handbook of statistics 29B

Sample Surveys: Inference and Analysis

Edited by
D. Pfeffermann
C.R. Rao

Handbook of Statistics

VOLUME 29

General Editor

C.R. Rao

ELSEVIER

Amsterdam • Boston • Heidelberg • London • New York • Oxford
Paris • San Diego • San Francisco • Singapore • Sydney • Tokyo

North-Holland is an imprint of Elsevier
Radarweg 29, PO Box 211, 1000 AE Amsterdam, The Netherlands
The Boulevard, Langford Lane, Kidlington, Oxford OX5 1GB, UK

First edition 2009

Copyright © 2009 by Elsevier B.V. All rights reserved.

No part of this publication may be reproduced, stored in a retrieval system
or transmitted in any form or by any means electronic, mechanical, photocopying,
recording or otherwise without the prior written permission of the publisher.

Permissions may be sought directly from Elsevier's Science & Technology Rights
Department in Oxford, UK: phone (+44) (0) 1865 843830; fax (+44) (0) 1865 853333;
email: permissions@elsevier.com. Alternatively you can submit your request online by
visiting the Elsevier web site at http://elsevier.com/locate/permissions, and selecting
Obtaining permission to use Elsevier material.

Notice
No responsibility is assumed by the publisher for any injury and/or damage to persons
or property as a matter of products liability, negligence or otherwise, or from any use
or operation of any methods, products, instructions or ideas contained in the material herein.

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

Library of Congress Cataloging-in-Publication Data
A catalog record for this book is available from the Library of Congress

ISBN: 978-0-444-53124-7
ISSN: 0169-7161

For information on all North-Holland publications
visit our web site at *books.elsevier.com*

Typeset by: diacriTech, India

Printed and bound in Hungary
09 10 11 10 9 8 7 6 5 4 3 2 1

Working together to grow
libraries in developing countries
www.elsevier.com | www.bookaid.org | www.sabre.org
ELSEVIER BOOK AID International Sabre Foundation

Preface to Handbook 29B

Thirty five years ago, the Central Bureau of Statistics in Israel held a big farewell party for the then retiring Prime Minister of Israel, Mrs Golda Meir. In her short thank you speech, the prime minister told the audience: "you are real magicians, you ask 1,000 people what they think, and you know what the whole country thinks". Magicians or not, this is what sample surveys are all about: to learn about the population from a (often small) sample, dealing with issues such as how to select the sample, how to process and analyse the data, how to compute the estimates, and face it, since we are not magicians, also how to assess the margin of error of the estimates.

Survey sampling is one of the most practiced areas of statistics, and the present handbook contains by far the most comprehensive, self-contained account of the state of the art in this area. With its 41 chapters, written by leading theoretical and applied experts in the field, this handbook covers almost every aspect of sample survey theory and practice. It will be very valuable to government statistical organizations, to social scientists conducting opinion polls, to business consultants ascertaining customers' needs and as a reference text for advanced courses in sample survey methodology. The handbook can be used by a student with a solid background in general statistics who is interested in learning what sample surveys are all about and the diverse problems that they deal with. Likewise, the handbook can be used by a theoretical or applied researcher who is interested in learning about recent research carried out in this broad area and about open problems that need to be addressed. Indeed, in recent years more and more prominent researchers in other areas of statistics are getting involved in sample survey research in topics such as small area estimation, census methodology, incomplete data and resampling methods.

The handbook consists of 41 chapters with a good balance between theory and practice and many illustrations of real applications. The chapters are grouped into two volumes. Volume 29A entitled "Design, Methods and Applications" contains 22 chapters. Volume 29B entitled "Inference and Analysis" contains the remaining 19 chapters. The chapters in each volume are further divided into three parts, with each part preceded by a short introduction summarizing the motivation and main developments in the topics covered in that part.

Volume 29A deals with sampling methods and data processing and considers in great depth a large number of broad real life applications. Part 1 is devoted to sampling and survey design. It starts with a general introduction of alternative approaches to survey sampling. It then discusses methods of sample selection and estimation, with separate chapters on unequal probability sampling, two-phase and multiple frame sampling,

surveys across time, sampling of rare populations and random digit dialling surveys. Part 2 of this volume considers data processing, with chapters on record linkage and statistical editing methods, the treatment of outliers and classification errors, weighting and imputation to compensate for nonresponse, and methods for statistical disclosure control, a growing concern in the modern era of privacy conscious societies. This part also has a separate chapter on computer software for sample surveys. The third part of Volume 29A considers the application of sample surveys in seven different broad areas. These include household surveys, business surveys, agricultural surveys, environmental surveys, market research and the always intriguing application of election polls. Also considered in this part is the increasing use of sample surveys for evaluating, supplementing and improving censuses.

The present volume 29B is concerned with inference and analysis, distinguishing between methods based on probability sampling principles ("design-based" methods), and methods based on statistical models ("model-based" methods). Part 4 (the first part of this volume) discusses alternative approaches to inference from survey data, with chapters on model-based prediction of finite population totals, design-based and model-based inference on population model parameters and the use of estimating functions and calibration for estimation of population parameters. Other approaches considered in this part include the use of nonparametric and semi-parametric models, the use of Bayesian methods, resampling methods for variance estimation, and the use of empirical likelihood and pseudo empirical likelihood methods. While the chapters in Part 4 deal with general approaches, Part 5 considers specific estimation and inference problems. These include design-based and model-based methods for small area estimation, design and inference over time and the analysis of longitudinal studies, categorical data analysis and inference on distribution functions. The last chapter in this part discusses and illustrates the use of scatterplots with survey data. Part 6 in Volume 29B is devoted to inference under informative sampling and to theoretical aspects of sample survey inference. The first chapter considers case-control studies which are in common use for health and policy evaluation research, while the second chapter reviews several plausible approaches for fitting models to complex survey data under informative sampling designs. The other two chapters consider asymptotics in finite population sampling and decision-theoretic aspects of sampling, bringing sample survey inference closer to general statistical theory.

This extensive handbook is the joint effort of 68 authors from many countries, and we would like to thank each one of them for their enormous investment and dedication to this extensive project. We would also like to thank the editorial staff at the North-Holland Publishing Company and in particular, Mr. Karthikeyan Murthy, for their great patience and cooperation in the production of this handbook.

Danny Pfeffermann
C. R. Rao

Contributors: Vol. 29B

Binder, David A., *Methodology Branch, Statistics Canada, 100 Tunney's Pasture Driveway, Ottawa ON K1A 0T6; e-mail: dbinder49@hotmail.com* (Ch. 24).

Breidt, F. Jay, *Department of Statistics, Colorado State University, Fort Collins, CO 80523-1877; e-mail: jbreidt@stat.colostate.edu* (Ch. 27).

Datta, Gauri S., *Department of Statistics, University of Georgia, Athens GA 30602-7952, USA; e-mail: gaurisdatta@gmail.com* (Ch. 32).

Dorfman, Alan H., *Office of Survey Methods Research, U.S. Bureau of Labor Statistics, Washington, D.C., U.S.A., 20212; e-mail: dorfman.alan@bls.gov* (Ch. 36).

Gershunskaya, Julie, *U.S. Bureau of Labor Statistics, 2 Massachusetts Avenue, NE, Washington, DC 20212, USA; e-mail: gershunskaya.julie@bls.gov* (Ch. 28).

Ghosh, Malay, *Dept. of Statistics, University of Florida, Gainesville, Florida, 32611-8545, USA; e-mail: ghoshm@stat.ufl.edu* (Ch. 29).

Godambe, V. P., *Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Ontario, Canada N2L 3G1; e-mail: vpgodamb@uwaterloo.ca* (Ch. 26).

Graubard, Barry I., *Biostatistics Branch, National Cancer Institute, Executive Plaza South Bldg, 6120 Executive Blvd, Room 8024, Bethesda, MD, 20892, USA; e-mail: graubarb@mail.nih.gov* (Ch. 37).

Jiang, Jiming, *Department of Statistics, University of California, Davis, CA 95616, USA; e-mail: jiang@wald.ucdavis.edu* (Ch. 28).

Korn, Edward L., *Biometric Research Branch, National Cancer Institute, Executive Plaza North Bldg, 6130 Executive Blvd, Room 8128, Bethesda, MD, 20892, USA; e-mail: korne@mail.nih.gov* (Ch. 37).

Kott, Phillip S., *RTI International, 6110 Executive Blvd., Suite 902, Rockville, MD 20852; e-mail: pkott@rti.org* (Ch. 25).

Lahiri, Partha, *Joint Program in Survey Methodology, 1218 Lefrak Hall, University of Maryland, College Park, MD 20742, USA; e-mail: plahiri@survey.umd.edu* (Ch. 28).

Lehtonen, Risto, *Department of Mathematics and Statistics, University of Helsinki, P.O. Box 68 (Gustaf Hällströmin katu 2b), FI-00014 University of Helsinki, Finland; e-mail: risto.lehtonen@helsinki.fi* (Ch. 31).

McLaren, Craig, *Head, Retail Sales Branch, Office for National Statistics, United Kingdom; e-mail: chmclaren@hotmail.com* (Ch. 33).

Nathan, Gad, *Department of Statistics, Hebrew University, Mt Scopus, 91905 Jerusalem, Israel; e-mail: gad@huji.ac.il* (Ch. 34, Introduction to Part 5).

Opsomer, Jean, *Department of Statistics, Colorado State University, Fort Collins, CO 80523-1877; e-mail: jopsomer@stat.colostate.edu* (Introduction to Part 4; Ch. 27).

Pfeffermann, Danny, *Department of Statistics, Hebrew University of Jerusalem, Jerusalem 91905, Israel; and Southampton Statistical Sciences Research Institute, University of Southampton, Southampton, SO17 1BJ, United Kingdom; e-mail: msdanny@huji.ac.il* (Ch. 39, Introduction to Part 5, 6).

Prášková, Zuzana, *Department of Probability and Mathematical Statistics, Faculty of Mathematics and Physics, Charles University in Prague, Sokolovská 83, 186 75 Prague, Czech Republic; e-mail: praskova@karlin.mff.cuni.cz* (Ch. 40).

Rao, J.N.K., *School of Mathematics and Statistics, Carleton University, Colonel by Drive Ottawa, Ontario K1S 5B6, Canada; e-mail: jrao34@rogers.com* (Ch. 30).

Rinott, Yosef, *Department of Statistics, The Hebrew University, Jerusalem 91905, Israel; e-mail: rinott@mscc.huji.ac.il* (Ch. 41).

Roberts, Georgia, *Methodology Branch, Statistics Canada, 100 Tunney's Pasture Driveway, Ottawa ON K1A 0T6; e-mail: Georgia.Roberts@statcan.gc.ca* (Ch. 24).

Scott, Alastair, *Department of Statistics, University of Auckland, 38 Princes Street, Auckland, New Zealand 1010; e-mail: a.scott@auckland.ac.nz* (Ch. 38).

Sen, Pranab Kumar, *Department of Biostatistics, University of North Carolina at Chapel Hill, NC 27599-7420, USA; e-mail: pksen.bios.unc.edu* (Ch. 40).

Steel, David., *Director, Centre for Statistical and Survey Methodology, University of Wollongong, Australia; e-mail: dsteel@uow.edu.au* (Ch. 33).

Sverchkov, Michail, *U. S. Bureau of Labor Statistics and BAE Systems IT, 2 Massachusetts Avenue NE, Suite 1950, Washington, DC, 20212; e-mail: Sverchkov. Michael@bls.gov* (Ch. 39).

Thompson, M. E., *Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Ontario, Canada N2L 3G1; e-mail: methomps@uwaterloo.ca* (Ch. 26).

Valliant, Richard, *Research Professor, Joint Program in Survey Methodology, University of Maryland and Institute for Social Research, University of Michigan, 1218 Lefrak Hall, College Park MD 20742; e-mail:rvalliant@survey.umd.edu* (Ch. 23).

Veijanen, Ari, *Statistics Finland, Työpajankatu 13, Helsinki, FI-00022 Tilastokeskus, Finland; e-mail: ari.veijanen@stat.fi* (Ch. 31).

Wild, Chris, *Department of Statistics, University of Auckland, 38 Princes Street, Auckland, New Zealand 1010; e-mail: c.wild@auckland.ac.nz* (Ch. 38).

Wu, Changbao, *Department of Statistics and Actuarial Science University of Waterloo 200 University Avenue West Waterloo, Ontario N2L 3G1 Canada. e-mail: cbwu@uwaterloo.ca* (Ch. 30).

Contributors: Vol. 29A

Beaumont, Jean-François, *Statistical Research and Innovation Division, Statistics Canada, 100 Tunney's Pasture Driveway, R.H. Coats building, 16th floor, Ottawa (Ontario), Canada K1A 0T6; e-mail: Jean-Francois.Beaumont@statcan.gc.ca* (Ch. 11).

Berger, Yves G., *Southampton Statistical Sciences Research Institute, University of Southampton, Southampton, SO17 1BJ, United Kingdom; e-mail: Y.G.Berger@soton.ac.uk* (Ch. 2).

Bethlehem, Jelke, *Statistics Netherlands, Methodology Department, The Hague, The Netherlands; e-mail: jbtm@cbs.nl* (Ch. 13).

Biemer, Paul P., *RTI International, P.O. Box 12194, Research Triangle Park, NC 27709-2194; and University of North Carolina, Odum Institute for Research in Social Science, Chapel Hill, NC; e-mail: ppb@rti.org* (Ch. 12, Introduction to Part 2).

Brewer, Kenneth, *School of Finance and Applied Statistics, College of Business and Economics, L.F. Crisp Building (Building 26), Australian National University, A.C.T. 0200, Australia; e-mail: ken.brewer@anu.edu.au* (Ch. 1).

Brick, J. Michael, *Westat and Joint Program in Survey Methodology, University of Maryland, 1650 Research Blvd, Rockville, MD, 20850; e-mail: mikebrick@westat.com* (Ch. 8).

Chowdhury, Sadeq, *NORC, University of Chicago, 4350 East-West Highway, Suite 800, Bethesda, MD 20814; e-mail: sadeqc@yahoo.com* (Ch. 7).

Christman, Mary C., *University of Florida, Department of Statistics, Institute of Food and Agricultural Science, Gainesville, Florida; e-mail: mcxman@ufl.edu* (Ch. 6).

De Waal, Ton, *Department of Methodology, Statistics Netherlands, PO Box 24500, 2490 HA The Hague, The Netherlands; e-mail: t.dewaal@cbs.nl* (Ch. 9).

Frankovic, Kathleen A., *Survey and Election Consultant, 3162 Kaiwiki Rd., Hilo, HI 96720; e-mail: kaf@cbsnews.com* (Ch. 22).

Fuller, Wayne A., *Center for Survey Statistics and Methodology, Department of Statistics, Iowa State University, Ames, IA 50011; e-mail: waf@iastate.edu* (Ch. 3).

Gambino, Jack G., *Household Survey Methods Division, Statistics Canada, Ottawa, Canada K1A 0T6; e-mail: jack.gambino@statcan.gc.ca* (Ch. 16, Introduction to Part 3).

Glickman, Hagit, *National Authority of Measurement and Evaluation in Education (RAMA), Ministry of Education, Kiryat Hamemshala, Tel Aviv 67012, Israel; e-mail: hglickman.rama@education.gov.il* (Ch. 21).

Gregoire, Timothy, Weyerhaeuser, J.P. Jr., *Professor of Forest Management, School of Forestry and Environmental Studies, Yale University, 360 Prospect Street, New Haven, CT 06511-2189; e-mail: timothy.gregoire@yale.edu* (Ch. 1).

Haziza, David, *Département de Mathématiques et de Statistique, Université de Montréal, Pavillon André-Aisenstadt, 2920, chemin de la Tour, bureau 5190, Montréal, Québec H3T 1J4, Canada; e-mail: David.haziza@umontreal.ca* (Ch. 10).

Hidiroglou, Michael A., *Statistical Research and Innovation Division, Statistics Canada, Canada, K1A 0T6; e-mail: Mike.Hidiroglou@statcan.gc.ca* (Ch. 17).

House, Carol C., *National Agricultural Statistics Service, U.S. Department of Agriculture, Washington, DC, USA; e-mail: Carol.House@usda.gov* (Ch. 18).

Kalton, Graham, *Westat, 1600 Research Blvd., Rockville, MD 20850; e-mail: grahamkalton@westat.com* (Ch. 5).

Kelly, Jenny, *NORC, University of Chicago, 1 North State Street, Suite 1600, Chicago, IL 60602; e-mail: Kelly-Jenny@norc.org* (Ch. 7).

