\pi}^{(1)}(\mathbf{x})$ does not necessarily equal to one. The Bayes classifier and its plug-in estimates are given by

$$\psi_B(\mathbf{x}) = \begin{cases} 1 & \text{if } \mathbf{P}(Y = 1 \mid \mathbf{X} = \mathbf{x}) > 0.5 \\ 0 & \text{otherwise} \end{cases}$$

and

$$\psi_N(\mathbf{x}) = \begin{cases} 1 & \text{if } \hat{\pi}^{(1)}(\mathbf{x}) > \hat{\pi}^{(0)}(\mathbf{x}) \\ 0 & \text{otherwise} \end{cases}. \quad (9.3.3)$$

We will then obtain the following standard results (see Chapter 2 of Devroye, Györfi and Lugosi, 1996).

Lemma 9.1 *Given $\pi(\mathbf{x})$, $\hat{\pi}^{(1)}(\mathbf{x})$, $\hat{\pi}^{(0)}(\mathbf{x})$, $\text{err}(\psi_B)$, and $\text{err}_N(\psi_N)$ defined above, we have:*

(i) For any classifier $\psi : \mathbb{R}^d \rightarrow \{0, 1\}$,

$$\text{err}(\psi_B) \leq \text{err}(\psi).$$

(ii) Let μ be the probability measure of $\mathbf{X}$. Then

$$\begin{aligned} \text{err}_N(\psi_N) - \text{err}(\psi_B) &\leq \int_{\mathbb{R}^d} |\pi(\mathbf{x}) - \hat{\pi}^{(1)}(\mathbf{x})| \mu(d\mathbf{x}) \\ &\quad + \int_{\mathbb{R}^d} |(1 - \pi(\mathbf{x})) - \hat{\pi}^{(0)}(\mathbf{x})| \mu(d\mathbf{x}). \end{aligned} \quad (9.3.4)$$

The first part in Lemma 9.1 says that the Bayes classifier gives the lowest misclassification error. The upper bound given in the second part is quite often used to establish consistency results.

The results in Lemma 9.1 can now be applied to the combined classification setup where the map $\tilde{\psi}$ acts on the L -dimensional vector of classifiers $\{\psi_{N,1}(\mathbf{X}), \dots, \psi_{N,L}(\mathbf{X})\}$.

Let the conditional probability in the combined classification setup be given by

$$\pi(\psi_{N,1}(\mathbf{X}), \dots, \psi_{N,L}(\mathbf{X})) = \mathbf{P}(Y = 1 | \psi_{N,1}(\mathbf{X}), \dots, \psi_{N,L}(\mathbf{X}))$$

and define the combined classifier

$$\tilde{\psi}^*(\psi_{N,1}(\mathbf{x}), \dots, \psi_{N,L}(\mathbf{x})) = \begin{cases} 1 & \text{if } \pi(\psi_{N,1}(\mathbf{X}), \dots, \psi_{N,L}(\mathbf{X})) > 0.5 \\ 0 & \text{otherwise.} \end{cases} \quad (9.3.5)$$

Now let $\pi^{(1)}(\psi_{N,1}(\mathbf{x}), \dots, \psi_{N,L}(\mathbf{x}))$ and $\pi^{(0)}(\psi_{N,1}(\mathbf{x}), \dots, \psi_{N,L}(\mathbf{x}))$ be some data-based versions of $\pi(\psi_{N,1}(\mathbf{X}), \dots, \psi_{N,L}(\mathbf{X}))$ and $1 - \pi(\psi_{N,1}(\mathbf{X}), \dots, \psi_{N,L}(\mathbf{X}))$ where their sum does not necessarily equal to one. We will note here that the data set T_N has been used twice; first to find the classifiers $\psi_{N,l}(\mathbf{x})$, $l = 1, \dots, L$ and a second time to find the estimates $\pi^{(1)}(\psi_{N,1}(\mathbf{x}), \dots, \psi_{N,L}(\mathbf{x}))$ and $\pi^{(0)}(\psi_{N,1}(\mathbf{x}), \dots, \psi_{N,L}(\mathbf{x}))$. The following lemma is a counterpart of Lemma 9.1, for the combined classification setup.

Lemma 9.2 (Upper Bound of Discretized Combined Classifier) *Let $\tilde{\psi}$ be any combined classifier and $\tilde{\psi}^*$ be the combined classifier defined in (9.3.5). Let $\text{err}_N(\tilde{\psi}^*) = \mathbf{P}\{\tilde{\psi}^*(\psi_{N,1}(\mathbf{x}), \dots, \psi_{N,L}(\mathbf{x})) \neq Y \mid T_N\}$ and $\text{err}_N(\tilde{\psi}) = \mathbf{P}\{\tilde{\psi}(\psi_{N,1}(\mathbf{x}), \dots, \psi_{N,L}(\mathbf{x})) \neq Y \mid T_N\}$.*

(i) *Then $\text{err}_N(\tilde{\psi}) - \text{err}_N(\tilde{\psi}^*) \geq \text{a.s. } 0$.*

(ii) *For the combined classifier*

$$\begin{aligned} \tilde{\psi}_N &= \tilde{\psi}_N(\psi_{N,1}(\mathbf{x}), \dots, \psi_{N,L}(\mathbf{x})) \\ &= \begin{cases} 1 & \text{if } \pi^{(1)}(\psi_{N,1}(\mathbf{x}), \dots, \psi_{N,L}(\mathbf{x})) \\ & > \pi^{(0)}(\psi_{N,1}(\mathbf{x}), \dots, \psi_{N,L}(\mathbf{x})) \\ 0 & \text{otherwise,} \end{cases} \end{aligned}$$

we have the following bound

$$\begin{aligned} &\text{err}_N(\tilde{\psi}_N) - \text{err}_N(\tilde{\psi}^*) \\ &\leq \text{a.s.} \int_{\mathbb{R}^d} |\pi^{(1)}(\psi_{N,1}(\mathbf{x}), \dots, \psi_{N,L}(\mathbf{x})) - \mathbf{P}(Y = 1 | \mathbf{X} = \mathbf{x})| \mu(d\mathbf{x}) \\ &\quad + \int_{\mathbb{R}^d} |\pi^{(0)}(\psi_{N,1}(\mathbf{x}), \dots, \psi_{N,L}(\mathbf{x})) - \mathbf{P}(Y = 0 | \mathbf{X} = \mathbf{x})| \mu(d\mathbf{x}), \end{aligned} \tag{9.3.6}$$

where μ is the probability measure of $\mathbf{X}$.

The main difference between Lemma 9.1 and Lemma 9.2 is the fact that, unlike the Bayes error in the single classifier setup, the error of the best combined classifier $\text{err}_N(\tilde{\psi}^*)$ is a function of the data set T_N .

Let $\mathcal{P}_N = \pi_N(T_N) = \{A_{N,1}, \dots, A_{N,2^L}\}$ be a random partition of $\mathfrak{R}^d$, induced by the data set T_N , where the i^{th} cell of the partition $A_{N,i}$, $i = 1, \dots, 2^L$ is given by

$$A_{N,i} = \{\mathbf{x} \in \mathfrak{R}^d : (\psi_{N,1}(\mathbf{x}), \dots, \psi_{N,L}(\mathbf{x})) = \mathbf{t}_i \text{ for one and only one } \mathbf{t}_i \in \{0, 1\}^L\},$$

where $A_{N,i} \cap A_{N,j} = \emptyset$ for $i \neq j$ and $\cup_{i=1}^{2^L} A_{N,i} = \mathfrak{R}^d$. Let $A_N[\mathbf{x}]$ be the unique cell that contains the point $\mathbf{x}$. Also, let $T_N^{(\text{obs})} = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_N, y_N)\}$ be the observed values of the data set T_N and let

$$\Omega_N = \{\pi : \pi = \pi_N(T_N^{(\text{obs})}), \text{ for } T_N^{(\text{obs})} \in \{\mathfrak{R}^d \times \{0, 1\}\}^N\}$$

be the nonrandom family of partitions of $\mathfrak{R}^d$ associated with π_N .

The consistency results for the combined classification setup, which is an extension of Theorem 8.8 and given in Mojirsheibani (1999), are now stated in the following theorem.

Theorem 9.1 (Mojirsheibani, 1999) *If, as $N \rightarrow \infty$ and the number of individual classifiers $L = L_N$ is allowed to increase with N such that,*

1. $L_N / \log N \rightarrow 0$
2. $N^{-1} \log \Delta_N(\Omega_N, N) \rightarrow 0$ and
3. for every $\gamma > 0$ and $\delta \in (0, 1)$,

$$\inf_{\mathbf{S} \subseteq \mathfrak{R}^d : \mu(\mathbf{S}) \geq 1-\delta} \mu\{\mathbf{x} : \text{diam}(A_N[\mathbf{x}] \cap \mathbf{S}) > \gamma\} \rightarrow \text{a.s. } 0,$$

then the combined classification rule $\tilde{\psi}_N$ is strongly consistent in the sense that

$$\text{err}_N(\tilde{\psi}_N) - \text{err}_N(\tilde{\psi}^*) \rightarrow \text{a.s. } 0.$$

It is also shown that $\tilde{\psi}_N$ is, asymptotically, at least as good as any one of the individual classifiers, even if one of them is already strongly consistent.