Lavallée, Pierre, *Social Survey Methods Division, Statistics Canada, Canada, K1A 0T6; e-mail: pierre.lavallee@statcan.gc.ca* (Ch. 17).

Legg, Jason C., *Division of Global Biostatistics and Epidemiology, Amgen Inc., 1 Amgen Center Dr. Newbury Park, CA 91360; e-mail: jlegg@amgen.com* (Ch. 3).

Lohr, Sharon L., *Department of Mathematics and Statistics, Arizona State University, Tempe, AZ 85287-1804, USA; e-mail: sharon.lohr@asu.edu* (Ch. 4, Introduction to Part 1).

Marker, David A., *Westat, 1650 Research Blvd., Rockville Maryland 20850; e-mail: DavidMarker@Westat.com* (Ch. 19).

Montaquila, Jill M., *Westat and Joint Program in Survey Methodology, University of Maryland, 1650 Research Blvd, Rockville, MD, 20850; e-mail: jillmontaquila@westat.com* (Ch. 8).

Naidu, Gurramkonda M., *Professor Emeritus, College of Business & Economics, University of Wisconsin-Whitewater, Whitewater, WI 53190; e-mail: naidug@uww.edu* (Ch. 20).

Nirel, Ronit, *Department of Statistics, The Hebrew University of Jerusalem, Mount Scopus, Jerusalem 91905, Israel; e-mail: nirelr@cc.huji.ac.il* (Ch. 21).

Nusser, S. M., *Department of Statistics, Iowa State University, Ames, IA, USA; e-mail: nusser@iastate.edu* (Ch. 18).

Panagopoulos, Costas, *Department of Political Science, Fordham University, 441 E. Fordham Rd., Bronx, NY 10458; e-mail: costas@post.harvard.edu* (Ch. 22).

Rivest, Louis-Paul, *Departement de mathématiques et de statistique, Université Laval, Cité universitaire, Québec (Québec), Canada G1K 7P4; e-mail: lpr@mat.ulaval.ca* (Ch. 11).

Shapiro, Robert Y., *Department of Political Science and Institute for Social and Economic Research and Policy, Columbia University, 420 West 118th Street, New York, NY 10027; e-mail: rys3@columbia.edu* (Ch. 22).

Silva, Pedro Luis do Nascimento, *Southampton Statistical Sciences Research Institute, University of Southampton, UK; e-mail: pedrolns@soton.ac.uk* (Ch. 16).

Skinner, Chris, *Southampton Statistical Sciences Research Institute, University of Southampton, Southampton SO17 1BJ, United Kingdom; e-mail: C.J.Skinner@soton.ac.uk* (Ch. 15).

Stevens, Don L. Jr., *Statistics Department, Oregon State University, 44 Kidder Hall, Corvallis, Oregon, 97331; e-mail: stevens@stat.oregonstate.edu* (Ch. 19).

Tillé, Yves, *Institute of Statistics, University of Neuchâtel, Pierre à Mazel 7, 2000 Neuchâtel, Switzerland; e-mail: yves.tille@unine.ch* (Ch. 2).

Velu, Raja, *Irwin and Marjorie Guttag Professor, Department of Finance, Martin J. Whitman School of Management, Syracuse University, Syracuse, NY 13244-2450; e-mail:rpvelu@syr.edu* (Ch. 20).

Winkler, William E., *Statistical Research Division, U.S. Census Bureau, 4600 Silver Hill Road, Suitland, MD 20746; e-mail: william.e.winkler@census.gov* (Ch. 14).

Wolter, Kirk, *NORC at the University of Chicago, and Department of Statistics, University of Chicago, 55 East Monroe Street, Suite 3000, Chicago, IL 60603; e-mail: wolter-kirk@norc.uchicago.edu* (Ch. 7).

Introduction to Part 4

Jean D. Opsomer

1. Introduction

As the chapters in Part 3 of this volume clearly illustrate, surveys are an important source of primary data in a large array of disciplines, ranging from natural resources and ecology to social and health sciences. In all those areas, the original focus on the estimation of well-defined finite-population quantities has been supplemented more recently by the use of survey data for fitting statistical models.

When targeting finite-population quantities, design-based inference remains the dominant approach in use today. But even within this context, modeling plays an important role in estimation and inference. Some of the main uses of modeling are to improve the precision of survey estimators at the population level or for domains, to calibrate estimates so that they "match" control quantities such as census numbers or estimates from other surveys, and to adjust estimators for nonresponse or measurement error. Purely model-based inference for finite-population quantities is also possible, as will be discussed below.

In inference on statistical models, the finite population is often not of primary interest. In fact, the design randomization itself is often viewed as a nuisance by data analysts, because their focus is on estimating and evaluating the model and its associated sources of uncertainty. Nevertheless, there is a clear recognition that model fitting with survey data needs to take account of the sampling design features. Inference on models is usually undertaken using one of the two modes: a pure model-based approach or a "hybrid" approach blending design-based and model-based methods.

Sitting at the intersection between model-based and design-based inference, the analysis of survey data can be a complicated affair, even though it certainly did not start off that way. Indeed, the original principles underlying randomization-based sample selection and corresponding inverse-probability weighted estimation are among the simplest concepts in statistics. The complexities arise because survey data analyses need to account not only for this randomization uncertainty but also for the effect of modeling. The next section provides a brief look at the different modes of inference with survey data. Section 3 gives an overview of the chapters in Part 4, which address a range of both traditional and modern methods for estimation of finite-population quantities, model quantities, or both.

2. Modes of inference with survey data

Consider a finite population $U = \{1, \ldots, i, \ldots, N\}$. Associated with each element i is a (possibly vector-valued) variable of interest y_i. For now, we will focus on the population mean $\overline{y}_N = \sum_U y_i/N$ as a finite-population quantity of interest. A probability sample $s \subset U$ is drawn according to a sampling design $p(\cdot)$, where $p(s) = \Pr[\text{sample } s \text{ is selected}]$. Let $\pi_i = \Pr[i \in s] = \sum_{s:i\in s} p(s) > 0$. When the sampling design is the only source of randomness explicitly accounted for in estimation and inference, the mode of inference is usually referred to as design-based or, somewhat less frequently, randomization-based.

Pure design-based estimators of $\overline{y}_N$ are given by the Horvitz–Thompson estimator $\hat{y}_{\text{HT}} = N^{-1}\sum_s y_i/\pi_i$ and the Hájek estimator $\hat{y}_{\text{HA}} = \hat{N}^{-1}\sum_s y_i/\pi_i$, where $\hat{N} = \sum_s 1/\pi_i$. Both estimators have a long history of applications in survey estimation because of their simplicity and desirable statistical properties. They are design unbiased ($\hat{y}_{\text{HT}}$), or approximately design unbiased ($\hat{y}_{\text{HA}}$), without reliance on a model for the y_i, and inference can be performed using the fact that for sufficiently large samples, the distribution of the estimators is well approximated by the Gaussian distribution. See Chapter 40 of this handbook for asymptotics in finite-population sampling. For most sampling designs, explicit formulas are available for constructing estimators for the variance of $\hat{y}_{\text{HT}}$. These, together with the Gaussian distribution, can then be used for constructing confidence intervals for $\overline{y}_N$ (the case of $\hat{y}_{\text{HA}}$ will be treated below).

For at least as long as design-based estimation have been used, the pure design-based estimators $\hat{y}_{\text{HT}}$ and $\hat{y}_{\text{HA}}$ have been supplemented by alternative design-based estimators that take advantage of additional information about the population, such as the ratio estimator and the poststratified estimator. Fuller (2002) gives an overview of the development of the more general class of regression estimators, which have traditionally relied on linear models. Let $\boldsymbol{x}_i$ represent a vector of auxiliary variables for element i and $\overline{\boldsymbol{x}}_N$ its known population mean. A general form of the regression estimator is given by

$$\hat{y}_{\text{reg}} = \hat{y}_{\text{HT}} + (\overline{\boldsymbol{x}}_N - \hat{\boldsymbol{x}}_{\text{HT}})\hat{\boldsymbol{B}}, \tag{1}$$

where $\hat{\boldsymbol{B}}$ is a sample-based quantity whose exact form varies, depending on the specific version of regression estimation being considered. For $\hat{y}_{\text{reg}}$ to be a design-consistent estimator for $\overline{y}_N$, the only requirement on $\hat{\boldsymbol{B}}$ is that it converges in probability to a constant, which allows a great deal of flexibility in the formulation of the regression estimator. Note also that the Horvitz–Thompson estimators in (1) can be replaced by the Hájek versions.

Depending on the choice of the auxiliary variables $\boldsymbol{x}_i$ and the specific choice of $\hat{\boldsymbol{B}}$, the regression estimator can be significantly more efficient than the Horvitz–Thompson and Hájek estimators, while maintaining design consistency. Most methods define $\hat{\boldsymbol{B}}$ as a vector of estimated regression coefficients for a linear regression model relating y_i and $\boldsymbol{x}_i$. A simple version of such a model can be written as

$$y_i = \boldsymbol{x}_i\boldsymbol{\beta} + \varepsilon_i, \tag{2}$$

with the ε_i representing independent and identically distributed (iid) zero-mean random variables. When working in the design-based paradigm, such a model is sometimes called a *working model*, because the regression estimator in (1) remains design consistent

even if the model is not a correct representation of the relationship between y_i and $\boldsymbol{x}_i$. However, the smaller the deviations between the y_i and the working model mean $\boldsymbol{x}_i\boldsymbol{\beta}$, the more efficient is the regression estimator.

The linear working model in (2) can be replaced by more complicated regression models, such as nonparametric or nonlinear models, which can be more appropriate depending on the nature of the data. In addition, using a working model is not the only approach to improve the efficiency of estimators of finite-population quantities with respect to the randomization distribution. Calibration estimators (Deville and Särndal, 1992) form a broad class of estimators that include regression estimators based on linear models and the raking ratio estimator as special cases. Calibration estimators do not use models in their construction. They have the same form as the Horvitz–Thompson estimator but with new weights w_i such that the "distance" between the calibrated estimator, written as $\hat{y}_{\text{cal}} = N^{-1}\sum_s w_i y_i$, and the purely design-weighted estimator $\hat{y}_{\text{HT}} = N^{-1}\sum_s \pi_i^{-1} y_i$ is minimized, subject to a set of calibration constraints. The "distance" referred to here is a sample-based function of the weights of the form $\sum_s q(w_i, \pi_i^{-1})$ for some suitable function $q(\cdot,\cdot)$, and the calibration constraints are given as $\sum_s w_i\boldsymbol{x}_i = \sum_U \boldsymbol{x}_i$. Several chapters in Part 4 explore the working model and calibration approaches for improving the efficiency of survey estimators.

The previous discussion focused on the design-based mode of estimation for simple finite-population quantities that can be expressed as population means. For more complicated population quantities that can be written as functions of population means, design-based estimation is ordinarily based on estimating the individual means using the methods just described and plugging the estimators into the same function. For example, a population ratio of two variables, $R_N = \overline{y}_N/\overline{x}_N$ can be estimated by $\hat{R} = \hat{y}_{\text{HT}}/\hat{x}_{\text{HT}}$. The estimator $\hat{y}_{\text{HA}}$ is actually the ratio of two Horvitz–Thompson estimators and follows this principle. The same principle applies to more complicated finite-population quantities including population regression coefficients or those implicitly defined by population estimating equations. This is further explored in some of the chapters in Part 4.

Just like for $\hat{y}_{\text{HT}}$, the Gaussian distribution is used as an approximate distribution for all the estimators mentioned so far. However, with the exception of $\hat{y}_{\text{HT}}$, none of the estimators mentioned above has an explicit variance expression for general designs. Inferential statements such as confidence intervals are therefore based on "delta method" arguments, in which an approximate variance is used instead of the exact variance. A number of methods are available for estimating the approximate variance, including Taylor linearization followed by estimation of the unknown population quantities, and several replication methods, including the jackknife and bootstrap methods. These methods are described in further chapters in Part 4.

It is also possible to consider a model-based mode of inference. In this case, we start from a model such as (2), which is assumed to be a correct representation of the data. Finite-population quantities such as $\overline{y}_N$ are now random with respect to the model, and "estimation" of $\overline{y}_N$ involves estimation of model parameters such as $\boldsymbol{\beta}$ in (2), followed by model-based prediction. We will, therefore, return to this mode after discussing inference for model parameters.

When model parameters are the target of inference with survey data, the data analyst starts from a statistical model such as (2), and based on that model, a suitable estimator is applied to the observed sample data to estimate the parameters. A crucial consideration

in this context is whether the model is correct for the sample data. While the issue of model correctness exists of course in any statistical data analysis, it is more pressing for survey data because of the effect of the sampling design.

To see this, consider again model (2) and assume that it holds in the finite population, in the sense that the pairs $(\boldsymbol{x}_i, y_i), i \in U$, follow the linear mean model and have the assumed iid error distribution. Under this model and for a simple random sampling design, the ordinary least squares (OLS) estimator is suitable for estimating $\boldsymbol{\beta}$, in the sense of having the smallest variance among the linear unbiased estimators. Now suppose that the sample s is drawn by a sampling design that tends to select pairs in which, for a given value of $\boldsymbol{x}_i$, the value of y_i is larger than expected under the iid model error distribution. In this case, the OLS estimator will be biased and any inference based on this estimator will be incorrect, because the assumed model is not correct for the sample data. If the model for the data in the sample is the same as that in the population, the sampling design is often called *ignorable*, and conversely, if the sample model deviates from the population model, the design is called nonignorable or informative. A complete treatment of inference under informative sampling is given in Chapter 39.

Note that ignorability of the sampling design does not require the full distributional specification of the data be unaffected by the design. For instance, a sampling design that tends to select pairs $(\boldsymbol{x}_i, y_i)$ with larger values for some or all the elements in $\boldsymbol{x}_i$ is ignorable for the estimation of $\boldsymbol{\beta}$ in the model (2), as long as it does not affect the linear relationship between y_i and $\boldsymbol{x}_i$, and hence the validity of the OLS estimator for $\boldsymbol{\beta}$. This is despite the fact that under this design, the joint distribution of the $(\boldsymbol{x}_i, y_i)$ in the sample is clearly different from their distribution in the population.

Several different approaches are available for performing statistically valid inference for model parameters. A first possible approach, which is rarely fully satisfactory in practice, is to assume that the statistical model holds for the sample itself and to base the inference on that model. The issue of ignorability is avoided because the population from which the sample originated is not considered. While statistical analysis in this manner is readily performed, a major concern is how to generalize the results beyond the particular sample at hand, because the representativeness of the sample data with respect to a broader context, whether a finite population or a hypothetical data generating mechanism, is not established.

A second possible approach is to specify the model in such a way that it holds for both the sample and the population. This approach is typically preferable over the previous one, because it ensures both the statistical validity of the estimator and the generalizability of the inference. This can be done in a number of different ways, the simplest of which is to check whether ignorability of the design holds and if not, adapt the model so that it is valid for both the sample and the population. For instance, in the linear regression model (2), ignorability of the design can often be achieved by including the so-called *design variables* such as stratification and/or cluster indicators in the covariate vector $\boldsymbol{x}$. We refer to chapters in Part 4 and Chapter 39 for a discussion of testing for ignorability and model adaptation strategies. While simple in concept, this "model extension" approach to achieve ignorability has drawbacks as well, further described in Chapter 39 but briefly summarized here. One key issue is that it requires access to all the design variables, which might not be available to secondary data analysts because of data confidentiality concerns. While it might be possible to use the sampling weights instead of the design variables, the weights might be not be sufficient

to achieve ignorability. If the design variables are available, another problem is that the extended model will now contain variables that are not part of the original model, complicating the interpretation of the fitted model. Finally, this model extension with design variables will not be applicable in all situations. A typical example of this is when the inclusion in the sample depends on the outcome variables (the y_i in the regression context).

A third approach is to acknowledge the fact that two different models will hold at the sample and the population level, but connect them with each other by using an additional model for the inclusion probabilities. Based on the relationships between these three models, it is possible to obtain sample-based parameter estimates that are valid for the population model. For a complete discussion of this relatively recent approach to inference for model parameters with survey data, we refer to Chapter 39.