9.4 Kernel-based Combined Classifier

9.4.1 Fixed Smoothing Parameter

The class-majority vote combined classifier in equation (9.3.1) is not smooth in the sense that the vote associated with $\mathbf{X}_i, i = 1, \dots, N$ is either 0 or 1 according to whether

$$\mathbf{I}\{\psi_{N,l}(\mathbf{X}_i) = \psi_{N,l}(\mathbf{x}), l = 1, \dots, L\}$$

is 0 or 1. So, perhaps, a more flexible combined classification procedure can be formed in which the votes are allowed to take values in the interval $[0,1]$. This leads to a class of kernel-based combined classifiers, discussed in Mojirsheibani (2000), where the optimal member of the class is selected using a data-splitting approach. The steps may be summarized as follows.

Consider the data set $T_N = \{(\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_N, Y_N)\}$. Suppose we randomly split the data set into a training sample T_n of size n and a testing set T_m of size m , where $n+m = N$. Let $\psi_{n,1}, \dots, \psi_{n,L}$ be L individual classifiers, based on the training sample T_n . For fixed $\alpha > 0$ consider the exponential kernels given by

$$\mathbf{K}_h(t) = e^{-(t/h)^\alpha}, h > 0.$$

We use exponential kernels here because there are known asymptotic results for ex-

ponential kernels. Define the class of combined classifiers

$$\tilde{\Psi}_{n,h} = \{\tilde{\psi}_{n,h} \mid h \geq 0\}$$

where $\tilde{\psi}_{n,h}$ is defined by

$$\begin{aligned} \tilde{\psi}_{n,h}(\mathbf{x}) &\equiv \tilde{\psi}_{n,h}(\psi_{n,1}(\mathbf{x}), \dots, \psi_{n,L}(\mathbf{x})) \\ &= \begin{cases} \operatorname{argmax}_{1 \leq k \leq K} \sum_{\{i: 1 \leq i \leq n, Y_i = k\}} \mathbf{K}_h(\sum_{l=1}^L \mathbf{I}\{\psi_{n,l}(\mathbf{X}_i) \neq \psi_{n,l}(\mathbf{x})\}), & \text{if } h > 0 \\ \operatorname{argmax}_{1 \leq k \leq K} \sum_{\{i: 1 \leq i \leq n, Y_i = k\}} \mathbf{I}\{\psi_{n,l}(\mathbf{X}_i) = \psi_{n,l}(\mathbf{x}), l = 1, \dots, L\}, & \text{if } h = 0. \end{cases} \end{aligned} \quad (9.4.1)$$

We note that, when $h = 0$, the above kernel-based combined classifier reduces to the class-majority vote combined classifier defined in equation (9.3.1), in which case we can use the entire data set T_N as the training sample. In the case of ties, we take $\tilde{\psi}_{n,h}$ to be the smallest k for which (9.4.1) holds. The selected combined classifier, denoted by $\tilde{\psi}_N^K$, is the classifier that minimizes

$$\widehat{\text{err}}_N(\tilde{\psi}_{n,h}) = \frac{1}{m} \sum_{j=1}^m \mathbf{I}\{\tilde{\psi}_{n,h}(\mathbf{X}_{n+j}) \neq Y_{n+j}\}, \quad (9.4.2)$$

where the subscript N shows that both $\tilde{\psi}_N^K$ and $\widehat{\text{err}}_N$ depend on the entire data set T_N . In practice, $\tilde{\psi}_N^K$ is selected from a search over a grid of values of h , say $\{h_{n,1}, \dots, h_{n,c(n)}\}$, where $c(n)$ represents the cardinality of the grid. Note that $c(n)$ typically increases with n , while the distances between two adjacent grid points decreases with n .

The following theorem (Mojirsheibani, 2000) shows that the selected combined classifier, $\tilde{\psi}_N^K$, is (asymptotically) at least as good as each individual classifier $\psi_{N,l}$. We note that the individual classifiers are denoted by $\psi_{N,1}, \dots, \psi_{N,L}$ instead of $\psi_{n,1}, \dots, \psi_{n,L}$

because, if one wants to select the best individual classifier instead of the combined classifier, then there is no need to split the data set T_N . If one of the individual classifiers is already strongly consistent, then $\tilde{\psi}_N^K$ will also be strongly consistent.

Theorem 9.2 (Consistency of Fixed Kernel Combined Classifier) *For the class of combined classifiers $\tilde{\psi}_{n,h}$, $h \geq 0$, defined by (9.4.1), let $\tilde{\psi}_N^K$ be the selected combined classifier that minimizes (9.4.2). If, as N (and thus n and m) $\rightarrow \infty$,*

1. $m^{-1} \log n \rightarrow 0$,
2. $L_n / \log n \rightarrow 0$,
3. $n^{-1} \log \Delta_n(\Omega_n, n) \rightarrow 0$ and
4. for every $\gamma > 0$ and $\delta \in (0, 1)$,

$$\inf_{\mathbf{S} \subseteq \mathbb{R}^d, \mu(\mathbf{S}) \geq 1-\delta} \mu\{\mathbf{x} : \text{diam}(A_n[\mathbf{x}] \cap \mathbf{S}) > \gamma\} \rightarrow_{\text{a.s.}} 0,$$

then $\tilde{\psi}_N^K$ is, asymptotically, at least as good as any one of the individual classifiers $\psi_{N,l}$, $l = 1, \dots, L$ in the sense that:

$$\limsup_{N \rightarrow \infty} \left\{ \text{err}_N(\tilde{\psi}_N^K) - \text{err}_N(\psi_{N,l}) \right\} \leq_{\text{a.s.}} 0,$$

where $\text{err}_N(\tilde{\psi}_N^K) = \mathbf{P}\{\tilde{\psi}_N^K(\psi_{N,1}(\mathbf{x}), \dots, \psi_{N,L}(\mathbf{x})) \neq Y \mid (\mathbf{X}_i, Y_i), i = 1, \dots, N\}$.

Corollary: *If one of the individual classifiers is already strongly (weakly) consistent, under the conditions of Theorem 9.2, then*

$$\text{err}_N(\tilde{\psi}_N^K) \rightarrow_{\text{a.s.}} \text{err}(\psi_B)$$

where $\text{err}(\psi_B)$ is the Bayes' error defined in equation (8.2.2).

9.4.2 Variable Smoothing Parameters

We will now extend the kernel-based combined classifier in (9.4.1) by allowing the smoothing parameter h to be different for each of the K classes. The motivation for changing the smoothing parameter for each of the K classes is that this is more flexible than the fixed kernel, and it may lead to improvements in the method in terms of minimizing the error rates. That is, for fixed $\alpha > 0$ we will now consider the exponential kernel defined by

$$\mathbf{K}_{h_k}(t) = e^{-(t/h_k)^\alpha}, \quad h_k > 0 \text{ for all } k = 1, \dots, K.$$

Then the class of combined classifiers is given by

$$\tilde{\Psi}_{n,\mathbf{h}} = \{\tilde{\psi}_{n,\mathbf{h}} \mid \mathbf{h} = (h_1, \dots, h_K), \quad h_k \geq 0 \text{ for } k = 1, \dots, K\}$$

where $\tilde{\psi}_{n,\mathbf{h}}$ is defined by

$$\begin{aligned} \tilde{\psi}_{n,\mathbf{h}}(\mathbf{x}) &\equiv \tilde{\psi}_{n,\mathbf{h}}(\psi_{n,1}(\mathbf{x}), \dots, \psi_{n,L}(\mathbf{x})) \\ &= \begin{cases} \operatorname{argmax}_{1 \leq k \leq K} \sum_{\{i: 1 \leq i \leq n, Y_i = k\}} \mathbf{K}_{h_k}(\sum_{l=1}^L \mathbf{I}\{\psi_{n,l}(\mathbf{X}_i) \neq \psi_{n,l}(\mathbf{x})\}), & \text{if } h_k > 0 \forall k \\ \operatorname{argmax}_{1 \leq k \leq K} \sum_{\{i: 1 \leq i \leq n, Y_i = k\}} \mathbf{I}\{\psi_{n,l}(\mathbf{X}_i) = \psi_{n,l}(\mathbf{x}), \quad l = 1, \dots, L\}, & \text{if } \exists h_k = 0. \end{cases} \end{aligned} \quad (9.4.3)$$

In the case of ties, we take $\tilde{\psi}_{n,\mathbf{h}}(\psi_{n,1}(\mathbf{x}), \dots, \psi_{n,L}(\mathbf{x}))$ to be the smallest k for which (9.4.3) holds. Let $\tilde{\psi}_{N,\mathbf{h}}$ be the combined classifier that minimizes the error rate

$$\widehat{\operatorname{err}}_N(\tilde{\psi}_{n,\mathbf{h}}) = \frac{1}{m} \sum_{j=1}^m \mathbf{I}\{\tilde{\psi}_{n,\mathbf{h}}(\mathbf{X}_{n+j}) \neq Y_{n+j}\}. \quad (9.4.4)$$

To find the combined classifier $\tilde{\psi}_{N,\mathbf{h}}$, the search will be performed over a grid of values of h_k , say $\{h_{k,1}, \dots, h_{k,c(k)}\}$, simultaneously for $k = 1, \dots, K$, where $c(k)$ denotes the cardinality of the grid for h_k .

We will now present the large sample results of the proposed combined classifier $\tilde{\psi}_{N,h}$, as defined by (9.4.3) and (9.4.4).

9.4.3 Large Sample Results

Using the definitions of random (data-based) partitions of $\mathfrak{R}^d$ and shatter coefficients in Section 8.8, we will examine the large sample performance of the proposed kernel-based combined classifier with different smoothing parameters for each of the K classes.

Let $\mathcal{P}_n = \pi_n(T_n) = \{A_{n,1}, \dots, A_{n,2^L}\}$ be a random partition of $\mathfrak{R}^d$ induced by the training sample T_n , where $A_{n,i}$ is the i^{th} cell of the partition defined by

$$A_{n,i} = \{\mathbf{x} \in \mathfrak{R}^d : (\psi_{n,1}(\mathbf{x}), \dots, \psi_{n,L}(\mathbf{x})) = \mathbf{t}_i \text{ for one and only one } \mathbf{t}_i \in \{1, \dots, K\}^L\},$$

such that $A_{n,i} \cap A_{n,j} = \emptyset$ for $i \neq j$ and $\cup_{i=1}^{2^L} A_{n,i} = \mathfrak{R}^d$. Define $A_n[\mathbf{x}]$ as the unique cell that contains the point $\mathbf{x}$ and let $\mathbf{t}[\mathbf{x}]$ be the value of $\mathbf{t} \in \{1, \dots, K\}^L$ corresponding to this cell. Let $T_n^{(\text{obs})} = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)\}$ be the observed values of the training sample T_n . Associated with the partitioning rule π_n consider the nonrandom family of partitions of $\mathfrak{R}^d$ given by

$$\Omega_n = \{\pi : \pi = \pi_n(T_n^{(\text{obs})}), T_n^{(\text{obs})} \in \{\mathfrak{R}^d \times \{1, \dots, K\}\}^n\}$$

For any family of partitions Ω of $\mathfrak{R}^d$, let $\Delta_n(\Omega, n)$ be the n^{th} shatter coefficient of Ω defined in (8.8.1).

The following theorem gives the conditions under which the combined classifier

$\tilde{\psi}_{N,\mathbf{h}}$, obtained by minimizing (9.4.4), is at least as good as any individual classifier $\psi_{N,l}$, $l = 1, \dots, L$.

Theorem 9.3 (Variable Kernel Combined Classifier) *Consider the class of combined classifiers $\tilde{\psi}_{n,\mathbf{h}}$, $h_k \geq 0$ for $k = 1, \dots, K$, defined by (9.4.3) and let $\tilde{\psi}_{N,\mathbf{h}}$ be the combined classifier that minimizes (9.4.4). If, as N (and thus n and m) $\rightarrow \infty$,*

$$(a) \quad m^{-1} \log n \rightarrow 0,$$

$$(b) \quad L_n / \log n \rightarrow 0,$$

$$(c) \quad \lim_{n \rightarrow \infty} \text{err}_n(\psi_{n,l}) \text{ exists almost surely,}$$

$$(d) \quad n^{-1} \log \Delta_n(\Omega_n, n) \rightarrow 0, \text{ and}$$

$$(e) \quad \text{for every } \gamma > 0 \text{ and } \delta \in (0, 1),$$

$$\inf_{\mathbf{S} \subseteq \mathbb{R}^d: \mu(\mathbf{S}) \geq 1-\delta} \mu\{\mathbf{x} : \text{diam}(A_n[\mathbf{x}] \cap \mathbf{S}) > \gamma\} \rightarrow_{\text{a.s.}} 0,$$

where μ is the probability measure of $\mathbf{X}$, then $\tilde{\psi}_{N,\mathbf{h}}$ is, asymptotically, at least as good as any one of the individual classifiers in the following sense: for any individual classifier $\psi_{N,l}$ one has

$$\mathbf{P} \left\{ \limsup_{N \rightarrow \infty} \{ \text{err}_N(\tilde{\psi}_{N,\mathbf{h}}) - \text{err}_N(\psi_{N,l}) \} \leq 0 \right\} = 1$$

where $\text{err}_N(\tilde{\psi}_{N,\mathbf{h}}) = \mathbf{P}\{\tilde{\psi}_{N,\mathbf{h}}(\psi_{n,1}(\mathbf{x}), \dots, \psi_{n,L}(\mathbf{x})) \neq Y \mid (\mathbf{X}_i, Y_i), i = 1, \dots, N\}$.

The following is an immediate corollary to the above theorem.

Corollary: *Suppose that the conditions of Theorem 9.3 hold. If one of the individual classifiers is already strongly (weakly) consistent, then so is the combined*

classifier $\tilde{\psi}_{N,h}$, that is,

$$\text{err}_N(\tilde{\psi}_{N,h}) \rightarrow_{\text{a.s.}} \text{err}(\psi_B),$$

where $\text{err}(\psi_B)$ is the Bayes' error given by (8.2.2).

Proof of Theorem 9.3

First, we write

$$\begin{aligned} \text{err}_N(\tilde{\psi}_{N,h}) - \text{err}_N(\psi_{N,l}) &= \text{err}_N(\tilde{\psi}_{N,h}) - \inf_{\tilde{\psi}_{n,h} \in \tilde{\Psi}_{n,h}} \text{err}_n(\tilde{\psi}'_{n,h}) \\ &\quad + \inf_{\tilde{\psi}'_{n,h} \in \tilde{\Psi}_{n,h}} \text{err}_n(\tilde{\psi}'_{n,h}) - \text{err}_n(\tilde{\psi}_{n,0}) \\ &\quad + \text{err}_n(\tilde{\psi}_{n,0}) - \text{err}_n(\psi_{n,l}) \\ &\quad + \text{err}_n(\psi_{n,l}) - \text{err}_N(\psi_{N,l}) \\ &= \mathbf{I}_N + \mathbf{II}_n + \mathbf{III}_n + \mathbf{IV}_N, \end{aligned}$$

where

$$\text{err}_N(\tilde{\psi}_{N,h}) = \mathbf{P}\{\tilde{\psi}_{N,h}(\mathbf{X}) \neq Y \mid (\mathbf{X}_i, Y_i), i = 1, \dots, N\},$$

$$\text{err}_n(\tilde{\psi}_{n,h}) = \mathbf{P}\{\tilde{\psi}_{n,h}(\mathbf{X}) \neq Y \mid (\mathbf{X}_i, Y_i), i = 1, \dots, n\},$$

$$\text{err}_n(\psi_{n,l}) = \mathbf{P}\{\psi_{n,l}(\mathbf{X}) \neq Y \mid (\mathbf{X}_i, Y_i), i = 1, \dots, n\},$$

$$\text{and } \text{err}_N(\psi_{N,l}) = \mathbf{P}\{\psi_{N,l}(\mathbf{X}) \neq Y \mid (\mathbf{X}_i, Y_i), i = 1, \dots, N\}.$$

To prove the theorem, it is enough to show the following:

$$[\mathbf{A}] \quad \mathbf{I}_N \xrightarrow{\text{a.s.}} 0, \text{ as } N \rightarrow \infty,$$

$$[\mathbf{B}] \quad \mathbf{II}_n \leq 0,$$

$$[\text{C}] \limsup_{n \rightarrow \infty} \mathbf{III}_n \leq \text{a.s. } 0,$$

$$[\text{D}] \limsup_{N \rightarrow \infty} \mathbf{IV}_N = \text{a.s. } 0.$$

We will deal with each of the four parts separately.

[A] We want to show $\mathbf{I}_N \xrightarrow{\text{a.s.}} 0$, as $N \rightarrow \infty$.

$$\begin{aligned} \mathbf{I}_N &= \text{err}_N(\tilde{\psi}_{N,\mathbf{h}}) - \inf_{\tilde{\psi}'_{n,\mathbf{h}} \in \tilde{\Psi}_{n,\mathbf{h}}} \text{err}_n(\tilde{\psi}'_{n,\mathbf{h}}) \\ &= \left\{ \text{err}_N(\tilde{\psi}_{N,\mathbf{h}}) - \widehat{\text{err}}_N(\tilde{\psi}_{n,\mathbf{h}}) \right\} + \left\{ \widehat{\text{err}}_N(\tilde{\psi}_{n,\mathbf{h}}) - \inf_{\tilde{\psi}'_{n,\mathbf{h}} \in \tilde{\Psi}_{n,\mathbf{h}}} \text{err}_n(\tilde{\psi}'_{n,\mathbf{h}}) \right\} \\ &\leq 2 \sup_{\tilde{\psi}_{n,\mathbf{h}} \in \tilde{\Psi}_{n,\mathbf{h}}} | \widehat{\text{err}}_N(\tilde{\psi}_{n,\mathbf{h}}) - \text{err}_n(\tilde{\psi}_{n,\mathbf{h}}) |, \end{aligned} \quad (9.4.5)$$

using the results of Vapnik and Chervonenkis (1974), given in Lemma 8.1. For any $\epsilon > 0$, we have

$$\begin{aligned} \mathbf{P}(\mathbf{I}_N > \epsilon) &\leq \mathbf{P} \left\{ \sup_{\tilde{\psi}_{n,\mathbf{h}} \in \tilde{\Psi}_{n,\mathbf{h}}} | \widehat{\text{err}}_N(\tilde{\psi}_{n,\mathbf{h}}) - \text{err}_n(\tilde{\psi}_{n,\mathbf{h}}) | > \frac{\epsilon}{2} \right\} \\ &= \mathbf{E} \left\{ \mathbf{P} \left\{ \sup_{\tilde{\psi}_{n,\mathbf{h}} \in \tilde{\Psi}_{n,\mathbf{h}}} | \widehat{\text{err}}_N(\tilde{\psi}_{n,\mathbf{h}}) - \text{err}_n(\tilde{\psi}_{n,\mathbf{h}}) | > \frac{\epsilon}{2} \mid (\mathbf{X}_i, Y_i), i = 1, \dots, n \right\} \right\} \end{aligned} \quad (9.4.6)$$

where we will then look for a distribution-free bound for the inner conditional probability in (9.4.6). We will need the following results to bound the inner conditional probability.