A final approach for inference for model parameters combines design-based and model-based inference and is the approach most commonly used by survey statisticians. Under this approach, the finite population is viewed as a realization from a statistical model, often referred to as the *superpopulation model* in this case. A sample is drawn from this finite population according to the sampling design. The first step in constructing sample-based estimators for the superpopulation model parameters is to define population-level "estimates," which are the appropriate model-based estimates for the model parameters if the full population had been observed. Pfeffermann (1993) refers to those population-level estimates as "descriptive population quantities" (DPQ). Typically, DPQs can be written as functions of finite-population sums. Then, sample-based estimators of the DPQs are constructed by applying the design-based methods described earlier in this section, that is, by replacing the population sums by their respective sample-based estimators. For instance, for the linear model (2), the DPQ is the population-level OLS estimator of $\boldsymbol{\beta}$, which can be written as

$$\boldsymbol{B} = \left(\sum_U \boldsymbol{x}_i^T \boldsymbol{x}_i \right)^{-1} \sum_U \boldsymbol{x}_i^T y_i.$$

A sample-based estimator of the DPQ $\boldsymbol{B}$, based on Horvitz–Thompson estimation, is defined as

$$\hat{\boldsymbol{B}}_{\mathrm{HT}} = \left(\sum_s \frac{\boldsymbol{x}_i^T \boldsymbol{x}_i}{\pi_i} \right)^{-1} \sum_s \frac{\boldsymbol{x}_i^T y_i}{\pi_i},$$

which is then used as an estimator of $\boldsymbol{\beta}$ under the combined design-based and model-based approach.

Inference under this approach requires to explicitly account for two sources of randomness: the model-based randomness, accounting for the difference between the DPQ and the superpopulation model parameters, and the design-based randomness, accounting for the difference between the sample estimator and the DPQ. An important advantage of this hybrid approach to inference is that the question of ignorability does not occur, because no model specifications are required at the sample level. Disadvantages are that the combined inference mode is more cumbersome to apply in practice and that the resulting estimators are often less efficient than pure model-based estimators. It is also restricted to point estimators of model parameters and cannot be used for prediction,

say to predict y for an unobserved unit with covariates $\boldsymbol{x}$ or to predict a small area mean with no sample from that area. Several chapters in Part 4 further discuss issues involved in this approach.

We now briefly return to the estimation of finite-population quantities, where these quantities are predicted based on a model fitted to the sample data. As it is clear from the above discussion, statistical validity of this approach requires either a model that is valid for both the sample and the population (in order words, a model for which the design is ignorable) or a way to connect the sample model with the population model. A model specification with ignorable design is most often assumed by statisticians pursuing model-based estimation of finite-population quantities. Once such a model is found, it is possible to apply appropriate model-based prediction methods and perform inference based on the model. As a simple example, suppose once again that we want to estimate the finite population mean $\overline{y}_N$ and that the data in the sample and the population follow the linear model (2) with iid errors. An appropriate "estimator" for $\overline{y}_N$ under this model is the best linear unbiased predictor

$$\hat{y}_{\text{blup}} = \frac{1}{N}\left(\sum_s y_i + \sum_{s^C} \boldsymbol{x}_i \hat{\boldsymbol{B}}\right),$$

where s^C represents the complement of the sample s in the population U, and $\hat{\boldsymbol{B}}$ is the OLS estimator computed from the sample data $(\boldsymbol{x}_i, y_i)$, $i \in s$. Prediction under informative sampling was considered in Sverchkov and Pfeffermann (2004). Other models and predictors are also possible, as further discussed in several chapters in Part 4.

3. Overview of Part 4

Chapter 25 by Kott revisits regression estimation and calibration for finite-population quantities. Kott discusses a number of refinements and extensions of regression estimators, including the use of instrumental variables, nonlinear calibration, and the issue of optimal regression estimation. Kott's discussion focuses on design-based estimators and their properties with respect to their randomization distribution over repeated sampling, but he also considers model-based properties.

Chapter 23 by Valliant and Chapter 29 by Ghosh discuss model-based prediction of finite-population quantities, the former using a frequentist prediction approach and the latter using Bayesian estimators. Echoing the previous discussion of design ignorability in model-based inference, both chapters explicitly acknowledge the importance of ensuring that any sample-fitted model is also valid at the population level. Once this is accomplished, the full set of model-based techniques familiar to statisticians, including best linear unbiased prediction for linear models (as in Chapter 23) or hierarchical Bayesian inference (Chapter 29), become available to the data analyst.

Chapter 24 by Binder and Roberts covers inference for parameters under general models. They expand on the discussion of ignorability in Section 2 and then compare the model-based and "hybrid" modes of estimation and inference for model parameters. They describe an extended version of the superpopulation-population-sample inferential framework of the previous section and evaluate the properties of estimators in this framework.

Chapter 26 by Godambe and Thompson and Chapter 27 by Breidt and Opsomer review two classes of estimators that are widely used in general statistics but are still relatively uncommon in survey statistics. Godambe and Thompson discuss estimators that are defined as the solutions to inverse-probability-weighted estimating equations. This class of estimators includes common survey estimators such as ratio and distribution function estimators but also covers less familiar ones such as quantile estimators. They consider these estimators both when targeting finite-population quantities (themselves defined as solutions to population-level estimating equations) and when targeting superpopulation model parameters. Breidt and Opsomer cover applications of nonparametric methods in survey estimation. The purpose of nonparametric regression and density estimation methods is to describe features of a model without having to specify the full parametric form for it. Breidt and Opsomer describe several nonparametric regression estimators of finite-population quantities. They also discuss nonparametric estimation of superpopulation models using survey data.

Finally, Chapter 28 by Gershunskaya, Jiang, and Lahiri and Chapter 30 by Rao and Wu focus on uncertainty measures for survey estimators of finite-population quantities. Gershunskaya et al. review replication methods in a number of survey estimation contexts. These methods include jackknife and bootstrap replication and several extensions. These methods are particularly useful for variance estimation in surveys, because the analytic variance formulas for complex sampling designs are often cumbersome or difficult to derive in case of complex nonlinear estimators or because they require information that cannot be provided to survey data users due of confidentiality concerns. Confidence intervals for finite-population quantities are obtained by combining the approximate normality of the point estimators with the replication-based variance estimators. In contrast, Rao and Wu discuss a relatively novel way to perform inference from survey data by extending the concept of *empirical likelihood* to the survey context. An empirical likelihood assigns point mass p_i to each observed value y_i in a data set, and a maximum empirical likelihood estimator is obtained by searching for the values p_i that maximize the (log) likelihood, subject to constraints such as $p_i > 0$, $\sum p_i = 1$ and calibration constraints. An important application of empirical likelihood ideas is in the construction of confidence intervals based on probability level sets for likelihood ratio statistics. In the design-based context, the empirical likelihood is replaced by a pseudo empirical likelihood in which each observation is weighted by the inverse of its inclusion probability. The resulting method is applied to point estimation and to the construction of confidence intervals that do not rely on asymptotic normality of the estimators.

Sample Surveys: Inference and Analysis, Vol. 29B
ISSN: 0169-7161
© 2009 Elsevier B.V. All rights reserved
DOI: 10.1016/S0169-7161(09)00223-5

Model-Based Prediction of Finite Population Totals

Richard Valliant

1. Superpopulation models and some simple examples

A finite population is a collection of distinct units such as people, business establishments, schools, hospitals, or transactions over some period of time. A basic descriptive statistic for these collections is the total of some variable. Depending on the population, the total may be the number of persons who are employed, the total of expenditures on capital equipment, total salary costs for teachers, or total number discharges from hospitals during some time period. Another common descriptive statistic is the mean per unit, which is often estimated as a total divided by an estimate of the number of units that contribute to the total.

In many populations, particularly ones that have been previously sampled or surveyed, a frame of units is available along with some auxiliary data on each unit. In other cases, a full frame of all units is not on hand but can be constructed by sampling in stages and assembling a partial frame at each stage. In both the cases of single-stage and multistage sampling, auxiliary data may be used to construct efficient estimators of totals.

A superpopulation model is a way of formalizing the relationship between a target variable and auxiliary data. For example, in a survey of hospitals, the number of discharges of patients in a particular calendar quarter may be related to the number of beds in the hospital and the type of hospital (e.g., general medical and surgical, rehabilitation, children's hospital, military, etc.). Quantities of interest are modeled as being realizations of random variables with a particular joint probability distribution. For example, in the case of the hospital population, the model might be

$$Y_i = \beta x_i + \varepsilon_i; \quad i = 1, \ldots, N \tag{1}$$

where Y_i is the number of discharges for hospital i, x_i is the number of beds in the hospital, β is an unknown parameter, N is the number of hospitals in the population, and the ε_is are uncorrelated random errors with mean 0 and variance $\sigma^2 x_i$. This simple model says that the number of discharges is, on average, proportional to the number of beds and that the variability among discharges is larger for the hospitals with the

larger numbers of beds. More elaborate models might be more realistic. We might, for instance, want to use a different model for the different types of hospitals or to include a quadratic term in x.

Models such as this can be used for constructing estimators, determining sample sizes, and assessing the precision of estimates. The population total of the Y is $T = \sum_U Y_i$, where U denotes the set of N units in the population (or universe). If a sample s of n units is selected from the N in the population, we can observe the sample total, $T_s = \sum_s Y_i$. The total for the remainder of the population, $r = U - s$, is equal to $T_r = \sum_r Y_i$, which is unknown and must be estimated using the sample units.

A logical approach is to treat this as a prediction problem in regression, predict each nonsample Y, and then add the predictions. The best linear unbiased estimator of β under model (1) is $\hat{\beta} = \overline{Y}_s/\overline{x}_s$ where $\overline{Y}_s = \sum_s Y_i/n$ and $\overline{x}_s = \sum_s x_i/n$. Thus, an estimator of the Y-value for each unit in r is $\hat{Y}_i = \hat{\beta}x_i$. An intuitive estimator, or more properly, predictor, of T is then $\hat{T} = T_s + \sum_r \hat{Y}_i$. After some simplification, this predictor can be written as $\hat{T} = \hat{\beta}\sum_U x_i = N\overline{Y}_s\overline{x}_U/\overline{x}_s$ with $\overline{x}_U = \sum_U x_i/N$. This is known as the ratio estimator and is the best linear unbiased predictor (BLUP) of T under model (1).

Figure 1 illustrates the general situation in estimating a finite population total. The upper panel of the figure is a plot of Y versus x for a hypothetical population. The middle plot shows a sample from the same population. The gray circles mark the x values of the nonsample units. In the model-based approach, a prediction is made for each nonsample unit and the total of the predictions is added to the observed sample total to estimate T. In the lower panel, the weighted least squares regression line estimated from the sample is superimposed for the model $E_M(Y) = \beta x$, $\text{var}_M(Y) = \sigma^2 x$. The gray points on the line are the predicted values for the nonsample points. To get a good aggregate prediction of the total, predicting the realized value for each nonsample point is unnecessary; we need only estimate the mean via $\hat{Y}_i = \hat{\beta}x_i$. Naturally, if the model is incorrectly specified, the predictions can be poor, leading to a biased estimator of the total. Guarding against certain types of model misspecification has been a major concern in the prediction approach, as discussed in Section 4. As in any regression application, having a sample that covers the range of covariates in the population is important to have some assurance that an assumed model is a reasonable fit throughout the range.

A key feature of this approach is that only the model is used to generate the estimator. The model is also the basis for estimating the precision of $\hat{T}$ and for making inferences about T (e.g., through confidence intervals). There are no restrictions on how the sample is selected other than that the sample units must follow model (1). Said more prosaically: the model that holds for the sample must be the same one that holds for the population. There are cases of informative sampling where this requirement is violated; methods of handling such situations are discussed in Pfeffermann (1993, 1996), Sverchkov and Pfeffermann (2004) and Chapter 39. In the model-based approach or the Bayesian approach (see Chapter 29), there is no assumption that a random sampling plan is used. This is in contrast to the *design-based* theory of sampling (see Chapter 1) and the model-assisted theory (see Chapter 25) in which a random sampling plan must be used because it generates the distribution which is the basis for inference. However, use of randomization in the model-based and Bayesian approaches does have the same justification as in experimental design. Randomization insulates the sample designer from accusations of personal bias in selection, and it provides, in expectation, certain types of balance on covariates between the sample and nonsample units.

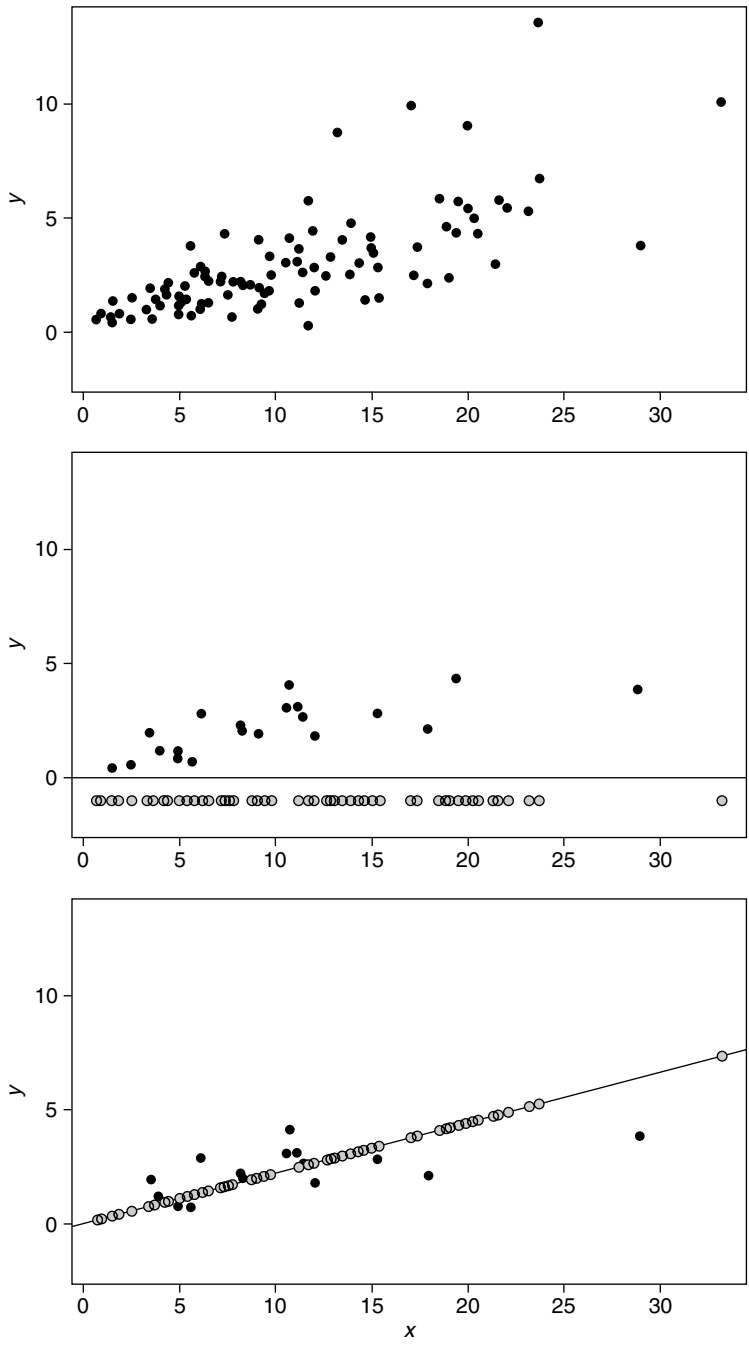


Fig. 1. A hypothetical population with one auxiliary (x) variable. The top panel shows the full population. The middle panel shows a sample from the population (black dots). The gray circles mark the nonsample units and their x' for which Y must be predicted. The lower panel shows the weighted least squares line estimated from the sample for the model $E_M(Y) = \beta x$, $\mathrm{var}_M(Y) = \sigma^2 x$. The gray dots are the predicted values for the nonsample units.

The appropriate approach to inference in finite population sampling has been an interesting source of debate and controversy over the last 50 years or so. This chapter does not address these foundational issues, but some discussion can be found in Basu (1971), Godambe (1955, 1966), Royall (1976, 1994), Smith (1976, 1984, 1994), and Valliant et al. (2000). This last reference also gives many of the technical details of results that are summarized in the rest of this chapter.