Let $T_n^{(\text{obs})} = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)\}$ be the observed training sample of size n and let $\tilde{\Psi}_{n,\mathbf{h}}^*$ be the class of sets of the form

$$\bigcup_{k=1}^K \left\{ \{\mathbf{x} : \tilde{\psi}_{n,\mathbf{h}}(\psi_{n,1}(\mathbf{x}), \dots, \psi_{n,L}(\mathbf{x})) \neq k\} \times \{k\} \right\}, \quad \tilde{\psi}_{n,\mathbf{h}} \in \tilde{\Psi}_{n,\mathbf{h}}.$$

Define

$$\Delta(\tilde{\Psi}_{n,\mathbf{h}}^*, r) = \max_{B: B \in \{\mathbb{R}^d \times \{1, \dots, K\}\}^r} \left\{ \text{number of different sets in } \{B \cap \tilde{\psi}_{n,\mathbf{h}}^* \mid \tilde{\psi}_{n,\mathbf{h}}^* \in \tilde{\Psi}_{n,\mathbf{h}}^*\} \right\}$$

as the r^{th} shatter coefficient of $\tilde{\Psi}_{n,h}^*$. Then, by an extension of equation (8.8.4) to the K -class classification problem, we obtain

$$\begin{aligned} & \mathbf{P} \left\{ \sup_{\tilde{\psi}_{n,h} \in \tilde{\Psi}_{n,h}} | \widehat{\text{err}}_N(\tilde{\psi}_{n,h}) - \text{err}_n(\tilde{\psi}_{n,h}) | > \frac{\epsilon}{2} \mid (\mathbf{X}_i, Y_i), i = 1, \dots, n \right\} \\ & \leq 4 \sup_{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)} \Delta(\tilde{\Psi}_{n,h}^*, 2m) e^{-m\epsilon^2/32} \end{aligned} \quad (9.4.7)$$

The bound in (9.4.7) above is useless unless $\Delta(\tilde{\Psi}_{n,h}^*, 2m)$ is $e^{o(m)}$ or smaller.

Let $\tilde{\Psi}_{n,h+}$ be the subclass of $\tilde{\Psi}_{n,h}$ obtained by removing the single classifier $\tilde{\psi}_{n,0}$ from $\tilde{\Psi}_{n,h}$. Note that $\tilde{\psi}_{n,0}$ is the combined classifier defined by equation (9.4.3) when a smoothing parameter $h_k = 0$ for some $k = 1, \dots, K$. Similarly, let $\tilde{\Psi}_{n,h+}^*$ be the subclass of $\tilde{\Psi}_{n,h}^*$ by removing $\tilde{\psi}_{n,0}$. Then from the third property of shatter coefficients in Section 8.7, we have

$$\Delta(\tilde{\Psi}_{n,h}^*, 2m) \leq \Delta(\tilde{\Psi}_{n,h+}^*, 2m) + 1.$$

In order to bound $\Delta(\tilde{\Psi}_{n,h+}^*, 2m)$ and subsequently $\Delta(\tilde{\Psi}_{n,h}^*, 2m)$, we will need to define the members of the subclass $\tilde{\Psi}_{n,h+}$ as follows:

$$\tilde{\psi}_{n,h}(\mathbf{x}) \equiv \tilde{\psi}_{n,h}(\psi_{n,1}(\mathbf{x}), \dots, \psi_{n,L}(\mathbf{x}))$$

$$= \begin{cases} 1 & \text{if} & f_{1,2}(\mathbf{x}) \geq 0, f_{1,3}(\mathbf{x}) \geq 0, \dots, f_{1,K}(\mathbf{x}) \geq 0, \\ 2 & \text{if} & f_{2,1}(\mathbf{x}) \geq 0, f_{2,3}(\mathbf{x}) \geq 0, \dots, f_{2,K}(\mathbf{x}) \geq 0, \\ \vdots & \vdots & \vdots \\ K-1 & \text{if} & f_{K-1,1}(\mathbf{x}) \geq 0, \dots, f_{K-1,K-2}(\mathbf{x}) \geq 0, f_{K-1,K}(\mathbf{x}) \geq 0, \\ K & \text{otherwise,} \end{cases}$$

where

$$\begin{aligned}
f_{a,b}(\mathbf{x}) &\equiv f_{a,b}(\mathbf{x}, (\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_n, Y_n)) \\
&= \sum_{\{i: 1 \leq i \leq n, Y_i = a\}} \mathbf{K}_{h_a} \left(\sum_{l=1}^L \mathbf{I}\{\psi_{n,l}(\mathbf{X}_i) \neq \psi_{n,l}(\mathbf{x})\} \right) \\
&\quad - \sum_{\{i: 1 \leq i \leq n, Y_i = b\}} \mathbf{K}_{h_b} \left(\sum_{l=1}^L \mathbf{I}\{\psi_{n,l}(\mathbf{X}_i) \neq \psi_{n,l}(\mathbf{x})\} \right) \\
&= \sum_{1 \leq i \leq n} \mathbf{I}\{Y_i = a\} \exp \left\{ -h_a^{-\alpha} \left(\sum_{l=1}^L \mathbf{I}\{\psi_{n,l}(\mathbf{X}_i) \neq \psi_{n,l}(\mathbf{x})\} \right)^\alpha \right\} \\
&\quad - \sum_{1 \leq i \leq n} \mathbf{I}\{Y_i = b\} \exp \left\{ -h_b^{-\alpha} \left(\sum_{l=1}^L \mathbf{I}\{\psi_{n,l}(\mathbf{X}_i) \neq \psi_{n,l}(\mathbf{x})\} \right)^\alpha \right\} \quad (9.4.8) \\
&= \sum_{i=1}^{n_a+n_b} e^{-h_a^{-\alpha} (\sum_{l=1}^L \mathbf{I}\{\psi_{n,l}(\mathbf{X}_i^*) \neq \psi_{n,l}(\mathbf{x})\})^\alpha} \times \\
&\quad \left[\mathbf{I}\{Y_i^* = a\} - \mathbf{I}\{Y_i^* = b\} e^{(h_a^{-\alpha} - h_b^{-\alpha}) (\sum_{l=1}^L \mathbf{I}\{\psi_{n,l}(\mathbf{X}_i^*) \neq \psi_{n,l}(\mathbf{x})\})^\alpha} \right] \\
&= \sum_{i=1}^{n_a+n_b} e^{-h_a^{-\alpha} (\sum_{l=1}^L \mathbf{I}\{\psi_{n,l}(\mathbf{X}_i^*) \neq \psi_{n,l}(\mathbf{x})\})^\alpha} W_i, \quad (9.4.9)
\end{aligned}$$

and

$$\begin{aligned}
n_a &= \sum_{i=1}^n \mathbf{I}\{Y_i = a\}, \quad n_b = \sum_{i=1}^n \mathbf{I}\{Y_i = b\}, \\
(\mathbf{X}_i^*, Y_i^*) &= (\mathbf{X}_i, Y_i) \mathbf{I}\{Y_i = a \text{ or } Y_i = b\} \text{ and} \\
W_i &= \mathbf{I}\{Y_i^* = a\} - \mathbf{I}\{Y_i^* = b\} e^{(h_a^{-\alpha} - h_b^{-\alpha}) (\sum_{l=1}^L \mathbf{I}\{\psi_{n,l}(\mathbf{X}_i^*) \neq \psi_{n,l}(\mathbf{x})\})^\alpha}.
\end{aligned}$$

We will now use some of the results on kernel complexity that appear in Devroye et. al. (1996, Chapter 25).

Definition 9.1 (Kernel Complexity) Let $\kappa_n(a, b)$ be the kernel complexity defined by

$$\begin{aligned}
\kappa_n(a, b) &= \sup_{\mathbf{x}, (\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)} \{ \text{number of sign changes of } f_{a,b}(\mathbf{x}) \text{ as } h_a \text{ and } h_b \\
&\quad \text{vary jointly from 0 to infinity} \}.
\end{aligned}$$

Now for each fixed $\mathbf{X}_j$, $j = 1, \dots, m$ and each $a \in \{1, \dots, K-1\}$, the set of $(K-1)$ functions

$$f_{a,1}(\mathbf{X}_j), \dots, f_{a,a-1}(\mathbf{X}_j), f_{a,a+1}(\mathbf{X}_j), \dots, f_{a,K}(\mathbf{X}_j)$$

jointly change sign at most

$$\kappa_n(a, 1) \times \dots \times \kappa_n(a, a-1) \times \kappa_n(a, a+1) \times \dots \times \kappa_n(a, K) = \prod_{\{1 \leq b \leq K, b \neq a\}} \kappa_n(a, b)$$

times, as h_a and h_b vary jointly from 0 to infinity. Therefore, for a testing set of size m , there are at most

$$m \times \prod_{\{1 \leq b \leq K, b \neq a\}} \kappa_n(a, b) + 1$$

different signs associated with the set of functions $f_{a,1}, \dots, f_{a,a-1}, f_{a,a+1}, \dots, f_{a,K}$, considered jointly. Since a takes $K-1$ values in the set $\{1, \dots, K-1\}$, we have

$$\Delta(\tilde{\Psi}_{n,h^+}^*, m) \leq \prod_{\{1 \leq a \leq K-1\}} [m \times \prod_{\{1 \leq b \leq K, b \neq a\}} \kappa_n(a, b) + 1].$$

We can use the following result from page 432 of Devroye et. al. (1996) to find a bound for $\kappa_n(a, b)$.

Lemma 9.3 *Let $a_1, \dots, a_n$ be fixed real numbers, and let $b_1, \dots, b_n$ be different non-negative reals. Then if $\alpha \neq 0$, the function*

$$g(x) = \sum_{i=1}^n a_i e^{(-b_i x^\alpha)}, \quad x \geq 0,$$

is either identically zero, or takes the value 0 at most n times.

Noting that the two parts in (9.4.8) are either identically zero, or take the value zero at most n_a or n_b times, we can find a bound for $\kappa_n(a, b)$ as follows,

$$\kappa_n(a, b) \leq 1 + 2 + \dots + [\max(n_a, n_b) + 1]$$

$$\begin{aligned}
&= 1 + \frac{(n_{\max})(n_{\max} + 1)}{2}, \text{ where } n_{\max} = \max(n_a, n_b) \\
&\leq \frac{n(n + 1)}{2} \text{ for all } a \text{ and } b \\
&\leq n^2.
\end{aligned}$$

We will then obtain the following loose upper-bound,

$$\begin{aligned}
\Delta(\Psi_{n, \mathbf{h}_k^+}^*, 2m) &\leq \prod_{\{1 \leq a \leq K-1\}} [2m \times \prod_{\{1 \leq b \leq K, b \neq a\}} n^2 + 1] \\
&= \prod_{\{1 \leq a \leq K-1\}} [2mn^{2(K-1)} + 1] \\
&= [2mn^{2(K-1)} + 1]^{K-1}.
\end{aligned} \tag{9.4.10}$$

From (9.4.6) and (9.4.7), we have shown that for any $\epsilon > 0$,

$$\mathbf{P}(\mathbf{I}_N > \epsilon) \leq 4 \left([2mn^{2(K-1)} + 1]^{K-1} + 1 \right) e^{-m\epsilon^2/32}.$$

An application of the Borel-Cantelli lemma, and the fact that $m^{-1} \log n \rightarrow 0$ as $N \rightarrow \infty$, shows that $\mathbf{I}_N \rightarrow_{\text{a.s.}} 0$ as $N \rightarrow \infty$.

[B] To show that $\mathbf{II}_n \leq 0$, we will have to observe that by the definition of infimum

$$\mathbf{II}_n \stackrel{\text{def}}{=} \inf_{\tilde{\psi}'_{n, \mathbf{h}} \in \tilde{\Psi}_{n, \mathbf{h}}} \text{err}_n(\tilde{\psi}'_{n, \mathbf{h}}) - \text{err}_n(\tilde{\psi}_{n, \mathbf{0}}) \leq 0$$

holds trivially.

[C] To prove $\limsup_{n \rightarrow \infty} \mathbf{III}_n \leq \text{a.s. } 0$, we will use some of the results of Lugosi and Nobel (1996) and Mojirsheibani (1999). For the ease of notation, we will focus on the two-class problem only in the following theorem. That is, we assume that $(\mathbf{X}, Y)$ is a random pair in $\mathbb{R}^d \times \{0, 1\}$.

Lemma 9.4 *Let $\tilde{\psi}_n$ be a combined classifier defined according to (9.3.1), based on the training sample of size n . Then, under conditions (b), (d) and (e) of Theorem 9.3,*

$$\limsup_{n \rightarrow \infty} \{ \text{err}_n(\tilde{\psi}_n) - \text{err}_n(\psi_{n,l}) \} \leq 0 \text{ a.s.},$$

where $\text{err}_n(\tilde{\psi}_n) = \mathbf{P}\{\tilde{\psi}_n(\psi_{n,1}(\mathbf{X}), \dots, \psi_{n,L}(\mathbf{X})) \neq Y \mid (\mathbf{X}_i, Y_i), i = 1, \dots, n\}$ and $\text{err}_n(\psi_{n,l}) = \mathbf{P}\{\psi_{n,l}(\mathbf{X}) \neq Y \mid (\mathbf{X}_i, Y_i), i = 1, \dots, n\}$.

Proof of Lemma 9.4

Let $\tilde{\psi}^*$ be defined by (9.3.5) with N replaced by n and observe that

$$\begin{aligned} \text{err}_n(\tilde{\psi}_n) - \text{err}_n(\psi_{n,l}) &= \text{err}_n(\tilde{\psi}_n) - \text{err}_n(\tilde{\psi}^*) + \text{err}_n(\tilde{\psi}^*) - \text{err}_n(\psi_{n,l}) \\ &= A_n + B_n \end{aligned}$$

where

$$A_n = \text{err}_n(\tilde{\psi}_n) - \text{err}_n(\tilde{\psi}^*)$$

$$B_n = \text{err}_n(\tilde{\psi}^*) - \text{err}_n(\psi_{n,l}) \text{ and}$$

$$\text{err}_n(\tilde{\psi}^*) = \mathbf{P}\{\tilde{\psi}^*(\psi_{n,1}(\mathbf{X}), \dots, \psi_{n,L}(\mathbf{X})) \neq Y \mid (\mathbf{X}_i, Y_i), i = 1, \dots, n\}.$$

From part (i) of Lemma 9.2 we have $B_n \leq_{\text{a.s.}} 0$ by setting $\tilde{\psi} = \psi_{n,l}$. To show

$A_n \rightarrow \text{a.s. } 0$ as $n \rightarrow \infty$, we need to write the combined classifier $\tilde{\psi}_n$ as

$$\begin{aligned} \tilde{\psi}_n &= \begin{cases} 1 & \text{if } \sum_{i=1}^n \prod_{l=1}^L \mathbf{I}\{\psi_{n,l}(\mathbf{X}_i) = \psi_{n,l}(\mathbf{x})\} \mathbf{I}\{Y_i = 1\} \\ & > \sum_{i=1}^n \prod_{l=1}^L \mathbf{I}\{\psi_{n,l}(\mathbf{X}_i) = \psi_{n,l}(\mathbf{x})\} \mathbf{I}\{Y_i = 0\} \\ 0 & \text{otherwise} \end{cases} \\ &= \begin{cases} 1 & \text{if } \pi^{(1)}(\psi_{n,1}(\mathbf{x}), \dots, \psi_{n,L}(\mathbf{x})) \\ & > \pi^{(0)}(\psi_{n,1}(\mathbf{x}), \dots, \psi_{n,L}(\mathbf{x})), \\ 0 & \text{otherwise} \end{cases} \end{aligned} \tag{9.4.11}$$

where

$$\pi^{(k)}(\psi_{n,1}(\mathbf{x}), \dots, \psi_{n,L}(\mathbf{x})) = \frac{\sum_{i=1}^n \prod_{l=1}^L \mathbf{I}\{\psi_{n,l}(\mathbf{X}_i) = \psi_{n,l}(\mathbf{x})\} \mathbf{I}\{Y_i = k\}}{n\mu(A_n[\mathbf{x}])} \quad (9.4.12)$$

for $k = 0, 1$ and $(\pi^{(1)} + \pi^{(0)})$ does not necessarily equal to 1. Now

$$\begin{aligned} & \prod_{l=1}^L \mathbf{I}\{\psi_{n,l}(\mathbf{X}_i) = \psi_{n,l}(\mathbf{x})\} \\ &= \mathbf{I}\{(\psi_{n,1}(\mathbf{X}_i), \dots, \psi_{n,L}(\mathbf{X}_i)) = (\psi_{n,1}(\mathbf{x}), \dots, \psi_{n,L}(\mathbf{x}))\} \\ &= \mathbf{I}\{(\psi_{n,1}(\mathbf{X}_i), \dots, \psi_{n,L}(\mathbf{X}_i)) = \mathbf{t}_n[\mathbf{x}]\} \\ &= \mathbf{I}\{\mathbf{X}_i \in A_n[\mathbf{x}]\}. \end{aligned}$$

Therefore (9.4.12) can be rewritten as

$$\pi^{(k)}(\psi_{n,1}(\mathbf{x}), \dots, \psi_{n,L}(\mathbf{x})) = \frac{\sum_{i=1}^n \mathbf{I}\{\mathbf{X}_i \in A_n[\mathbf{x}]\} \mathbf{I}\{Y_i = k\}}{n\mu(A_n[\mathbf{x}])}, \quad (9.4.13)$$

and consequently one has

$$\tilde{\psi}_n = \begin{cases} 1 & \text{if } \frac{\sum_{i=1}^n \mathbf{I}\{\mathbf{X}_i \in A_n[\mathbf{x}]\} \mathbf{I}\{Y_i = 1\}}{n\mu(A_n[\mathbf{x}])} \\ & > \frac{\sum_{i=1}^n \mathbf{I}\{\mathbf{X}_i \in A_n[\mathbf{x}]\} \mathbf{I}\{Y_i = 0\}}{n\mu(A_n[\mathbf{x}])}. \\ 0 & \text{otherwise} \end{cases} \quad (9.4.14)$$

From part (ii) of Lemma 9.2 it is enough to show that

$$\int_{\mathbb{R}^d} |\pi^{(k)}(\psi_{n,1}(\mathbf{x}), \dots, \psi_{n,L}(\mathbf{x})) - \mathbf{P}(Y = k | \mathbf{X} = \mathbf{x})| \mu(d\mathbf{x}) \longrightarrow 0 \text{ a.s.}$$

for $k = 0, 1$.

Fix $\epsilon > 0$ and let $r : \mathbb{R}^d \rightarrow \mathbb{R}$ be a continuous function with bounded support such that

$$\int_{\mathbb{R}^d} |\mathbf{P}(Y = k | \mathbf{X} = \mathbf{x}) - r(\mathbf{x})| \mu(d\mathbf{x}) < \epsilon. \quad (9.4.15)$$

We note that such functions r exist by the following lemma.