2. Prediction under the general linear model

Estimation of a total can be formulated for a general linear model and the BLUP derived. The finite population consists of N units, each of which has a value of a target variable y associated with it. The population vector of y is $\mathbf{y} = (y_1, \ldots, y_N)'$ and is treated as the realization of a random vector $\mathbf{Y} = (Y_1, \ldots, Y_N)'$. Our goal will be to estimate a linear combination of the y, $\boldsymbol{\gamma}'\mathbf{y}$, where $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_N)'$ is an N-vector of constants. If each $\gamma_i = 1$, then the target is the total; if $\gamma_i = 1/N$, the target is the mean. From the population of N units, a sample s of n units is selected, and the y values of the sample units are observed. Denote the set of nonsample units, that is, the remainder of the population, by r. For any sample s we can reorder the population vector y so that the first n elements are those in the sample and the last $N - n$ are those in the nonsample: $\mathbf{y} = (\mathbf{y}_s', \mathbf{y}_r')'$ where $\mathbf{y}_s$ is the n-vector of observed values and $\mathbf{y}_r$ the $N - n$ unobserved. The vector $\boldsymbol{\gamma}$ can also be partitioned into parts corresponding to the sample and nonsample units, $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_s', \boldsymbol{\gamma}_r')'$. The estimation target can now be expressed as $\boldsymbol{\gamma}'\mathbf{y} = \boldsymbol{\gamma}_s'\mathbf{y}_s + \boldsymbol{\gamma}_r'\mathbf{y}_r$, which is the realization of the random variable $\boldsymbol{\gamma}'\mathbf{Y} = \boldsymbol{\gamma}_s'\mathbf{Y}_s + \boldsymbol{\gamma}_r'\mathbf{Y}_r$. Because the sum for the sample units, $\boldsymbol{\gamma}_s'\mathbf{y}_s$, is known after the sample is selected and the values for its units observed, the problem of estimating $\boldsymbol{\gamma}'\mathbf{y}$ is logically equivalent to that of predicting the value, $\boldsymbol{\gamma}_r'\mathbf{y}_r$, of the unobserved random variable $\boldsymbol{\gamma}_r'\mathbf{Y}_r$.

The types of estimators considered here are linear combinations of the sample Y as defined in the following section.

DEFINITION 1. A *linear estimator* of $\theta = \boldsymbol{\gamma}'\mathbf{Y}$ is defined as $\hat{\theta} = \mathbf{g}_s'\mathbf{Y}_s$ where $\mathbf{g}_s = (g_1, \ldots, g_n)'$ is an n-vector of coefficients.

DEFINITION 2. The *estimation error* of an estimator $\mathbf{g}_s'\mathbf{Y}_s$ is $\hat{\theta} - \theta = \mathbf{g}_s'\mathbf{Y}_s - \boldsymbol{\gamma}'\mathbf{Y}$.

The estimation error can be rewritten in terms of the sample and nonsample units as

$$\begin{aligned}\mathbf{g}_s'\mathbf{Y}_s - \boldsymbol{\gamma}'\mathbf{Y} &= (\mathbf{g}_s' - \boldsymbol{\gamma}_s')\mathbf{Y}_s - \boldsymbol{\gamma}_r'\mathbf{Y}_r, \\ &= \mathbf{a}'\mathbf{Y}_s - \boldsymbol{\gamma}_r'\mathbf{Y}_r\end{aligned}$$

where $\mathbf{a} = \mathbf{g}_s - \boldsymbol{\gamma}_s$. Using $\mathbf{g}_s'\mathbf{Y}_s$ to estimate $\boldsymbol{\gamma}'\mathbf{Y}$ is equivalent to using $\mathbf{a}'\mathbf{Y}_s$ to predict $\boldsymbol{\gamma}_r'\mathbf{Y}_r$. Consequently, finding a good $\mathbf{g}_s$ is equivalent to finding a good $\mathbf{a}$. This prediction problem can be studied under the general linear model M:

$$E_M(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta},\ \mathrm{var}_M(\mathbf{Y}) = \mathbf{V}, \tag{2}$$

where $\mathbf{X}$ is an $N \times p$ matrix of auxiliaries, $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown parameters, and $\mathbf{V}$ is a positive definite covariance matrix. This is also the setting for much of calibration

and model-assisted estimation (see Chapter 25). In much of this development, it is assumed that all auxiliary values are known for each unit in the population. In some special cases, this condition can be relaxed to require only that population totals of the components of $\mathbf{X}$ be known. The population elements are rearranged so that the first n elements of $\mathbf{Y}$ are those in the sample, and the first n rows of $\mathbf{X}$ are for units in the sample. Then, $\mathbf{X}$ and $\mathbf{V}$ can be expressed as

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_s \\ \mathbf{X}_r \end{bmatrix} \quad \mathbf{V} = \begin{bmatrix} \mathbf{V}_{ss} & \mathbf{V}_{sr} \\ \mathbf{V}_{rs} & \mathbf{V}_{rr} \end{bmatrix},$$

where $\mathbf{X}_s$ is $n \times p$, $\mathbf{X}_r$ is $(N-n) \times p$, $\mathbf{V}_{ss}$ is $n \times n$, $\mathbf{V}_{rr}$ is $(N-n) \times (N-n)$, $\mathbf{V}_{sr}$ is $n \times (N-n)$, and $\mathbf{V}_{rs} = \mathbf{V}'_{sr}$. Assume that $\mathbf{V}_{ss}$ is positive definite.

DEFINITION 3. The estimator $\hat{\theta}$ is unbiased (or, equivalently, *prediction unbiased* or *model unbiased*) for θ under a model M if $E_M(\hat{\theta} - \theta) = 0$.

DEFINITION 4. The *error variance* (or, equivalently, *prediction variance*) of $\hat{\theta}$ under a model M is $E_M(\hat{\theta} - \theta)^2$.

If the auxiliaries are known for every unit in the population, this implies that a sampling frame has been constructed that lists every unit in the survey universe. In a universe of elementary-level schools, auxiliaries could include the number of students and teachers, location of the school (urban, suburban, or rural), and total budget in a recent year. There are many applications in which a complete list of every population unit is not available. For example, in personal-interview surveys of households in countries without population registries, multistage sampling is often used. The sample is selected in a series of stages with each stage being a different type of unit. The first two or three stages may be successively smaller geographic areas. The last stage may be the households that are the target of the survey. Models for clustered populations are usually appropriate in such cases, as discussed in section 7.

The general prediction theorem (Royall, 1976), giving the BLUP of $\hat{\theta}$ under model (2) is

THEOREM 1. *Among linear, prediction-unbiased estimators $\hat{\theta}$ of θ, the error variance is minimized by*

$$\hat{\theta}_{\text{opt}} = \boldsymbol{\gamma}'_s \mathbf{Y}_s + \boldsymbol{\gamma}'_r \left[\mathbf{X}_r \hat{\boldsymbol{\beta}} + \mathbf{V}_{rs} \mathbf{V}_{ss}^{-1} \left(\mathbf{Y}_s - \mathbf{X}_s \hat{\boldsymbol{\beta}} \right) \right], \tag{3}$$

where $\hat{\boldsymbol{\beta}} = \mathbf{A}_s^{-1} \mathbf{X}'_s \mathbf{V}_{ss}^{-1} \mathbf{Y}_s$ with $\mathbf{A}_s = \mathbf{X}'_s \mathbf{V}_{ss}^{-1} \mathbf{X}_s$. The error variance of $\hat{\theta}$ is

$$\begin{aligned} \text{var}_M(\hat{\theta} - \theta) = {} & \boldsymbol{\gamma}'_r \left(\mathbf{X}_r - \mathbf{V}_{rs} \mathbf{V}_{ss}^{-1} \mathbf{X}_s \right) \mathbf{A}_s^{-1} \left(\mathbf{X}_r - \mathbf{V}_{rs} \mathbf{V}_{ss}^{-1} \mathbf{X}_s \right)' \boldsymbol{\gamma}_r \\ & + \boldsymbol{\gamma}'_r \left(\mathbf{V}_{rr} - \mathbf{V}_{rs} \mathbf{V}_{ss}^{-1} \mathbf{V}_{sr} \right) \boldsymbol{\gamma}_r. \end{aligned} \tag{4}$$

A feature of the BLUP is that it equals the weighted sum for the sample units, $\boldsymbol{\gamma}'_s \mathbf{Y}_s$, plus a predictor of the weighted sum for the nonsample units, $\boldsymbol{\gamma}'_r [\mathbf{X}_r \hat{\boldsymbol{\beta}} + \mathbf{V}_{rs} \mathbf{V}_{ss}^{-1} (\mathbf{Y}_s - \mathbf{X}_s \hat{\beta})]$. When the sample and nonsample units are uncorrelated, that is, $\mathbf{V}_{rs} = \mathbf{0}$, the BLUP simplifies to $\hat{\theta}_{\text{opt}} = \boldsymbol{\gamma}'_s \mathbf{Y}_s + \boldsymbol{\gamma}'_r \mathbf{X}_r \hat{\boldsymbol{\beta}}$. The assumption that sample and nonsample units are

uncorrelated will often be reasonable in populations where single-stage sampling is appropriate, such as institution or establishment sampling. However, in clustered populations, described later in this chapter, units within the same cluster will be correlated, a feature that must be accounted for in analysis.

To appreciate the formulation of the theorem as one of prediction, rather than estimation, it is instructive to look at the results for the optimum $\hat{\theta}$ if we minimize its variance, $\text{var}_M(\hat{\theta}) = \mathbf{g}_s'\mathbf{V}_{ss}\mathbf{g}_s$ instead of the error variance $\text{var}_M(\hat{\theta} - \theta)$. In that case, the minimum variance estimator is $\hat{\theta}^* = \boldsymbol{\gamma}'\mathbf{X}\hat{\boldsymbol{\beta}}$. In other words, the value for each unit in the population is estimated as its expected value from the estimated regression model. Contrast this to $\hat{\theta}_{\text{opt}}$ where the sum for the sample units, $\boldsymbol{\gamma}_s'\mathbf{Y}_s$, is used directly, and the sum for the nonsample units is predicted by the estimated regression mean, $\boldsymbol{\gamma}_r'\mathbf{X}_r\hat{\boldsymbol{\beta}}$, plus an adjustment based on sample residuals, $\boldsymbol{\gamma}_r'\mathbf{V}_{rs}\mathbf{V}_{ss}^{-1}(\mathbf{Y}_s - \mathbf{X}_s\hat{\boldsymbol{\beta}})$.

Many commonly used estimators can be derived by applying Theorem 1 to particular models. In the examples below, the estimation target is the finite population total $T = \sum_U Y_i$, implying that $\boldsymbol{\gamma} = 1_N$, a vector of N ones. Suppose that the model is $Y_i = \mu + \varepsilon_i$ with the ε_i being uncorrelated and $\varepsilon_i \sim (0, \sigma^2)$. Then, in model (2) $\boldsymbol{\beta} = \mu$, $\mathbf{X} = \mathbf{I}_N$, $\mathbf{V} = \sigma^2\mathbf{I}_N$, and $\hat{\boldsymbol{\beta}} = \overline{Y}_s \equiv \sum_s Y_i/n$. The BLUP is the expansion estimator, $\hat{T}_0 = \sum_s Y_i + \sum_r \overline{Y}_s = N\overline{Y}_s$, with the implied prediction for each nonsample unit being $\overline{Y}_s$. The error variance of the expansion estimator is

$$\text{var}_M\left(\hat{T}_0 - T\right) = \frac{N^2}{n}(1 - f)\sigma^2,$$

where $f = n/N$. This is also the usual, design-based variance formula under simple random sampling without replacement.

The model that leads to the ratio estimator, as noted earlier, is (1). The estimator itself is $\hat{T}_R = \hat{\beta}\sum_U x_i$ with $\hat{\beta} = \overline{Y}_s/\overline{x}_s$, and its error variance under the model is

$$\text{var}_M\left(\hat{T}_R - T\right) = \frac{N^2}{n}(1 - f)\frac{\overline{x}_r\overline{x}_U}{\overline{x}_s}\sigma^2,$$

where $\overline{x}_r$ is the mean of x for the nonsample units, $\overline{x}_U$ is the population mean, and $f = n/N$ is the sampling fraction.

The simple linear regression estimator comes from the model $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ with the ε_i being independent with mean 0 and variance σ^2. The BLUP is $\hat{T}_{\text{LR}} = N[\overline{Y}_s + \hat{\beta}_1(\overline{x}_U - \overline{x}_s)]$ where $\hat{\beta}_1 = \sum_s(x_i - \overline{x}_s)(Y_i - \overline{Y}_s)/\sum_s(x_i - \overline{x}_s)^2$. The error variance is $\text{var}_M(\hat{T}_{\text{LR}} - T) = \frac{N^2}{n}(1 - f)\sigma^2\left[1 + n(\overline{x}_s - \overline{x}_U)^2/(1 - f)\sum_s(x_i - \overline{x}_s)^2\right]$.

Another common estimator is the stratified expansion estimator. A set of strata is a collection of mutually exclusive groups that covers the entire population. Strata might be regions of a country, types of industries, or size classes of schools. Suppose that h denotes a stratum and that the model is $Y_{hi} = \mu_h + \varepsilon_{hi}$ with the ε_{hi} being uncorrelated and $\varepsilon_{hi} \sim (0, \sigma_h^2)$. The BLUP is $\hat{T}_{st} = \sum_{h=1}^H N_h\overline{Y}_{hs}$ where N_h is the number of population units in stratum h, $\overline{Y}_{hs} = \sum_{s_h} Y_{hi}/n_h$, s_h is the set of sample units in stratum h, and n_h is the number of sample units from that stratum. The error variance is $\text{var}_M(\hat{T}_{st} - T) = \sum_{h=1}^H N_h^2(1 - f_h)\sigma_h^2/n_h$ with $f_h = n_h/N_h$.

A final example is the mean-of-ratios estimator which flows from the model $Y_i = \beta x_i + \varepsilon_i$ where $\varepsilon_i \sim (0, \sigma^2 x_i^2)$. The BLUP is $\hat{T} = \sum_s Y_i + \hat{\beta}\sum_r x_i$ with

$\hat{\beta} = n^{-1} \sum_s Y_i/x_i$. The error variance is $\text{var}_M(\hat{T}_R - T) = [(N-n)^2 \bar{x}_r^2 / n + \sum_r x_i^2]\sigma^2$. When the sampling fraction is small, the BLUP is approximated by the mean-of-ratios estimator, $\hat{T} = N\bar{x}_U\hat{\beta}$.

The models that can be used in estimating a total are by no means limited to the simple ones mentioned above. A mixture of quantitative and qualitative auxiliaries along with interactions, nested structures, and other complexities may be needed in some populations. These possibilities are all within the scope covered by Theorem 1.

3. Estimation weights

When constructing a database from a sample survey, standard procedure is to have a *weight* associated with each unit in the sample that is used to calculate linear estimates. The weights are intended to be applicable to several Y variables of interest. For a single y variable, the $n \times 1$ optimal vector of coefficients in a linear estimator of a total, implied by Theorem 1, is

$$\mathbf{g}_s = \mathbf{V}_{ss}^{-1}\left[\mathbf{V}_{sr} - \mathbf{X}_s\mathbf{A}_s^{-1}\left(\mathbf{X}_r' - \mathbf{X}_s'\mathbf{V}_{ss}^{-1}\mathbf{V}_{sr}\right)\right]\mathbf{1}_r + \mathbf{1}_s,$$

where $\mathbf{1}_r$ and $\mathbf{1}_s$ are, respectively, vectors of $N-n$ and n 1's. Unit i is assigned a weight equal to the ith component of the vector $\mathbf{g}_s$. The optimal weight depends, through the covariance structure, on the particular y variable being considered and on the way that the population is split between the sample and nonsample units.

The simple examples from Section 2 are ones in which the weight for a sample unit does not depend on the particular Y being studied. For the expansion estimator, $g_i = N/n$ for all units in the sample. The ratio estimator has $g_i = N\bar{x}_U/(n\bar{x}_s)$, which is also the same for all $i \in s$. The linear regression estimator has $g_i = N[n^{-1} + (\bar{x}_U - \bar{x}_s)(x_i - \bar{x}_s)]\big/\sum_{j\in s}(x_j - \bar{x}_s)^2$. The weight for the stratified expansion estimator is N_h/n_h for $i \in s_h$ and for the mean-of-ratios estimator is $g_i = N\bar{x}_U/nx_i$. The simple linear regression and the mean-of-ratios estimators are cases where the weight depends on the particular sample unit through x_i.

Although common survey practice is to use the same weight to make an estimate for different y variables, this is justifiable from the model-based point-of-view only when the y follow the same general form of model. If one variable follows the expansion estimator model while another follows the regression estimator model, using the same weight for each is not generally sensible. However, in the case where the sample is balanced, in the sense given in Section 4, estimators of many forms, can in effect be subsumed under one form, so that per-unit weights are well-grounded.