Lemma 9.5 *For every probability measure μ on $\mathbb{R}^d$, the set of continuous functions with bounded support is dense in $L_p(\mu)$, where $L_p(\mu)$ denotes the set of all measurable functions satisfying $\int |f|^p d\mu < \infty$ for a fixed number $p \geq 1$ and measure space $(S, \mathcal{F}, \mu)$. In other words, for every $\epsilon > 0$ and $f \in L_p$ there is a continuous function with bounded support $g \in L_p$ such that*

$$\int |f - g|^p d\mu < \epsilon.$$

Define the auxiliary functions

$$\bar{\pi}^{(k)}(\mathbf{x}) = \frac{\mathbf{E}(\mathbf{I}\{Y = k\}\mathbf{I}\{\mathbf{X} \in A_n[\mathbf{x}]\}|T_n)}{\mu(A_n[\mathbf{x}])}$$

and

$$\bar{r}(\mathbf{x}) = \frac{\mathbf{E}(r(\mathbf{x})\mathbf{I}\{\mathbf{X} \in A_n[\mathbf{x}]\}|T_n)}{\mu(A_n[\mathbf{x}])},$$

noting that both $\bar{\pi}^{(k)}$ and $\bar{r}$ are piecewise constant on the cells of the random partition $\mathcal{P}_n$. We will then decompose the error of the first term on the right-hand side of (9.3.6) as follows:

$$\begin{aligned} & |\mathbf{P}(Y = 1|\mathbf{X} = \mathbf{x}) - \pi^{(1)}(\psi_{n,1}(\mathbf{x}), \dots, \psi_{n,L}(\mathbf{x}))| \\ & \leq |\mathbf{P}(Y = 1|\mathbf{X} = \mathbf{x}) - r(\mathbf{x})| \\ & \quad + |r(\mathbf{x}) - \bar{r}(\mathbf{x})| \\ & \quad + |\bar{r}(\mathbf{x}) - \bar{\pi}^{(1)}(\mathbf{x})| \\ & \quad + |\bar{\pi}^{(1)}(\mathbf{x}) - \pi^{(1)}(\psi_{N,1}(\mathbf{x}), \dots, \psi_{N,L}(\mathbf{x}))| \\ & = |\mathbf{F}_n| + |\mathbf{G}_n| + |\mathbf{H}_n| + |\mathbf{J}_n| \end{aligned} \tag{9.4.16}$$

where

$$\mathbf{F}_n = \mathbf{P}(Y = k|\mathbf{X} = \mathbf{x}) - r(\mathbf{x}),$$

$$\mathbf{G}_n = r(\mathbf{x}) - \tilde{r}(\mathbf{x}),$$

$$\mathbf{H}_n = \tilde{r}(\mathbf{x}) - \tilde{\pi}^{(1)}(\mathbf{x}),$$

$$\mathbf{J}_n = \tilde{\pi}^{(1)}(\mathbf{x}) - \pi^{(1)}(\psi_{N,1}(\mathbf{x}), \dots, \psi_{N,L}(\mathbf{x})).$$

Taking integrals of the terms in (9.4.16) we have:

(a) $\int_{\mathbb{R}^d} |\mathbf{F}_n| \mu(d\mathbf{x}) < \epsilon$ by the definition of r .

(b) For the third term

$$\begin{aligned} \int_{\mathbb{R}^d} |\mathbf{H}_n| \mu(d\mathbf{x}) &= \sum_{A \in \mathcal{P}_n} \int_A \left| \frac{\mathbf{E}(\mathbf{I}\{Y = 1\} \mathbf{I}\{\mathbf{X} \in A\} | T_n)}{\mu(A)} \right. \\ &\quad \left. - \frac{\mathbf{E}(r(\mathbf{x}) \mathbf{I}\{\mathbf{X} \in A\} | T_n)}{\mu(A)} \right| \mu(d\mathbf{x}) \\ &= \sum_{A \in \mathcal{P}_n} \left| \int_A \mathbf{P}(Y = 1 | \mathbf{X} = \mathbf{x}) \mu(d\mathbf{x}) - \int_A r(\mathbf{x}) \mu(d\mathbf{x}) \right| \\ &\leq \sum_{A \in \mathcal{P}_n} \int_A |\mathbf{P}(Y = 1 | \mathbf{X} = \mathbf{x}) - r(\mathbf{x})| \mu(d\mathbf{x}) \\ &= \int_{\mathbb{R}^d} |\mathbf{P}(Y = 1 | \mathbf{X} = \mathbf{x}) - r(\mathbf{x})| \mu(d\mathbf{x}) < \epsilon \end{aligned}$$

by (9.4.15).

(c) For the fourth term on the right-hand side of (9.4.16), let ν be the distribution of $(\mathbf{X}, \mathbf{I}\{Y = k\})$ on $\mathbb{R}^d \times \{0, 1\}$ and let ν_n be the empirical measure of $(\mathbf{X}_1, \mathbf{I}\{Y_1 = k\}), \dots, (\mathbf{X}_n, \mathbf{I}\{Y_n = k\})$. That is, for a fixed measurable set $A \subset \mathbb{R}^d \times \{0, 1\}$, $\nu(A) = \mathbf{P}\{(\mathbf{X}, Y) \in A\}$ and $\nu_n(A) = \frac{1}{n} \sum_{i=1}^n \mathbf{I}\{(\mathbf{X}_i, Y_i) \in A\}$. For each partition $\mathcal{P} = \{A_{n,1}, \dots, A_{n,2L}\} \in \Omega_n$, define a partition $\tilde{\mathcal{P}}$ of $\mathbb{R}^d \times \{0, 1\}$ by

$$\tilde{\mathcal{P}} = \{A_{n,1} \times \{0\}, \dots, A_{n,2L} \times \{0\}\} \cup \{A_{n,1} \times \{1\}, \dots, A_{n,2L} \times \{1\}\}$$

and let $\mathcal{B}_n = \{\tilde{\mathcal{P}} : \mathcal{P} \in \Omega_n\}$. Then

$$\int_{\mathbb{R}^d} |\tilde{\pi}^{(1)}(\mathbf{x}) - \pi^{(1)}(\psi_{n,1}(\mathbf{x}), \dots, \psi_{n,L}(\mathbf{x}))| \mu(d\mathbf{x})$$

$$\begin{aligned}
&= \sum_{A \in \mathcal{P}_n} \int_A \left| \frac{\sum_{i=1}^n \mathbf{I}\{Y_i = 1\} \mathbf{I}\{X_i \in A\}}{n \mu(A)} - \frac{\mathbf{E}(\mathbf{I}\{Y = 1\} \mathbf{I}\{X \in A\} | T_n)}{\mu(A)} \right| \mu(d\mathbf{x}) \\
&= \sum_{A \in \mathcal{P}_n} \left| \frac{1}{n} \sum_{i=1}^n \mathbf{I}\{Y_i = 1\} \mathbf{I}\{X_i \in A\} - \mathbf{E}(\mathbf{I}\{Y = 1\} \mathbf{I}\{X \in A\} | T_n) \right| \\
&= \sum_{A \in \mathcal{P}_n} |\nu_n(A \times \{1\}) - \nu(A \times \{1\})| \\
&\leq \sup_{\mathcal{P} \in \Omega_n} \sum_{A \in \mathcal{P}} |\nu_n(A \times \{1\}) - \nu(A \times \{1\})| \\
&= \sup_{\tilde{\mathcal{P}} \in \mathcal{B}_n} \sum_{B \in \tilde{\mathcal{P}}} |\nu_n(B) - \nu(B)|. \tag{9.4.17}
\end{aligned}$$

From condition (d) of Theorem 9.3 and the Corollary of Lemma 8.3, the right-hand side of (9.4.17) converges to zero with probability one. It then follows that

$$\int_{\mathfrak{R}^d} |\mathbf{J}_n| \mu(d\mathbf{x}) \longrightarrow 0 \text{ a.s.}$$

(d) For the second term on the right-hand side of (9.4.16) we get

$$\begin{aligned}
&\int_{\mathfrak{R}^d} |r(\mathbf{x}) - \tilde{r}(\mathbf{x})| \mu(d\mathbf{x}) \\
&= \sum_{A: \mu(A) \neq 0} \int_A \left| r(\mathbf{x}) - \frac{\mathbf{E}(r(\mathbf{x}) \mathbf{I}\{X \in A\} | T_n)}{\mu(A)} \right| \mu(d\mathbf{x}) \\
&= \sum_{A: \mu(A) \neq 0} \frac{1}{\mu(A)} \int_A |r(\mathbf{x}) \mu(A) - \mathbf{E}(r(\mathbf{x}) \mathbf{I}\{X \in A\} | T_n)| \mu(d\mathbf{x}) \\
&= \sum_{A: \mu(A) \neq 0} \frac{1}{\mu(A)} \int_A \left| r(\mathbf{x}) \int_A \mu(d\mathbf{y}) - \int_A r(\mathbf{y}) \mu(d\mathbf{y}) \right| \mu(d\mathbf{x}) \\
&\leq \sum_{A: \mu(A) \neq 0} \frac{1}{\mu(A)} \int_A \int_A |r(\mathbf{x}) - r(\mathbf{y})| \mu(d\mathbf{x}) \mu(d\mathbf{y}) \tag{9.4.18}
\end{aligned}$$

by Fubini's theorem.