All the examples we have considered share a certain common structure that will be discussed in more detail in the next section. Suppose that $\mathbf{V}$ is diagonal and that the ith diagonal element can be expressed as $v_i = \sigma^2 f(\mathbf{x}_i)$ with $f(\mathbf{x}_i) = \sum_{j=1}^p c_j x_{ij}$ being a known function and $\mathbf{x}_i$ the p-vector of auxiliaries for unit i. In matrix terms, suppose that $\mathbf{V1}_N = \mathbf{Xc}$ for a $p \times 1$ vector $\mathbf{c}$. The BLUP becomes $\hat{T} = N\bar{\mathbf{x}}_U\hat{\boldsymbol{\beta}}$ (see Lemma 1 below) where $\bar{\mathbf{x}}_U = (\bar{x}_{U1}, \ldots, \bar{x}_{Up})'$ is the vector of population means of the auxiliaries. Notice that this form of T is the same as would be obtained under the general linear model if we minimized $\text{var}_M(\hat{T})$ rather than $\text{var}_M(\hat{T} - T)$. Even if the variance condition $\mathbf{V1}_N = \mathbf{Xc}$ does not hold, $\hat{T} = N\bar{\mathbf{x}}_U\hat{\boldsymbol{\beta}}$ is still prediction unbiased under model (2). The weight for the ith sample unit is then $g_i = N\bar{\mathbf{x}}_U\left[\sum_s \mathbf{x}_i\mathbf{x}_i'/f(\mathbf{x}_i)\right]^{-1}\mathbf{x}_i/f(\mathbf{x}_i)$. This

weight depends on the variance structure only through $f(\mathbf{x}_i)$. However, these weights may differ for different Y variables because the vector $\mathbf{c}$ in the condition $\mathbf{V1}_N = \mathbf{Xc}$ may depend on Y. Another option would be to use the ordinary least squares slope estimate, $\hat{\boldsymbol{\beta}}^* = (\mathbf{X}_s'\mathbf{X}_s)^{-1}\mathbf{X}_s'\mathbf{Y}_s$, which ignores any nonconstant variance structure. The estimator, $\hat{T}^* = N\overline{\mathbf{x}}_U\hat{\boldsymbol{\beta}}^*$ with weights $g_i = N\overline{\mathbf{x}}_U(\mathbf{X}_s'\mathbf{X}_s)^{-1}\mathbf{x}_i$, is also prediction-unbiased under (2) but is inefficient.

4. Weighted balance and robustness

In this section, we describe the idea of balanced samples and show that BLUP of total based on models fulfilling certain conditions are bias-robust in weighted balanced samples. Furthermore, there exists a lower bound on the error variance of the BLUP under these conditions, and this bound is only achieved if the sample is balanced.

Any model used to generate an estimator should, at best, be considered a "working" model. That is, one that may be plausible based on data or prior knowledge but that may be wrong. If the working model is wrong, then the BLUP based on that model can be seriously biased. Consider the ratio estimator but suppose that the correct model includes an intercept: $E_{M^*}(Y_i) = \alpha + \beta x_i$. The bias of the ratio estimator under M^* is $E_{M^*}(\hat{T}_R - T) = N\alpha(\overline{x}_U/\overline{x}_s - 1)$. When the sample is balanced in the sense that $\overline{x}_s = \overline{x}_U$, the ratio estimator is still unbiased, even though the straight-line through the origin, working model is violated. If $\overline{x}_s$ is much smaller than $\overline{x}_U$, $\hat{T}_R$ will be an over-estimate on average; when $\overline{x}_s$ is larger than $\overline{x}_U$, the opposite is true. Some random sampling plans, like simple random sampling (*srs*), are balanced on average. That is, $E_\pi(\overline{x}_s) = \overline{x}_U$ where E_π denotes design-expectation. On the other hand, a particular sample that has been randomly selected may be far from balanced. Averaging across all simple random samples obscures the conditional bias of $\hat{T}_R$ in some samples. The unconditional nature of design-based calculations has historically been one of the main objections to the design-based approach.

A more general version of balance is defined below and plays a role both in bias protection and optimality. Models that satisfy the variance condition $\mathbf{V1}_N = \mathbf{Xc}$ play a key role in robustness and optimality. Let $M(\mathbf{X} : \mathbf{V})$ refer to the special case of the general linear model given by (2) with matrix $\mathbf{X}$ of auxiliary variables, and diagonal covariance matrix $\mathbf{V} = \text{diag}(\sigma_i^2)$, $i \in U$. Define $\mathbf{W}$ to be an $N \times N$ diagonal matrix and $\mathbf{W}_s$ to be the $n \times n$ diagonal submatrix for the sample units. When $\mathbf{V1}_N = \mathbf{Xc}$, the BLUP and its variance simplify as shown in Lemma 1.

LEMMA 1 (Royall, 1992). *If* $\mathbf{V1}_N = \mathbf{Xc}$ *for some vector* $\mathbf{c}$*, then the BLUP predictor and its error variance are* $\hat{T} = N\overline{\mathbf{x}}_U\hat{\boldsymbol{\beta}}$ *and* $var_M(\hat{T} - T) = \left(N^2\overline{\mathbf{x}}_U'\mathbf{A}_s^{-1}\overline{\mathbf{x}}_U - \mathbf{1}_N'\mathbf{V1}_N\right)$.

DEFINITION 5. The collection of samples that satisfy

$$\frac{1}{n}\mathbf{1}_s'\mathbf{W}_s^{-1/2}\mathbf{X}_s = \frac{\mathbf{1}_N'\mathbf{X}}{\mathbf{1}_N'\mathbf{W}^{1/2}\mathbf{1}_N} \tag{5}$$

will be denoted $B(\mathbf{X} : \mathbf{W})$, and said to be *balanced with respect to the weights root*$(\mathbf{W})$, *root*$(\mathbf{W})$ *balanced*, or to be *weighted balanced*.

The weighted balance condition can also be written as $n^{-1}\sum_s \mathbf{x}_i/w_i^{1/2} = \overline{x}/\overline{w}_U^{(1/2)}$ where $\overline{w}_U^{(1/2)} = \sum_U w_i^{1/2}/N$, the population mean of the square roots of the w weights.

The matrix $\mathbf{W}$ can be $\mathbf{V}^{-1}$ or another arbitrary weight matrix. When $\mathbf{W} = \mathbf{I}$, $B(\mathbf{X} : I)$ is the set of samples that are balanced on the columns of $\mathbf{X}$, that is, $\overline{\mathbf{x}}_s = \overline{\mathbf{x}}_U$. This type of balance made the ratio estimator unbiased when $E_{M^*}(Y_i) = \alpha + \beta x_i$ was the correct model. If the model for Y is polynomial in x, then $\overline{x}_s^{(j)} = \overline{x}_U^{(j)}$ where $\overline{x}_s^{(j)} = \sum_s x_i^j/n$ and $\overline{x}_U^{(j)} = \sum_U x_i^j/N$ for $j = 1, \ldots, J$ with J being the degree of the polynomial.

Theorem 2 below gives the lower bound on the error variance of the BLUP in a certain class of models and shows that the bound is achieved in a weighted balanced sample. Let $\mathcal{M}(\mathbf{X})$ denote the linear manifold generated by the columns of $\mathbf{X}$, that is the vector space spanned by all linear combinations of the columns of $\mathbf{X}$.

THEOREM 2 (Royall, 1992). *Under* $M(\mathbf{X} : \mathbf{V})$ *if both* $\mathbf{V}\mathbf{1}_N$ *and* $\mathbf{V}^{1/2}\mathbf{1}_N \in \mathcal{M}(\mathbf{X})$, *then*

$$\mathrm{var}_M[\hat{T}(\mathbf{X} : \mathbf{V}) - T] \geq \left[n^{-1}(N\overline{\sigma}_U)^2 - N\overline{\sigma}_U^{(2)}\right],$$

where $\overline{\sigma}_U = \sum_U \sigma_i/N$ *and* $\overline{\sigma}_U^{(2)} = \sum_U \sigma_i^2/N$. *The bound is achieved if and only if* $s \in B(\mathbf{X} : \mathbf{V})$, *in which case* $\hat{T} = \frac{N\overline{\sigma}_U}{n}\sum_s \frac{Y_i}{\sigma_i}$.

Note that the reduced form of the BLUP in the theorem does not depend on the x in the model. Suppose that, instead of $M(\mathbf{X} : \mathbf{V})$, the correct model is $M(\tilde{\mathbf{X}} : \mathbf{V})$ where $\tilde{\mathbf{X}}$ includes all of the columns in $\mathbf{X}$ plus some additional ones. If the sample is *root*$(\mathbf{V})$ balanced on the columns of $\tilde{\mathbf{X}}$, that is, $\frac{1}{n}1_s'\mathbf{V}_s^{-1/2}\tilde{\mathbf{X}}_s = \frac{\mathbf{1}_N'\tilde{\mathbf{X}}}{\mathbf{1}_N'\mathbf{V}^{1/2}\mathbf{1}_N}$, then the BLUP still reduces to $\hat{T} = N\overline{\sigma}_U\sum_s (Y_i/\sigma_i)/n$. As a result, weighted balanced sampling, using an augmented version of the auxiliary matrix $\mathbf{X}$**,** is robust to misspecification of the matrix of auxiliaries in the model.

As an illustration, consider a model in which the variance is proportional to x. If the model for Y is $Y_i = \beta_1 x_i^{1/2} + \beta_2 x_i + \varepsilon_i$ with the errors being uncorrelated and $\varepsilon_i \sim (0, \sigma^2 x_i)$. The conditions of Theorem 2 are satisfied because $\sigma_i^2 = \sigma^2 x_i$ and $\sigma_i = \sigma x_i^{1/2}$ are both in the column space of $\mathbf{X}$. Call the BLUP under this model $\hat{T}(x^{1/2}, x : x)$. The lower bound on its variance is

$$\left[n^{-1}\left(N\overline{x}_U^{(1/2)}\right)^2 - N\overline{x}_U\right]\sigma^2$$

with $\overline{x}_U^{(1/2)} = N^{-1}\sum_U x_i^{(1/2)}$. The lower bound is achieved in any sample that is balanced in the sense that $\overline{x}_s^{(1/2)} = \overline{x}_U/\overline{x}_U^{(1/2)}$ where $\overline{x}_s^{(1/2)} = n^{-1}\sum_s x_i^{1/2}$. Bias protection against general polynomial models is obtained by balancing on additional powers:

$$n^{-1}\sum_s x_i^{j-1/2} = \overline{x}_U^{(j)}/\overline{x}_U^{(1/2)} \quad \text{for} \quad j = 1, \ldots, J.$$

The additional powers correspond to the additional columns in $\tilde{\mathbf{X}}$ mentioned earlier. For example, if $\mathrm{var}_M(Y_i) = \sigma^2 x_i$ and the correct mean specification in the model is $E_M(Y_i) = \beta_1 x_i^{1/2} + \beta_2 x_i + \beta_3 x_i^2$, then the weighted balance conditions are $n^{-1}\sum_s x_i^{1/2} = \overline{x}_U/\overline{x}_U^{(1/2)}$ and $n^{-1}\sum_s x_i^{3/2} = \overline{x}_U^{(2)}/\overline{x}_U^{(1/2)}$. In such a sample, the BLUP under the working model

$M(x^{1/2}, x : x)$ is $\hat{T} = N\bar{x}_U^{(1/2)} \sum_s \left(Y_i / x_i^{1/2} \right) \Big/ n$ and is unbiased under both that working model and the extended model $M(x^{1/2}, x, x^2 : x)$.

Exact methods for obtaining weighted balanced samples are described in Chauvet and Tillé (2006) and Tillé (2006, Chapter 2). These methods are implemented in the R software package, `sampling` available at www.r-project.org.

Standard methods of probability proportional to size (pps) sampling (see Chapter 2) can also be used to approximate weighted balanced samples. Suppose that the size measure, according to which pps is carried out, is $\sigma_i = [\text{var}_M(Y_i)]^{1/2}$ so that, for a fixed sample size, the inclusion probability of unit i is $\pi_i = n\sigma_i / N\overline{\sigma}_U$. We refer to this as a pp(σ) plan. Then, the sample is balanced in design-expectation because

$$E_{\text{pps}} \left(n^{-1} \sum_s \mathbf{x}_i / \sigma_i \right) = \overline{\mathbf{x}}_U / \overline{\sigma}_U,$$

where pp(σ) is expectation with respect to pp(σ) sampling.

Consequently, a pp(σ) plan produces weighted balance on average across all possible pp(σ) samples. However, this may be far from true in any particular sample. One practical approach is select a pp(σ) sample and retain it if the sample moments, $n^{-1} \sum_s \mathbf{x}_i / \sigma_i$, are within some prespecified tolerances of $\overline{\mathbf{x}}_U / \overline{\sigma}_U$. This method of restricted sampling was illustrated first by Herson (1976) for simple random sampling and in Valliant et al. (2000) for various sampling plans, including srs and pps.

The formulas for the BLUP and its lower bound are also found in design-based theory. If the inclusion probability of unit i is $\pi_i = n\sigma_i / N\overline{\sigma}_U$ in a fixed size sample, then the BLUP, $\hat{T} = N\overline{\sigma}_U \sum_s (Y_i/\sigma_i)/n$, is the Horvitz–Thompson estimator $\sum_s y_i/\pi_i$. The variance bound is the one established by Godambe and Joshi (1965, Theorem 6.1) for the model-based expectation of the design-based variance of the Horvitz–Thompson estimator. Isaki and Fuller (1982) also showed that this bound is approached asymptotically by an estimator based on a regression model that includes the standard deviations and variances (of the Y given the x) in the column space of the $\mathbf{X}$ matrix, when the inclusion probabilities are proportional to the standard deviations.

One of the practical constraints in many survey applications is that a sample that is selected to be balanced does not remain so at the analysis stage. Losses due to nonresponse (see Chapter 9) and ineligible units can destroy balance. In such a case, careful modeling is essential, and it may be wise to include auxiliaries in the estimator beyond those in the working model if extra x are available. Accounting populations of transactions or bookkeeping entries are examples where balance can often be selected and maintained through data collection.

5. Variance estimation

Because the total T itself is the realization of a random variable, we want to estimate the mean square error or error variance, $E_M(\hat{T} - T)^2$. If $\hat{T}$ is unbiased, the error variance is just the variance $\text{var}_M(\hat{T} - T)$. Once we have an estimate v of this variance, a confidence interval for T of the form $\hat{T} \pm z v^{1/2}$ can be constructed, where z is the appropriate quantile of the standard normal distribution. This type of interval is justified because, under appropriate conditions, $\hat{T}$ will be asymptotically normal.

Deviations of the working model from models that might be better descriptions of the population values are a concern when estimating the variance. If the variance structure assumed in the working model is wrong, standard least squares variance estimators are vulnerable to bias. However, it is possible to construct variance estimators that are robust to this departure. Robustness to misspecification of $E_M(Y_i)$ can be achieved through balance, as described in Section 4, but this may be difficult, especially if there are omitted and unknown regressors that should be included in $E_M(Y_i)$. If $\hat{T}$ is biased, then $E_M(\hat{T}-T)^2$ will contain a bias-squared component that typically cannot be estimated.

The estimation error of any estimator of the total can be written as $\hat{T}-T=\hat{T}_r-T_r$, where T_r is the total for the nonsample units and $\hat{T}_r$ is an estimator of T_r based on the sample units. When the sample and nonsample units are independent, $\hat{T}_r=\mathbf{1}_r'\mathbf{X}_r\hat{\boldsymbol{\beta}}$, with $\hat{\boldsymbol{\beta}}$ a weighted least squares estimator of the underlying parameter $\boldsymbol{\beta}$. We can also write $\mathrm{var}_M(\hat{T}-T)=\mathrm{var}_M(\hat{T}_r-T_r)=\mathrm{var}_M(\hat{T}_r)+\mathrm{var}_M(T_r)$. Under typical sampling conditions, the first component, $\mathrm{var}_M(\hat{T}_r)$, has order $O((N-n)^2/n)$ while the second is $O(N-n)$. If $n/N\to 0$ as $n, N\to\infty$, the first component of the variance, $\mathrm{var}_M(\hat{T}_r)$, dominates and is, consequently, the more important one to estimate.

The idea behind robust variance estimation is fairly simple. Define the working model to be

$$E_M(Y_i)=\mathbf{x}_i'\boldsymbol{\beta},\quad \mathrm{var}_M(Y_i)=\sigma_i^2 \tag{6}$$

with the Y being uncorrelated. A weighted least squares variance estimator can be derived under this working model, but it will perform poorly when σ_i^2 is misspecified (Royall and Cumberland, 1978, 1981a). Consider the case where $\mathrm{var}_M(Y_i)=\psi_i$ rather than the σ_i^2 in the working model. The estimator of the nonsample total can be rewritten as $\hat{T}_r=\sum_s a_iY_i$ with $a_i=\mathbf{1}_r'\mathbf{X}_r\mathbf{A}_s^{-1}\mathbf{x}_i\big/\sigma_i^2$. The error variance is

$$\mathrm{var}_M(\hat{T}_r)+\mathrm{var}_M(T_r)=\sum_s a_i^2\psi_i+\sum_r\psi_i. \tag{7}$$

Separate estimators are needed for these two components.

Estimators of the ψ_i for the sample units can be constructed using the regression *residuals*, defined as $r_i=Y_i-\hat{Y}_i$ with $\hat{Y}_i=\hat{\boldsymbol{\beta}}\mathbf{x}_i$. Under the working model (6), $E_M(r_i^2)=\sigma_i^2(1-h_i)$ with $h_i=\mathbf{x}_i'\mathbf{A}_s^{-1}\mathbf{x}_i\big/\sigma_i^2$ being the leverage (e.g., see Belsley et al., 1980, p. 16, Hoaglin and Welsch, 1978). More generally, $E_M(r_i^2)\cong\psi_i$ since $h_i\to 0$ under some weak conditions as n increases. Thus, two choices for estimating ψ_i are $\hat{\psi}_i=r_i^2$ and $\hat{\psi}_i=r_i^2(1-h_i)$. Both of these are robust in the sense that $E_M(r_i^2)\cong\psi_i$ regardless of whether ψ_i is known or not. An estimator of the first term in (7) is then $v(\hat{T}_r)=\sum_s a_i^2\hat{\psi}_i$. A third choice which, on average is an overestimate of ψ_i, is $\hat{\psi}_i=r_i^2/(1-h_i)^2$. Finally, a choice that amounts to an aggregate adjustment to $\sum_s a_i^2r_i^2$ is $\hat{\psi}_i=r_i^2\frac{\sum_s a_i^2\sigma_i^2}{\sum_s a_i^2\sigma_i^2(1-h_i)}$. The choice $\hat{\psi}_i=r_i^2$ generates what is known as a "sandwich" estimator because the estimator of $\mathrm{var}_M(\hat{T}_r)$ can be written as

$$\sum_s a_i^2r_i^2=\mathbf{a}'\mathrm{diag}\left(r_i^2\right)\mathbf{a}$$

with $\mathbf{a}=(a_1,\ldots,a_n)'$ and $\mathrm{diag}(r_i^2)$ being the $n\times n$ diagonal matrix of squared residuals.

Estimating the variance of the nonsample total, $\sum_r \psi_i$, requires more assumptions. Residuals for the nonsample units are not available because the Y is unknown. One strategy is to take $\sum_s \hat{\psi}_i$ and inflate it to the nonsample size using quantities from the working model (6). A reasonable approach is to use

$$v(T_r) = \frac{\sum_r \sigma_i^2}{\sum_s \sigma_i^2} \sum_s \hat{\psi}_i.$$

An estimator of the error variance is then $v(\hat{T}) = \sum_s a_i^2 \hat{\psi}_i + \frac{\sum_r \sigma_i^2}{\sum_s \sigma_i^2} \sum_s \hat{\psi}_i$ with $\hat{\psi}_i$ being one of the three choices noted earlier.

The choice $\hat{\psi}_i = r_i^2/(1-h_i)^2$ is closely related to the jackknife. One version of the jackknife variance estimator is defined as

$$v_J(\hat{T}) = \frac{n-1}{n} \sum_s (\hat{T}_{(i)} - \hat{T})^2,$$

where $\hat{T}_{(i)}$ is the BLUP calculated after deleting unit i from the sample. The jackknife can be rewritten exactly as

$$v_J(\hat{T}) = \frac{n-1}{n} \left\{ \sum_s \left(\frac{a_i r_i}{1-h_i} \right)^2 - \frac{1}{n} \left(\sum_s \frac{a_i r_i}{1-h_i} \right)^2 \right\}.$$

The second term in the braces converges in probability to 0. As a result, the jackknife is approximately equal to $\sum_s \left(\frac{a_i r_i}{1-h_i} \right)^2$, corresponding to the choice , $\hat{\psi}_i = r_i^2/(1-h_i)^2$. Adding an estimator of the variance of the nonsample total, a jackknife estimator of the error variance is $v_J(\hat{T} - T) = v_J(\hat{T}) + \frac{\sum_r \sigma_i^2}{\sum_s \sigma_i^2} \sum_s \hat{\psi}_i$.

6. Models with qualitative auxiliaries

In many populations, some of the most useful auxiliaries are qualitative rather than quantitative. For example, in surveys of persons, demographic variables like age group, race-ethnicity, and gender are useful predictors of response variables. Quantitative x can also be used in combination with qualitative ones. Some numerical issues arise because $\mathbf{A}_s = \mathbf{X}_s' \mathbf{V}_{ss}^{-1} \mathbf{X}_s$ is not invertible when one or more columns of $\mathbf{X}_s$ are a linear combination of others. This can happen if, say, dummy variables are included for both male and female. However, under model (2) the BLUP of $\theta = \boldsymbol{\gamma}' \mathbf{Y}$ can still be found as

$$\hat{\theta}_{\text{opt}} = \boldsymbol{\gamma}_s' \mathbf{Y}_s + \boldsymbol{\gamma}_r' \left[\mathbf{X}_r \boldsymbol{\beta}^o + \mathbf{V}_{rs} \mathbf{V}_{ss}^{-1} (\mathbf{Y}_s - \mathbf{X}_s \boldsymbol{\beta}^o) \right], \tag{8}$$

where $\boldsymbol{\beta}^o = \mathbf{G} \mathbf{X}_s' \mathbf{V}_{ss}^{-1} \mathbf{Y}_s$ with $\mathbf{G}$ being a generalized inverse (also called a g-inverse) of A_s (Valliant et al., 2000, Theorem 7.4.1). Although $\mathbf{G}$ is not unique, the predictor (8) is invariant to the choice of $\mathbf{G}$. Analogous to the formula in Theorem 1, the error variance

of $\hat{\theta}_{\text{opt}}$ is

$$\begin{aligned}\text{var}_M(\hat{\theta}_{\text{opt}} - \theta) &= \boldsymbol{\gamma}_r'\left(\mathbf{X}_r - \mathbf{V}_{rs}\mathbf{V}_{ss}^{-1}\mathbf{X}_s\right)\mathbf{G}\left(\mathbf{X}_r - \mathbf{V}_{rs}\mathbf{V}_{ss}^{-1}\mathbf{X}_s\right)'\boldsymbol{\gamma}_r \\ &\quad + \boldsymbol{\gamma}_r'\left(\mathbf{V}_{rr} - \mathbf{V}_{rs}\mathbf{V}_{ss}^{-1}\mathbf{V}_{sr}\right)\boldsymbol{\gamma}_r\end{aligned}$$

When the Y have a common variance, $\mathbf{V} = \sigma^2\mathbf{I}$, the BLUP and its error variance simplify to

$$\begin{aligned}\hat{\theta}_{\text{opt}} &= \boldsymbol{\gamma}_s'\mathbf{Y}_s + \boldsymbol{\gamma}_r'\mathbf{X}_r\boldsymbol{\beta}^o \\ \text{var}_M(\hat{\theta}_{\text{opt}} - \theta) &= \sigma^2\boldsymbol{\gamma}_r'\mathbf{X}_r\mathbf{G}\mathbf{X}_r'\boldsymbol{\gamma}_r + \sigma^2\boldsymbol{\gamma}_r'\boldsymbol{\gamma}_r\end{aligned} \tag{9}$$

with $\mathbf{G}$ being the generalized inverse of $\mathbf{X}_s'\mathbf{X}_s$. If $\mathbf{V} = \sigma^2\mathbf{I}$ is used to generate $\hat{T}$, the weight vector simplifies to $\dot{\mathbf{g}}_s = \mathbf{X}_s\mathbf{G}\mathbf{X}_r'\mathbf{1}_r + \mathbf{1}_s$. If the correct model has a more general covariance matrix than $\sigma^2\mathbf{I}$, $\hat{T}$ using these weights is not optimal but will still be unbiased. In applications where qualitative auxiliaries are typically used, there may be insufficient knowledge to specify a covariance matrix more complicated than $\sigma^2\mathbf{I}$ so that (9) is a practical choice.

One example of using a qualitative auxiliary is a model that specifies a separate mean in different groups:

$$Y_{ij} = \beta_0 + \beta_i + \varepsilon_{ij},\ \varepsilon_{ij} \sim (0, \sigma^2), \quad i = 1, \ldots, I; \quad j = 1, \ldots, N_i$$

with the errors being uncorrelated. Groups are denoted by i and units within groups by j. This model is the same as a one-way analysis-of-variance model. The BLUP of the population total is $\hat{T} = \sum_{i=1}^{I} N_i\overline{Y}_{si}$, which is the stratified expansion estimator defined earlier. This estimator is also the Horvitz–Thompson estimator in stratified simple random sampling.

7. Clustered populations

Many naturally occurring populations exhibit clustering in which units that are, in some sense, near each other have similar characteristics. Households in the same neighborhood may tend to have similar incomes, education levels of the heads of household, and amounts of expenditures on food and clothing. Business establishments in the same industry and geographic area will pay similar wages to a given occupation because of competition. Students within the same school may have similar scores on achievement tests because of similar demographics and classroom instruction. This similarity among "nearby" units can express itself statistically as a correlation between the target variables for different units.

In clustered populations, the methods of data collection may also differ from the methods used in other populations. In a household survey, for example, a complete list of households to use for sampling may not be available, especially if the population is large. The households may be geographically dispersed so that field work can be more economically done when sample units are clustered together to limit travel costs. A practical, and widely used, technique is to select the sample in stages, using, at each stage, sampling units for which a complete list is available or can be compiled as part of field work. In the household example, geographic areas may be selected at the first

stage. At the second stage, each first stage sample unit may be further subdivided and a sample of the subdivisions selected. A list of the households in each sample subdivision is then compiled and data collected from each. In a student population, schools may be selected at the first stage, a list of classrooms compiled in each sample school, and a sample of students then drawn from a sample of classrooms. Although students are the units ultimately sampled, a complete list of students for each school in the universe is unlikely to be available while a list of schools often is. Selecting schools at the first stage is also sensible because survey costs may depend on the number of sampled schools more than the number of sampled students. Cooperation must be elicited at the school level and the more schools in the sample, the more the survey will cost.

In clustered populations, an intraclass correlation model may be useful. We present one of the simpler versions here. Suppose the population is divided into N nonoverlapping clusters. Cluster i contains M_i units with the total number of units in the population being $M = \sum_{i=1}^{N} M_i$. Associated with unit j in cluster is a random variable Y_{ij}. The total of the Y is $T = \sum_{i=1}^{N} \sum_{j=1}^{M_i} Y_{ij}$. A simple working model is

$$\begin{aligned} E_M(Y_{ij}) &= \mu \\ \operatorname{cov}_M(Y_{ij}, Y_{kl}) &= \begin{cases} \sigma_i^2 & i = k,\ j = l \\ \sigma_i^2 \rho_i & i = k,\ j \neq l \\ 0 & i \neq k \end{cases} \end{aligned} \tag{10}$$

Under this model, units have a variance that can differ among the clusters; different units within the same cluster are correlated; and units in different clusters are uncorrelated. In some populations, more elaborate models that involve several levels of clustering are better descriptions. A population of school districts, schools within districts, classes within schools, and students within classes is an example. Multilevel models can then be used as discussed in Pfeffermann et al. (1998) and Rabe-Hesketh and Skrondal (2006). Note that this type of multilevel structure should be accounted for in estimation even if all of the levels are not used as stages in sample selection.

In the discussion later, assume a two-stage sample is selected. At the first stage, a sample s of n clusters is picked; within sample cluster i, a sample s_i of m_i units is selected. The population total is then naturally represented as the sum of three parts: (i) the total for the sample units in the sample clusters, (ii) the total of the nonsample units in the sample clusters, and (iii) the total for the units in the nonsample clusters:

$$T = \sum_{i \in s} \sum_{j \in s_i} Y_{ij} + \sum_{i \in s} \sum_{j \notin s_i} Y_{ij} + \sum_{i \notin s} \sum_{j=1}^{M_i} Y_{ij}$$

The optimal predictor under model (10) is

$$\hat{T} = \sum_{s} \sum_{s_i} Y_{ij} + \sum_{s} \sum (M_i - m_i) \left[\omega_i \overline{Y}_{si} + (1 - \omega_i) \hat{\mu} \right] + \sum_{i \notin s} M_i \hat{\mu}, \tag{11}$$

where $\omega_i = m_i \rho_i / [1 + (m_i - 1)\rho_i]$, $\overline{Y}_{si} = \sum_{s_i} Y_{ij} / m_i$, and $\hat{\mu} = \sum_s u_i \overline{Y}_{si}$ is a weighted average of the sample means with weights

$$u_i = \frac{m_i \big/ \left\{ \sigma_i^2 [1 + (m_i - 1)\rho_i] \right\}}{\sum\limits_{s} m_i \big/ \left\{ \sigma_i^2 [1 + (m_i - 1)\rho_i] \right\}}.$$

The estimator of the nonsample total within sample cluster i is a kind of composite estimator between the sample mean for that cluster $\overline{Y}_{si}$, and the overall weighted mean, $\hat{\mu}$.

A practical limitation to using (11) is that the intraclass correlations, ρ_i, and the variance components, σ_i^2, are unknown and must be estimated. More workable estimators are in the class, $\hat{T} = \sum_s g_i \overline{Y}_{si}$. An estimator in this class is prediction-unbiased if $\sum_s g_i = M$. One member of the class is $\hat{T}_p = (M/n) \sum_s \overline{Y}_{si}$, which is also the Horvitz–Thompson estimator under a plan where clusters are selected with probabilities proportional to size, that is, $nM_i \big/ \sum_{k=1}^{N} M_k$, and an equal probability sample is selected within each sample cluster. As in single-stage sampling, a probability sampling plan is not required to construct this estimator. The vital requirement for prediction-unbiasedness is that model (10) holds for both the sample and population.

Regression models can also be used in clustered populations. If $E_M(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$ and $\mathbf{W}$ is an arbitrary weight matrix, then a prediction-unbiased estimator of the total is

$$\hat{T} = \mathbf{1}_s' \mathbf{Y}_s + \mathbf{1}_r' X_r \hat{\boldsymbol{\beta}} \tag{12}$$

with $\hat{\boldsymbol{\beta}} = \mathbf{A}_s^{-1} \mathbf{X}_s' \mathbf{W}_s \mathbf{Y}_s$ if $\mathbf{A}_s = \mathbf{X}_s' \mathbf{W}_s \mathbf{X}_s$ is nonsingular. If an over-parameterized model is used, then we take $\boldsymbol{\beta}^o = \mathbf{G}\mathbf{X}_s' \mathbf{W}_s \mathbf{Y}_s$, where $\mathbf{G}$ is a generalized inverse of $\mathbf{A}_s$. The weight vector corresponding to $\hat{T}$, when a g-inverse is used, is

$$\mathbf{g}_s = \mathbf{W}_s \mathbf{X}_s \mathbf{G} \mathbf{X}_r' \mathbf{1}_r + \mathbf{1}_s. \tag{13}$$

As in previous sections, the x in the model can include qualitative variables, quantitative ones, and interactions.

The estimator in (12) is unbiased under a model with $E_M(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$, and can also be expressed as

$$\hat{T} = \mathbf{1}_s' \mathbf{Y}_s + \mathbf{T}_{xr}' \hat{\boldsymbol{\beta}}, \tag{14}$$

where $\mathbf{T}_{xr} = \sum_{i \in s} \sum_{j \notin s_i} \mathbf{x}_{ij} + \sum_{i \notin s} \sum_{j=1}^{M_i} \mathbf{x}_{ij}$ is the vector of nonsample totals of the auxiliary variables. If the population totals of the auxiliaries, $\mathbf{T}_x = \sum_{i=1}^{N} \sum_{j=1}^{M_i} \mathbf{x}_{ij}$, are known from the frame, a census, or some other source, then the nonsample total $\mathbf{T}_{xr}$ can be obtained by subtraction. An alternative model-unbiased predictor is $\hat{T} = \mathbf{T}_x' \hat{\boldsymbol{\beta}}$. In fact, if the model contains an intercept and $\mathbf{V} = \sigma^2 \mathbf{I}$, the BLUP reduces to $\hat{T} = \mathbf{T}_x' \hat{\boldsymbol{\beta}}$ due to Lemma 1 mentioned earlier.