Fix $\delta \in (0, 1)$ and let $\gamma > 0$ be chosen so that if $A \subseteq \mathfrak{R}^d$ satisfies $\text{diam}(A) < \gamma$, then $|r(\mathbf{x}) - r(\mathbf{y})| < \delta$ for every $\mathbf{x}, \mathbf{y} \in A$. In addition, let $K < \infty$ be a constant such that $|r(\mathbf{x})| \leq K$ for every $\mathbf{x} \in \mathfrak{R}^d$. Let $\mathbf{S} \subset \mathfrak{R}^d$ be such that $\mu(\mathbf{S}) \geq 1 - \delta$. If $\text{diam}(A \cap \mathbf{S}) \geq \gamma$, then

$$\frac{1}{\mu(A)} \int_A \int_A |r(\mathbf{x}) - r(\mathbf{y})| \mu(d\mathbf{x}) \mu(d\mathbf{y}) \leq 2K\mu(A). \tag{9.4.19}$$

If, on the other hand, $\text{diam}(A \cap \mathbf{S}) < \gamma$, then we can write

$$\frac{1}{\mu(A)} \int_A \int_A |r(\mathbf{x}) - r(\mathbf{y})| \mu(d\mathbf{x}) \mu(d\mathbf{y}) \leq \delta \mu(A). \quad (9.4.20)$$

Summing over the cells $A \in \mathcal{P}_n$ with $\mu(A) > 0$ and substituting (9.4.19) and (9.4.20) into (9.4.18), we have shown that

$$\begin{aligned} & \int_{\mathbb{R}^d} |r(\mathbf{x}) - \tilde{r}(\mathbf{x})| \mu(d\mathbf{x}) \\ & \leq \sum_{A \in \mathcal{P}_n: \text{diam}(A \cap \mathbf{S}) \geq \gamma} 2K\mu(A) + \sum_{A \in \mathcal{P}_n: \text{diam}(A \cap \mathbf{S}) < \gamma} \delta \mu(A) \\ & \leq 2K\mu\{\mathbf{x} : \text{diam}(A_n[\mathbf{x}] \cap \mathbf{S}) \geq \gamma\} + \delta. \end{aligned} \quad (9.4.21)$$

Taking the infimum of both sides of (9.4.21) over $\mathbf{S} \subset \mathbb{R}^d$ with $\mu(\mathbf{S}) \geq 1 - \delta$ and letting $n \rightarrow \infty$ gives

$$\int_{\mathbb{R}^d} |r(\mathbf{x}) - \tilde{r}(\mathbf{x})| \mu(d\mathbf{x}) \leq \delta \text{ a.s.}$$

by condition (e) of Theorem 9.3. That is,

$$\limsup_{n \rightarrow \infty} \int_{\mathbb{R}^d} |r(\mathbf{x}) - \tilde{r}(\mathbf{x})| \mu(d\mathbf{x}) \leq \delta \text{ a.s.}$$

In summary, we have shown that

$$\limsup_{n \rightarrow \infty} \int_{\mathbb{R}^d} |\mathbb{P}(Y = 1 | \mathbf{X} = \mathbf{x}) - \pi^{(1)}(\psi_{n,1}(\mathbf{x}), \dots, \psi_{n,L}(\mathbf{x}))| \mu(d\mathbf{x}) \leq 2\epsilon + \delta \text{ a.s.}$$

By symmetry, we can also show that

$$\limsup_{n \rightarrow \infty} \int_{\mathbb{R}^d} |\mathbb{P}(Y = 0 | \mathbf{X} = \mathbf{x}) - \pi^{(0)}(\psi_{n,1}(\mathbf{x}), \dots, \psi_{n,L}(\mathbf{x}))| \mu(d\mathbf{x}) \leq 2\epsilon + \delta \text{ a.s.}$$

Since ϵ and δ are arbitrary, the proof of Lemma 9.4 is complete. ■

Noting that $\tilde{\psi}_n$ is the same as $\tilde{\psi}_{n,0}$ defined by equation (9.4.3) when one or more of the smoothing parameters is zero (that is, $h_k = 0$ for some k , $k = 1, \dots, K$), it follows from Lemma 9.4 that

$$\limsup_{n \rightarrow \infty} \mathbf{III}_n \leq \text{a.s. } 0.$$

[D] Finally, we want to show that $\lim_{N \rightarrow \infty} \mathbf{IV}_N = \text{a.s. } 0$, which follows immediately from condition (c) of Theorem 9.3 and the fact that $n \leq N$.

Hence, parts [A]-[D] complete the proof of Theorem 9.3.

■

Chapter 10

Conclusions and Future Research

10.1 Conclusions

In Part I of the thesis we studied the estimation of a population total or a vector of finite population parameters using auxiliary information that may be available at the estimation stage. These auxiliary information can be incorporated into the estimation procedure by using calibration methods, such as poststratification and generalised regression (GREG), that make adjustments to the basic survey design weights. We then use these calibration adjusted weights in forming estimates of means and totals for the variables of interest collected in the survey. In particular, we studied the estimation of the variance of calibration estimators of a population total or a vector of finite population parameters, such as regression parameters, using the Estimating Function (EF) method of Hu and Kalbfleisch (2000). We then apply this EF method to the jackknife procedure for both the full response case and the nonresponse case, where some data in the survey are missing.

For the full response case, we presented the asymptotic normality of the calibration estimator of a population total and the asymptotic consistency of the alternative Taylor linearization variance estimator that involves the calibration weights. We also showed that the EF resampling variance estimators are identical to linearized resampling variance estimators.

For the estimation of a vector of finite population parameters under the full response case, we presented asymptotic design-based consistency and normality of the basic estimator of the vector of finite population parameters. We showed that the EF jackknife variance estimator of the calibration estimator of the vector of finite population parameters is approximately equal to the linearized jackknife and the Taylor linearization variance estimator. This approximation holds for the variance estimator of the calibration estimator of a subset of the vector of finite population parameters.

For unit nonresponse with multiple weighting classes, we studied the calibration estimator of a population total or a vector of finite population parameters. We showed that the EF jackknife variance estimator and the linearized jackknife variance estimator are approximately the same.

For item nonresponse, we use the idea of adjusting the imputed values whenever a cluster is deleted and construct “correct” jackknife variance estimators for the calibration estimator of a population total or a vector of finite population parameters, under mean imputation. We showed that the EF jackknife variance estimator is approximately equal to the linearized jackknife variance estimator.

We evaluated the loss in efficiency of the BRR variance estimator in the linear case, when two or more groups are employed to obtain balanced replicates.

In Part II of the thesis we examined the coverage accuracy of confidence intervals for the mean, by comparing the studentized t -statistic with an alternative approximate pivotal statistic proposed by Godambe and Thompson (1999).

In Part III of the thesis, we extended the method of kernel-based combined classification procedures (Mojirsheibani, 2000) to accomodate varying smoothing parameters.

10.2 Future Research

We will now discuss briefly some areas of further research.

Imputation

In Part I of the thesis we considered only mean imputation. Yung and Rao (2000) and Yung, Hidioglou and Rao (2001) have used the Rao-Shao jackknife method to construct linearized jackknife variance estimators for the calibration (GREG) estimator under mean and hot-deck imputation. Results for the EF jackknife method should be obtained for hot-deck imputation and other imputation methods.

Dual Frames

Lohr and Rao (2000) studied jackknife variance estimation in dual frame surveys, where samples are drawn independently from two overlapping frames that together cover the population of interest. Perhaps the EF jackknife method can be applied to this dual frame sampling setup.

Combined Classification

For combined classification we may want to use the idea of iterated combined classification suggested in Mojirsheibani (1999). Suppose we have L_2 different methods available for combining the individual classifiers. We combine the L different classifiers to produce the first combined classifier, which may be regarded as the $(L + 1)^{\text{th}}$ classifier. In the second iteration we then combine the available $(L + 1)$ classifiers to produce a new combined classifier and the procedure continues. In the i^{th} iteration, $1 \leq i \leq L_2$, we combine the available $(L + i - 1)$ classifiers to produce the i^{th} combined classifier. We would expect the i^{th} combined classifier to outperform all of the previous classifiers.

References

- Binder, D.A. (1983). On the Variances of Asymptotically Normal Estimators from Complex Surveys, *International Statistical Review*, **51**, 279-292.
- Binder, D.A. (1996). Linearization methods for single phase and two phase samples: A cookbook approach, *Survey Methodology*, **22**, 17-22.
- Boos, D.D. (1992). On Generalized Score Tests, *Amer. Statist.*, **46**, 327-333.
- Breiman, L. (1995). Stacked Regression, *Machine Learning*, **24**, 49-64.
- Breiman, L., Friedman, J., Olsen, R. and Stone, C. (1984). *Classification and Regression Trees*. Wadsworth International, Belmont, CA.
- Chen, J., Rao, J.N.K. and Sitter, R.R. (2000). Efficient Random Imputation for Missing Data in Complex Surveys, *Statistica Sinica*, **10**, 1153-1169.
- Cochran, W.G. (1977). *Sampling Techniques*. 3rd ed. New York: Wiley.
- Deville, J.C., and Särndal, C.E. (1992). Calibration Estimators in Survey Sampling, *Journal of the American Statistical Association*, **87**, 376-382.
- Devroye, L. (1988). Automatic Pattern Recognition: A Study of the Probability of Error, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **10**, 530-543.
- Devroye, L. and Krzyzak, A. (1989). An equivalence theorem for L_1 convergence of the kernel regression estimate, *Journal of Statistical Planning and Inference*, **23**, 71-82.