The estimator in (14) is especially useful in populations where a full frame of the units to be surveyed is not available, but response variables are known to depend on auxiliaries. Even though the auxiliary values are not known for all individual units in the population, the x can be collected for the units that are in the sample. As long as the population totals, $\mathbf{T}_x$, are on hand, (14) or $\hat{T} = \mathbf{T}_x' \hat{\boldsymbol{\beta}}$ can be constructed. An example of this is the model:

$$E_M(Y_{ij}) = \mu_c \tag{15}$$

when unit ij is in group $c (c = 1, \ldots, C)$.

Often group membership of individual units is unknown at the time of sampling and can only be determined when the data are collected. A group can cut across clusters and, in design-based sampling, is usually called a *poststratum* if the membership of a unit in

a group is determined *after* the sample is selected. A particular example of this would be a two-way model with interaction:

$$E_M(Y_{ij}) = \mu + \alpha_k + \beta_\ell + (\alpha\beta)_{k\ell},$$

where unit (ij) is in level k of the first factor (say, age group) and level ℓ of the second factor (say, sex). Men and women of different ages may, for example, have different average incomes in a population of households. A sample of clusters is selected and households are listed within the sample clusters. The age and sex of household members, and possibly other important explanatory variables, are collected only for individuals who are in the sample. However, the total numbers in the population of males and females in different age groups may be known from a census or from demographic projections.

Under the poststratification model (15), the estimator $\hat{T} = \mathbf{1}_s'\mathbf{Y}_s + \mathbf{T}_{xr}'\hat{\boldsymbol{\beta}}$ reduces to

$$\hat{T}_{\text{PS}} = \sum_{c=1}^{C} M_c \overline{Y}_{sc},$$

where M_c is the number of units in the population in group c, $\overline{Y}_{sc} = \sum_{i\in s} \sum_{j\in s_{ic}} Y_{ij}/m_c$, m_c is the number of sample units in group c across all sample clusters, and s_{ic} is the set of sample units in sample cluster i that are also in group c. Note that $\hat{T}_{\text{PS}}$ is, in general, different from the poststratified estimator that flows from the model-assisted general regression estimator (see Chapter 25).

Variance estimation in samples from clustered populations is, as might be expected, more complicated than in unclustered populations. First, consider the simple model

$$\begin{aligned} E_M(Y_{ij}) &= \mu \\ \text{cov}_M(Y_{ij}, Y_{kl}) &= \begin{cases} \sigma^2 & i = k,\ j = l \\ \sigma^2\rho & i = k,\ j \neq l \\ 0 & \text{otherwise} \end{cases} \end{aligned} \tag{16}$$

which is a special case of (10). Consider the general class of estimators, $\hat{T} = \sum_s g_i \overline{Y}_{si}$. The estimation error is $\hat{T} - T = \sum_s g_i \overline{Y}_{si} - \left(\sum_s M_i \overline{Y}_i + \sum_r M_i \overline{Y}_i\right)$. A general form of the error variance is

$$\text{var}_M(\hat{T} - T) = B_1 - 2B_2 + B_3, \tag{17}$$

where $B_1 = \text{var}_M(\hat{T}) = \sum_{i\in s} g_i^2 \text{var}_M(\overline{Y}_{si})$, $B_2 = \sum_{i\in s} g_i M_i \text{cov}_M(\overline{Y}_{si}, \overline{Y}_i)$, and $B_3 = \sum_{i=1}^{N} M_i^2 \text{var}_M(\overline{Y}_i)$. If the first-stage sampling fraction is negligible, and certain other population and sample quantities are bounded, the B_1 term dominates the variance (see Valliant et al., 2000, Theorem 9.1.1).

The formulation in (17) is different than the one usually found in design-based texts. For example, take the case of a two-stage sample of clusters and elements with simple random sampling without replacement used at both stages. The π-estimator of the total is $\hat{t}_\pi = (N/n)\sum_s M_i \overline{Y}_{si}$, which has the form $\hat{T} = \sum_s g_i \overline{Y}_{si}$ with $g_i = NM_i/n$. The

design-variance of $\hat{t}_\pi$, that is, the variance in repeated sampling (see Särndal et al., 1992, Chapter 4) is

$$\text{var}_\pi(\hat{t}_\pi) = N^2 \frac{1 - n/N}{n} S_1^2 + \frac{N}{n} \sum_{U_1} M_i^2 \frac{1 - m_i/M_i}{m_i} S_{2i}^2 \tag{18}$$

with

$$S_1^2 = \sum_{i=1}^{N} \left(M_i \overline{Y}_{Ui} - N^{-1} \sum_{i=1}^{N} M_i \overline{Y}_{Ui} \right)^2 \Big/ (N-1),$$

$$\overline{Y}_{Ui} = \sum_{j=1}^{M_i} Y_{ij} \big/ M_i, \text{ and}$$

$$S_{2i}^2 = \sum_{j=1}^{M_i} \left(Y_{ij} - \overline{Y}_{Ui} \right)^2 \Big/ (M_i - 1).$$

If the cluster sizes and means, M_i and $\overline{Y}_{Ui}$, are bounded as the population and sample sizes of clusters, N and n, become large and the sampling fraction of clusters is small, then the first term in (18) is dominant with order N^2/n. The first term is a variance between cluster totals. In (17) the dominant term, $\sum_s g_i^2 \text{var}_M(\overline{Y}_{si}) = (N/n)^2 \sum_s M_i^2 \text{var}_M(\overline{Y}_{si})$ is a weighted combination of within-cluster variances but also has order N^2/n. However, as noted subsequently, variance estimators that flow from design-based and prediction theories are similar in this example.

A robust estimator of the dominant term in the error variance can be constructed under the general model

$$\begin{aligned} &E_M(Y_{ij}) = \mu \quad i = 1, \ldots, N; \quad j = 1, \ldots, M_i \\ &\text{cov}_M(Y_{ij}, Y_{k\ell}) = 0, \quad i \neq k. \end{aligned} \tag{19}$$

This model says that units in different clusters are uncorrelated. But, it is less specific than (16), because it imposes no further constraints on the covariance structure within clusters. Each unit within a cluster may have a different variance, for example, and different pairs of units may have different correlations. Define the residual for cluster i as $r_i = \overline{Y}_{si} - \hat{\mu}$ where $\hat{\mu} = \hat{T}/M$. When $E_M(Y_{ij}) = \mu$, $E_M(r_i) = 0$ and

$$E_M(r_i^2) = \text{var}_M\left(\overline{Y}_{si}\right) \left(1 - \frac{2g_i}{M}\right) + \frac{1}{M^2} \sum_{i' \in s} g_{i'}^2 \text{var}_M\left(\overline{Y}_{si'}\right)$$

for $i = 1, \ldots, n$.

When $g_i = O(N/n)$ and $M_i = O(1)$, we have $g_i/M = O(n^{-1})$. Thus, r_i^2 is an approximately unbiased estimator of $\text{var}_M(\overline{Y}_{si})$ under model (19). A sandwich variance estimator is then simply $v_R = \sum_s g_i^2 r_i^2$. This estimator is approximately unbiased and consistent under either the working model (10) or the more general model (19) when the sample of clusters is large and their sampling fraction is small. It is possible to estimate the two less important terms in (17) but the component estimators are unbiased only under more restrictive working models than (19).

Returning to the case of a two-stage cluster sample with simple random sampling at each stage, it is instructive to compare v_R to a variance estimator often used in design-based practice. As noted earlier, the π-estimator is $\hat{t}_\pi = (N/n)\sum_s M_i \overline{Y}_{si}$. The *ultimate cluster* variance estimator (see Chapter 1) in this example can be written as

$$v_\pi(\hat{t}_\pi) = \left(\frac{N}{n}\right)^2 \sum_s M_i^2 \left(\overline{Y}_{si} - \frac{\hat{t}_\pi}{NM_i}\right)^2.$$

Strictly speaking this estimator is appropriate when the sample of clusters is selected with replacement but is often used for without replacement sampling. In this case, we have

$$v_R = \left(\frac{N}{n}\right)^2 \sum_s M_i^2 \left(\overline{Y}_{si} - \frac{\hat{t}_\pi}{N\overline{M}}\right)^2$$

with $\overline{M} = M/N$. If $M_i = \overline{M}$, then $v_\pi = v_R$. Such correspondences can frequently be found in special cases. But, because prediction theory conditions on the obtained sample, unlike design-based theory, variance estimators resulting from the two theories are often quite different (e.g., see Royall, 1986; Royall and Cumberland, 1978, 1981a)

Variance estimators can also be developed when $E_M(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$ and the predictor $\hat{T} = \mathbf{1}_s'\mathbf{Y}_s + \mathbf{1}_r'\mathbf{X}_r\hat{\boldsymbol{\beta}} = \mathbf{g}_s'\mathbf{Y}_s$ is used. The weight vector, $\mathbf{g}_s$, is given in (13). As in the case of estimation under the common mean model, we can construct a simple, sandwich estimator that is consistent under a reasonably general variance specification. Analogous to (19), consider the model:

$$\begin{aligned} &E_M\left(Y_{ij}\right) = \mathbf{x}_{ij}'\boldsymbol{\beta} \qquad i = 1, \ldots, N; \qquad j = 1, \ldots, M_i \\ &\operatorname{cov}_M\left(Y_{ij}, Y_{k\ell}\right) = 0, \quad i \neq k. \end{aligned} \tag{20}$$

This model assumes that the regression specification, $E_M(Y_{ij})$, used to construct $\hat{T}$ is correct. Units in different clusters are assumed to be uncorrelated, but the variance–covariance structure within each cluster is arbitrary. The estimation error of $\hat{T}$ is

$$\hat{T} - T = \sum_s \mathbf{g}_i'\mathbf{Y}_{si} - \left(\sum_s M_i\overline{Y}_i + \sum_r M_i\overline{Y}_i\right)$$

where $\mathbf{g}_i = (g_{i1}, \ldots, g_{im_i})'$ is the part of the weight vector for the sample cluster i and $\mathbf{Y}_{si} = (Y_{i1}, \ldots, Y_{im_i})'$ is the data for the sample units from sample cluster i. The error variance is

$$\operatorname{var}_M(\hat{T} - T) = B_1 - 2B_2 + B_3 \tag{21}$$

where $B_1 = \sum_s \mathbf{g}_i'\operatorname{var}_M(\mathbf{Y}_{si})\mathbf{g}_i$, $B_2 = \sum_s \mathbf{g}_i'\operatorname{cov}_M(\mathbf{Y}_{si}, \mathbf{Y}_i)\mathbf{1}_{M_i}$ with $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{iM_i})'$, and $B_3 = \sum_{i=1}^N M_i^2\operatorname{var}_M(\overline{Y}_i)$. As in the common mean model, the B_1 term dominates under some reasonable conditions.

To construct a robust estimator of this dominant term, define the residual for sample unit ij to be $r_{ij} = Y_{ij} - \mathbf{x}_{ij}'\hat{\boldsymbol{\beta}}$ where the estimator of slope is either $\hat{\boldsymbol{\beta}} = \mathbf{A}_s^{-1}\mathbf{X}_s'\mathbf{W}_s\mathbf{Y}_s$ or $\boldsymbol{\beta}^o = \mathbf{G}\mathbf{X}_s'\mathbf{W}_s\mathbf{Y}_s$. The vector of residuals for sample cluster i is $\mathbf{r}_i = (r_{i1}, \ldots, r_{im_i})'$.

Define $\mathbf{g}_i = (g_{i1}, \ldots, g_{im_i})'$, which is the part of $\mathbf{g}_s$ associated with sample cluster i. The sandwich variance estimator

$$v_R(\hat{T}) = \sum_s \left(\mathbf{g}_i'\mathbf{r}_i\right)^2 = \sum_s \mathbf{g}_i'\left(\mathbf{r}_i\mathbf{r}_i'\right)\mathbf{g}_i.$$

is approximately unbiased under the general model (20).

There are also adjusted versions of the sandwich variance estimator. The adjustments involve the hat matrix $\mathbf{H} = \mathbf{X}_s\mathbf{G}\mathbf{X}_s'\mathbf{W}_s$ and are covered in Valliant et al. (2000, Chapter 9). The jackknife variance estimator, where one cluster is deleted at a time, also involves parts of the hat matrix. One version of the jackknife estimator of the variance of $\hat{T}_r$ is

$$v_J(\hat{T}_r) = \frac{n-1}{n}\sum_{i\in s}(\hat{T}_{r(i)} - \hat{T}_{r(\bullet)})^2$$

where $\hat{T}_{r(i)}$ is the estimate of T_r found after omitting cluster i and $\hat{T}_{r(\bullet)} = n^{-1}\sum_s \hat{T}_{r(i)}$. Rather than mechanically deleting a cluster, computing $\hat{T}_{r(i)}$, and cycling through all sample clusters, the following alternative computational form can be used:

$$v_J(\hat{T}_r) = \frac{n-1}{n}\left\{\sum_s \left(\mathbf{a}_i'\mathbf{P}_i^{-1}\mathbf{r}_i\right)^2 - n^{-1}\left[\sum_s \mathbf{a}_i'\mathbf{P}_i^{-1}\mathbf{r}_i\right]^2\right\} \tag{22}$$

where $\mathbf{a}_i$ is the part of $\mathbf{a}_s = \mathbf{W}_s\mathbf{X}_s\mathbf{G}\mathbf{X}_r'\mathbf{1}_r$ associated with cluster i, $\mathbf{P}_i = \mathbf{I}_{m_i} - \mathbf{H}_{ii}$, $\mathbf{I}_{m_i}$ is the $m_i \times m_i$ identity matrix, $\mathbf{H}_{ii} = \mathbf{X}_{si}\mathbf{G}\mathbf{X}_{si}'\mathbf{W}_{si}$, $\mathbf{X}_{si}$ is the $m_i \times p$ matrix of auxiliaries for the sample units in sample cluster i, and $\mathbf{W}_{si}$ is the $m_i \times m_i$ part of the $\mathbf{W}$ matrix for sample cluster i. The jackknife (22) is a consistent estimator of the dominant term in the variance (21).

8. Estimation under nonlinear models

The preceding sections of this chapter have described the estimation of totals and the variances of the estimators assuming that a linear working model is reasonable. An obvious situation where a nonlinear model may be better is when Y is an indicator for whether a unit has a characteristic or not. For example, we might want to estimate the total number of persons with a chronic health condition like osteoarthritis. For such 0–1 Y variables, logistic or some other type of nonlinear model is usually a better fit than a linear model.

Standard survey practice is to estimate the total of a binary variable with a linear estimator of the form $\hat{T} = \sum_s w_i Y_i$ as described in previous sections. This type of estimator can be prediction-unbiased if a linear model holds, but can be seriously biased if, say, the correct underlying model is logistic. One problem with using a linear model for a binary variable in the presence of auxiliaries is that the predicted value for a given unit does not have to be confined to [0,1], as a probability should be.

Estimators of totals can be developed under nonlinear models that are very similar in appearance to the BLUP introduced earlier (Valliant, 1985). Related model-assisted work is found in Lehtonen and Veijanen (1998). Suppose that the population vector of target values is $\mathbf{Y} = (Y_1, \ldots, Y_N)'$ where each Y can be continuous or binary. Denote the vector of expected values of $\mathbf{Y}$ by $\mathbf{f}(\boldsymbol{\beta}) = (f(\mathbf{x}_1;\boldsymbol{\beta}), \ldots, f(\mathbf{x}_N;\boldsymbol{\beta}))$ where f is

a nonlinear function of the components of the $p \times 1$ parameter vector $\boldsymbol{\beta}$ and $\mathbf{x}_i$ is a p-vector of auxiliaries associated with unit i. The $N \times N$ covariance matrix for $\mathbf{Y}$ is $\mathbf{V}$. As in earlier sections, after a sample is selected, the population can be split into the sample s and the nonsample r. The full specification of the model for $\mathbf{Y}$ is

$$\begin{aligned} E_M(\mathbf{Y}) &= \mathbf{f}(\boldsymbol{\beta}) = [\mathbf{f}_s(\boldsymbol{\beta})', \mathbf{f}_r(\boldsymbol{\beta})']' \\ \text{var}_M(\mathbf{Y}) &= \mathbf{V} = \begin{bmatrix} \mathbf{V}_{ss} & \mathbf{V}_{sr} \\ \mathbf{V}_{rs} & \mathbf{V}_{rr} \end{bmatrix} \end{aligned} \tag{23}$$

where $\mathbf{f}(\boldsymbol{\beta})$ and $\mathbf{V}$ are decomposed in the obvious way. We also need the vectors and matrices of first partial derivatives defined by

$$\begin{aligned} \mathbf{F}_i(\boldsymbol{\beta}) &= \left[\frac{\partial f(\mathbf{x}_i; \boldsymbol{\beta})}{\partial \beta_1}, \ldots, \frac{\partial f(\mathbf{x}_i; \boldsymbol{\beta})}{\partial \beta_p} \right]' \text{ for } i = 1, \ldots N, \text{ and} \\ \mathbf{F}(\boldsymbol{\beta}) &= [\mathbf{F}_1(\boldsymbol{\beta}), \ldots, \mathbf{F}_N(\boldsymbol{\beta})]' \\ &= [\mathbf{F}_s(\boldsymbol{\beta})', \mathbf{F}_r(\boldsymbol{\beta})']' \end{aligned}$$

where $\mathbf{F}_s(\boldsymbol{\beta})$ is the $n \times p$ matrix of first partial derivatives for the sample units and $\mathbf{F}_r(\boldsymbol{\beta})$ is the $(N-n) \times p$ matrix of partials for the nonsample units. In subsequent formulas, the argument $\boldsymbol{\beta}$ will sometimes be suppressed in $\mathbf{F}_i(\boldsymbol{\beta})$, $\mathbf{F}_s(\boldsymbol{\beta})$, and $\mathbf{F}_r(\boldsymbol{\beta})$ for compactness of notation.