- Devroye, L., Györfi, L., and Lugosi, G. (1996). *A Probabilistic Theory of Pattern Recognition*. Springer-Verlag, New York.
- Dippo, C.S., Fay, R.E., and Morganstein, D.H. (1984). Computing variances from complex samples with replicate weights. In *Proc. Survey Res. Meth. Sect., Am. Statist. Assoc.*, pp. 489-494. Washington DC: American Statistical Association.
- Efron, B. (1979). Bootstrap Methods: Another look at Jackknife, *The Annals of Statistics*, **7**, 1-26.
- Fisher, R. (1936). The case of multiple measurements in taxonomic problems. *Annals of Eugenics*, **7**, part II:179-188.
- Fix, E. and Hodges, J. (1951). Discriminatory Analysis. Nonparametric discrimination: Consistency properties. Technical Report 4, Project Number 21-49-004, USAF School of Aviation Medicine, Randolph Field, TX.
- Flyer, P., Rust, K. and Morganstein, D. (1989). Complex Survey Variance Estimation and Contingency Table Analysis using Replication, In *Proc. Survey Res. Meth. Sect., Am. Statist. Assoc.*, pp.110-119. Alexandria, VA: American Statistical Association.
- Godambe, V.P., and Thompson, M.E. (1986). Parameters of Superpopulation and Survey Population: Their Relationships and Estimation, *International Statistical Review*, **54**, 127-138.
- Godambe, V.P., and Thompson, M.E. (1999). A New Look at Confidence Intervals in Survey Sampling, *Survey Methodology*, **25**, 161-173.

- Gupta, V.K. and Nigam, A.K. (1987). Mixed orthogonal arrays for variance estimation with unequal numbers of primary selections per stratum, *Biometrika*, **74**, 735-742.
- Hall, P. (1992). *The Bootstrap and Edgeworth Expansion*. Springer-Verlag, New York.
- Hidiroglou, M.A., Rao, J.N.K. and Yung, W. (1999). Variance Computation for Complex Surveys using Estimating Equations, *Proceedings of the Survey Methods Section, Statistical Society of Canada*, pp. 3-9.
- Hidiroglou, M.A., Rao, J.N.K. and Yung, W. (2000). Variance Estimation and Quasi-Score Tests for Complex Surveys using Estimating Equations, *Proceedings of the Survey Research Methods Section, American Statistical Association*, to appear.
- Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables, *Journal of the American Statistical Association*, **58**, 13-30.
- Hu, F. and Kalbfleisch, J.D. (2000). The Estimating Function Bootstrap, *The Canadian Journal of Statistics*, **28**, 449-499.
- Kalton, G. and Kasprzyk, D. (1986). The Treatment of Missing Survey Data, *Survey Methodology*, **12**, 1-16.
- Kish, L. and Frankel, R.M. (1970). Balanced repeated replications for standard errors, *Journal of the American Statistical Association*, **65**, 1071-1094.

- Kovacevic, M.S. and Binder, D.A. (1997). Variance Estimation for Measures of Income Inequality and Polarization - the estimating equations approach, *J. Off. Statist.*, **13**, 41-58.
- Krewski, D.R. (1977). Linearization and Replication Methods in Finite Population Sampling, unpublished Ph.D. thesis, Carleton University, Ottawa, Canada.
- Krewski, D.R. (1978). On the stability of some replication variance estimators in the linear case, *J. Statist. Plan. Inf.*, **2**, 45-51.
- Krewski, D.R. and Rao, J.N.K. (1981). Inference from Stratified Samples: Properties of the Linearization, Jackknife and Balance Repeated Replication Methods, *The Annals of Statistics*, **9**, 1010-1019.
- LeBlanc, M., and Tibshirani, R. (1996). Combining Estimates in Regression and Classification, *Journal of the American Statistical Association*, **91**, 1641-1650.
- Lehmann, E.L. (1999). *Elements of Large-Sample Theory*, Springer-Verlag, New York.
- Lipsitz, S.R., Dear, K.B.G. and Zhao, L. (1994). Jackknife estimators of variance for parameter estimates from estimating equations with applications to clustered survival data, *Biometrics*, **50**, 842-846.
- Lohr, S. (1999). *Sampling: Design and Analysis*. Duxbury Press.
- Lohr, S. and Rao, J.N.K., (2000). Inference From Dual Frame Surveys, *Journal of the American Statistical Association*, **95**, 271-280.

- Lugosi, G., and Nobel, A.B. (1996). Consistency of Data-Driven Histogram Methods for Density Estimation and Classification, *The Annals of Statistics*, **24**, 687-706.
- McCarthy, P.J. (1966). Replication: an approach to the analysis of data from complex surveys, *Vital and Health Statistics*, Series 2, No. 14. Washington, D.C.: U.S. Government Printing Office.
- McCarthy, P.J. (1969). Pseudo-replication: Half-samples, *International Statistical Review*, **37**, 239-264.
- Mojirsheibani, M. (1997). A Consistent Combined Classification Rule, *Statistics and Probability Letters*, **36**, 43-47.
- Mojirsheibani, M. (1999). Combining Classifiers Via Discretization, *Journal of the American Statistical Association*, **94**, 600-609.
- Mojirsheibani, M. (2000). A kernel-based combined classification rule, *Statistics and Probability Letters*, **48**, 411-419.
- Pollard, D. (1984). *Convergence of Stochastic Processes*. Springer, New York.
- Rao, C.R. (1948). Large Sample Tests of Statistical Hypotheses Concerning Several Parameters With Applications to Problems of Estimation, *Proceedings of the Cambridge Philosophical Society*, **44**, 50-57.
- Rao, J.N.K. (1985). Conditional Inference in Survey Sampling, *Survey Methodology*, **11**, 15-31.

- Rao, J.N.K. (1997). Developments in sample survey theory: an appraisal, *The Canadian Journal of Statistics*, **25**, 1-21.
- Rao, J.N.K. and Wu, C.F.J. (1988). Resampling Inference with Complex Survey Data, *Journal of the American Statistical Association*, **83**, 231-241.
- Rao, J.N.K., and Shao, J. (1992). Jackknife variance estimation with survey data under hot deck imputation, *Biometrika*, **79**, 811-822.
- Rao, J.N.K., and Shao, J. (1999). Modified balanced repeated replication for complex survey data, *Biometrika*, **86**, 403-415.
- Rao, J.N.K., Wu, C.F.J. and Yue, K. (1992) Some recent work on resampling methods for complex surveys, *Survey Methodology*, **18**, 209-217.
- Rao, J.N.K., Scott, A.J. and Skinner, C.J. (1998). Quasi-score Test with Survey Data, *Statistica Sinica*, **8**, 1059-1070.
- Rao, J.N.K., Jocelyn, W. and Hidiroglou, M.A. (2001). Confidence Interval Coverage Properties for Regression Estimators in Uni-Phase and Two-Phase Sampling, Tech. Rep., Business Survey Methods Division, Statistics Canada.
- Rao, J.N.K., Yung, W. and Hidiroglou, M.A. (2002). Estimating Equations for the Analysis of Survey Data Using Poststratification Information, *Sankhyā*, in press.
- Rubin-Bleuer, S. (1997). Research Sabbatical Report, Business Survey Methods Division, Statistics Canada.

- Rubin-Bleuer, S. and Schiopu-Kratina, I. (2001). Joint design and model based inference for finite populations, Tech. Rep., Business Survey Methods Division, Statistics Canada.
- Rust, K.F. and Rao, J.N.K. (1996). Variance estimation for complex surveys using replication techniques, *Statistical Methods in Medical Research*, **5**, 283-310.
- Särndal, C.E., Swensson, B. and Wretman, J.H. (1989). The weighted residual technique for estimating the variance of the general regression estimator of the finite population total, *Biometrika*, **76**, 527-537.
- Särndal, C.E., Swensson, B. and Wretman, J.H. (1992). *Model Assisted Survey Sampling*. Springer-Verlag, New York.
- Shao, J. (1999). *Mathematical Statistics*. Springer-Verlag, New York.
- Shao, J. and Tu, D. (1995). *The Jackknife and Bootstrap*. Springer, New York.
- Stone, M. (1974). Cross-Validation Choice and Assessment of Statistical Predictions (with discussion), *Journal of the Royal Statistical Society, Ser. B*, **36**, 111-147.
- Stone, C. (1977). Consistent nonparametric regression, *Annals of Statistics*, **13**, 689-705.
- Stuart, A. and Ord, J.K. (1993). *Kendall's Advanced Theory of Statistics. Vol. 1, Sixth Edition*. New York: John Wiley & Sons Inc.
- Valliant, R. (1993). Poststratification and Conditional Variance Estimation, *Journal of the American Statistical Association*, **88**, 89-96.

- Vapnik, V.N. and Chevonenkis, A. (1974). *Theory of Pattern Recognition*. Moscow: Nauka.
- Wilks, S.S. (1962). *Mathematical Statistics*. New York: John Wiley & Sons Inc.
- Wolpert, D. (1992). Stacked Generalization, *Neural Networks*, **5**, 241-259.
- Wu, C.F.J. (1991). Balanced repeated replications based on mixed orthogonal arrays, *Biometrika*, **78**, 181-188.
- Yung, W. (1996). Contributions to Poststratification in Stratified Multistage Samples, unpublished Ph.D. thesis, Carleton University, Ottawa, Canada.
- Yung, W. and Rao, J.N.K. (1996). Jackknife Linearization Variance Estimators under Stratified Multi-Stage Sampling, *Survey Methodology*, **22**, 23-31.
- Yung, W. and Rao, J.N.K. (2000). Jackknife Variance Estimation Under Imputation for Estimators Using Poststratification Information, *Journal of the American Statistical Association*, **95**, 903-915.
- Yung, W., Hidirolou, M.A. and Rao, J.N.K. (2001). Variance Estimation for Estimating Equations in the Presence of Missing Values, *Proceedings of the Survey Methods Section, Statistical Society of Canada*, to appear.