If $\boldsymbol{\beta}$ were known, the BLUP of T is simply the sample sum of the Y plus the BLUP of the nonsample sum as stated in the following theorem.

THEOREM 1. *Under model (23) with $\boldsymbol{\beta}$ known, among linear estimators of the form $\hat{T} = \mathbf{g}_s'\mathbf{Y}_s$ satisfying $E_M(\hat{T} - T) = 0$, the error variance $E_M(\hat{T} - T)^2$ is minimized by*

$$\hat{T}^* = \sum_s Y_i + \mathbf{1}_r' \left[\mathbf{f}_r(\boldsymbol{\beta}) + \mathbf{V}_{rs}\mathbf{V}_{ss}^{-1} (\mathbf{Y}_s - \mathbf{f}_s(\boldsymbol{\beta})) \right]$$

where $\mathbf{1}_r$ is an $(N-n)$-vector of all 1.

When $\boldsymbol{\beta}$ is unknown, an estimator must be used. The standard estimator of $\boldsymbol{\beta}$ in a nonlinear regression problem is obtained by generalized least squares (GLS). The GLS estimator is the value $\hat{\boldsymbol{\beta}}$ that minimizes the sum of squares

$$[\mathbf{Y}_s - \mathbf{f}_s(\boldsymbol{\beta})]'\mathbf{V}_{ss}^{-1}[\mathbf{Y}_s - \mathbf{f}_s(\boldsymbol{\beta})]. \tag{24}$$

Differentiating (24) with respect to $\boldsymbol{\beta}$ and setting the result to 0 leads to this set of p estimating equations in the p unknowns $\beta_1, \ldots, \beta_p$:

$$\mathbf{F}_s(\boldsymbol{\beta})'\mathbf{V}_{ss}^{-1}[\mathbf{Y}_s - \mathbf{f}_s(\boldsymbol{\beta})] = \mathbf{0}. \tag{25}$$

These must be solved iteratively to find an estimator of $\boldsymbol{\beta}$. If Y is binary, the maximum likelihood estimator (MLE) is also found by solving a system like (25).

When estimating the population total, the obvious candidate comes from substituting $\hat{\boldsymbol{\beta}}$ into the estimator from Theorem 3 giving

$$\hat{T} = \sum_s Y_i + \mathbf{1}_r' \left\{ \mathbf{f}_r(\hat{\boldsymbol{\beta}}) + \mathbf{V}_{rs}\mathbf{V}_{ss}^{-1}[\mathbf{Y}_s - \mathbf{f}_s(\hat{\boldsymbol{\beta}})] \right\}. \tag{26}$$

Notice that if $\mathbf{f}(\boldsymbol{\beta}) = \mathbf{X}\boldsymbol{\beta}$, then (26) reduces to the BLUP under the general linear model in Theorem 1. An approximation to the error variance of $\hat{T}$ is:

$$\begin{aligned}\mathrm{var}_M(\hat{T} - T) \cong\ & \mathbf{1}_r' \left(\mathbf{F}_r - \mathbf{V}_{rs}\mathbf{V}_{ss}^{-1}\mathbf{F}_s\right) \left(\mathbf{F}_s'\mathbf{V}_{ss}^{-1}\mathbf{F}_s\right)^{-1} \left(\mathbf{F}_r - \mathbf{V}_{rs}\mathbf{V}_{ss}^{-1}\mathbf{F}_s\right)' \mathbf{1}_r \\ & + \mathbf{1}_r' \left(\mathbf{V}_{rr} - \mathbf{V}_{rs}\mathbf{V}_{ss}^{-1}\mathbf{V}_{sr}\right) \mathbf{1}_r.\end{aligned}$$

Some details of the approximation are given in Valliant (1985). When the observations are all independent, then the error variance approximation simplifies to

$$\mathrm{var}_M(\hat{T} - T) \cong \mathbf{1}_r'\mathbf{F}_r \left(\mathbf{F}_s'\mathbf{V}_{ss}^{-1}\mathbf{F}_s\right)^{-1} \mathbf{F}_r'\mathbf{1}_r + \mathbf{1}_r'\mathbf{V}_{rr}\mathbf{1}_r. \tag{27}$$

In the particular instance of independent Bernoulli random variables, the model is

$$\begin{aligned} E_M(Y_i) &= f(\mathbf{x}_i;\boldsymbol{\beta}) \\ \mathrm{var}_M(Y_i) &= f(\mathbf{x}_i;\boldsymbol{\beta})[1 - f(\mathbf{x}_i;\boldsymbol{\beta})], \end{aligned} \tag{28}$$

where $0 \leq f(\mathbf{x}_i;\boldsymbol{\beta}) \leq 1$. The covariance matrix under this model is automatically unknown because we assume that the parameter $\boldsymbol{\beta}$ is unknown. When the Y are independent, the estimator of the total reduces to

$$\hat{T} = \sum_s Y_i + \mathbf{1}_r'\mathbf{f}_r(\hat{\boldsymbol{\beta}}).$$

Under model (28), the MLE of $\boldsymbol{\beta}$ can be calculated using standard methods like the Newton–Raphson algorithm. A variance estimator is found by substituting the estimator $\hat{\boldsymbol{\beta}}$ into (27) to obtain

$$v_{(\hat{T}-T)} = \mathbf{1}_r'\mathbf{F}_r(\hat{\boldsymbol{\beta}}) \left[\mathbf{F}_s(\hat{\boldsymbol{\beta}})'\mathbf{V}_{ss}^{-1}\mathbf{F}_s(\hat{\boldsymbol{\beta}})\right]^{-1} \mathbf{F}_r(\hat{\boldsymbol{\beta}})'\mathbf{1}_r + \mathbf{1}_r'\hat{\mathbf{V}}_{rr}\mathbf{1}_r,$$

where $\hat{\mathbf{V}}_{ss} = \mathrm{diag}[\mathbf{f}_s(\hat{\boldsymbol{\beta}})(1 - \mathbf{f}_s(\hat{\boldsymbol{\beta}}))]$ and $\hat{\mathbf{V}}_{rr} = \mathrm{diag}[\mathbf{f}_r(\hat{\boldsymbol{\beta}})(1 - \mathbf{f}_r(\hat{\boldsymbol{\beta}}))]$.

Use of nonlinear models in estimating totals is seldom if ever used because of the inconvenient form of $\hat{T}$. For example, If the superpopulation model were logistic, that is, $f(\mathbf{x}_i;\boldsymbol{\beta}) = \exp(\mathbf{x}_i'\boldsymbol{\beta})/[1 + \exp(\mathbf{x}_i'\boldsymbol{\beta})]$, then the estimated total would be

$$\begin{aligned}\hat{T} &= \sum_s Y_i + \mathbf{1}_r'\mathbf{f}_r(\hat{\boldsymbol{\beta}}) \\ &= \sum_s Y_i + \sum_r \frac{\exp\left(\mathbf{x}_i'\hat{\boldsymbol{\beta}}\right)}{\left[1 + \exp\left(\mathbf{x}_i'\hat{\boldsymbol{\beta}}\right)\right]}\end{aligned}$$

This cannot be expressed as a weighted summation of the sample Y and, thus, does not dovetail with standard survey practice where a single weight is used for all types of estimates. Nevertheless, estimators based on nonlinear models may have advantages in particular applications, especially for totals of rare characteristics where there is some danger of a linear estimator having implied predictions that are outside of [0,1].

Sample Surveys: Inference and Analysis, Vol. 29B
ISSN: 0169-7161
© 2009 Elsevier B.V. All rights reserved
DOI: 10.1016/S0169-7161(09)00224-7

Design- and Model-Based Inference for Model Parameters

David A. Binder and Georgia Roberts

1. Introduction and scope

When survey data are being analyzed, it is common to formulate the questions of interest as relationships among parameters of a statistical model. We refer to studies that make inferences on model parameters as analytic studies, in contrast to descriptive studies where the relationships refer only to finite population characteristics; see Kalton (1983). Commonly used statistical models are linear regression models, logistic regression models, generalized linear models, and hierarchical linear models. When the questions of interest are based on parameters of a statistical model, populations satisfying these models are conceptually infinite. It is often assumed that the values of the variables in the finite population from which the observed survey sampled units were selected are outcomes resulting from sampling from this infinite population. The model may or may not contain variables related to the survey design.

In this chapter, we use a frequency-based framework to discuss the issues associated with making inferences about model parameters from survey data that have been obtained from a probability-sampling scheme. There are many similarities between the randomization assumptions used in this chapter and those in Chapter 39, but in Chapter 39 there is more attention paid to the sample likelihood function. In this chapter, we do not discuss statistical models used for estimating finite population quantities, such as is the case in small area estimation (see Chapters 31 and 32), or model-based prediction of finite population quantities, as discussed in Chapter 23.

In Section 2, we explain what is meant by the survey population, by the populations targeted by the survey producers and by the populations of interest to the researcher who is fitting a statistical model to the survey data. In Section 3, we focus on the probability randomization distribution of the observations as a basis for frequency-based statistical inference. This distribution is based on the statistical model that generated the values of the variables of interest in a finite population and the sampling plan used to select the sample from that population. In Section 4, we discuss the properties of model- and design-based estimators in the context of a model-design-based randomization framework. Situations where the design-based approach can be problematic are discussed in Section 5, including cases of large sampling fractions, small sample sizes, estimating

parameters for models that include random effects, population-based case-control studies, and making appropriate design-based inferences in event history analysis models. In Section 6, we briefly discuss the estimation of design-based variances of estimates of model parameters. The particular problem of integrating data from more than one survey is given in Section 7. In Section 8, we provide some final remarks.

The more technical details are confined to Section 4 and part of Section 5, so that readers who wish to read the less technical discussions can skip over these parts. Table 1 in Section 4 summarizes many of the main concepts described in that section.

2. Survey populations and target populations

In sample surveys, information is collected from a sample of units from a finite population. It is common for the sampling plan to be complex, which we define as any sampling plan where the units are selected using a design that is not simple random sampling. To select the units for the sample using a complex probability-based sampling plan, a sampling frame is constructed, and a probability sample is taken using this frame to lead to the ultimate units observed in the survey. (See Part 1 of this Handbook for more details.)

A *descriptive* study is one where the quantities of interest are characteristics of a finite population, such as population totals, means, proportions, or other ratios. On the other hand, in an *analytical* study, the quantities of interest are related to the parameters of a statistical model, such as the coefficients of a regression model.

When a survey is conducted, the survey producer[1] targets a particular population of inference (or a particular set of populations of inference). It is important to distinguish between two types of populations: the survey population and the target populations. The survey population comprises all the units that are eligible for selection in the survey sample.[2] The survey population is always finite. However, the survey population and the finite population being targeted by the survey producer may not necessarily coincide. The units observed in the survey depend on the sampling frame, which may suffer from imperfections due to coverage or classification errors. Also, for operational reasons, the survey producer may exclude certain population units from being eligible for inclusion in the sample. On the other hand, for a statistical study, either descriptive or analytic, the researcher's target populations are the populations about which he wishes to draw conclusions and are often different from either the survey population or the survey producer's target population. The target population that is appropriate for a particular analysis may depend on the quantities being estimated, since more than one target population can be studied from the same survey. The quantities being estimated are based on the purpose of the study and on how the estimates will shed light on the questions of analytic interest.

[1] In this chapter, we use the term survey producer to refer to the agencies or organizations that select the sample, collect the survey data, process the data, and produce the files to be used for the production of survey estimates and for making inferences about the finite populations targeted by the survey.

[2] Note that the definition of the survey population might depend on the particular survey variables being included in the study, since some survey variables may not be collected for certain subpopulations. As an example, we might not ask some labor-related questions from those who are not currently employed.

In this chapter, we consider analytic studies where a researcher is interested in making inferences on the parameters of a statistical model. It is presumed that the realizations of the random variables generated by such a model have given rise to the values of the characteristics of interest in the finite population from which the sample was selected. Graubard and Korn (2002) gave several references where the primary interest is in the parameters of the model. Also, examples and further discussion may be found in monographs by Chambers and Skinner (2003), Korn and Graubard (1999), and Lehtonen and Pahkinen (2004). If the finite population could be completely observed, then an estimate of the model parameters of interest, based on all the values of the finite population, would be available. These estimates are *finite population quantities that are associated with parameters of the model.* Generally, the choice of estimator is based on its statistical properties under the assumed model.

We denote the parameters of the model by $\boldsymbol{\theta}$ and the finite population quantities associated with these parameters by $\boldsymbol{\theta}_N$. As an example, if the statistical model assumed to have generated values of the characteristics of interest in the N units in a population is a standard linear regression model, $\mathbf{y}_N = \mathbf{X}_N\boldsymbol{\theta} + \boldsymbol{\varepsilon}_N$, where $\boldsymbol{\varepsilon}_N$ is a vector of independent and identically distributed normal errors, the finite population quantities associated with the regression coefficients could be the ordinary least-squares estimator of the regression coefficients, based on the complete finite population values; that is, $\boldsymbol{\theta}_N = (\mathbf{X}_N'\mathbf{X}_N)^{-1}\mathbf{X}_N'\mathbf{y}_N$. These finite population quantities are descriptive characteristics of the population. Note that even if the model is only approximately true, the researcher may consider these finite population quantities to be useful descriptive measures for the finite population.

We provide some examples of possible studies that illustrate the concepts of survey population and target population and the differences between descriptive and analytic studies:

(a) A descriptive study where the survey producer's target population is either the survey population or differs from the survey population due to frame overcoverage (such as duplication of persons on the frame, persons on the frame but not resident in the geographic area covered by the survey, etc.), or due to frame undercoverage (such as those units that were deliberately excluded from selection from the frame and those units that were missing from the frame).

Example 2.1. Suppose that we are interested in studying (i) the average expenses per acre for Canadian farmers who used organic farming techniques for vegetables in 2002 and (ii) whether the average of expenses per acre in the province of Ontario differs from that in the province of Quebec. Our data source is a cross-sectional survey based on a frame consisting of farmers in Canada operating in 2001. The survey producer's target population was actually farmers in Canada operating in 2002. Survey questions included information about organic farming techniques used in 2002. In this case, our analysis is primarily descriptive and the population of interest consists of those farmers who engaged in organic farming in 2002 among all the farmers in the finite population targeted by the survey producer. We see that some differences between the survey population and the target population are due to frame imperfections.

Example 2.2. Suppose that we are interested in studying residents of the United States living in households, aged 25–40 years, who were overweight in 1993. We would like to know (i) what percentage of these persons were still overweight in 2001 and (ii) whether males and females differ with respect to these characteristics. To study these questions, we use a longitudinal survey where the finite population targeted by the survey consists of residents of the United States living in households in 1993 and where data pertaining to years 1993 and 2001 are collected from the sampled individuals (such as by interviewing the people every 2 years beginning in 1993). In this case, we again have a descriptive analysis, but here the population of interest consists of the finite population units targeted by the survey in 1993, who were overweight in 1993. Since we have a longitudinal survey, we can observe the overweight status both in 1993 and in 2001 for units in the observed sample.

(b) A descriptive study where the researcher's target finite population is larger than the one targeted by the survey producer, but its characteristics can be represented by the characteristics of the finite population targeted by the survey producer.

Example 2.3. Suppose we believe, based on clinical or other studies, that the prevalence of chronic back pain in Canada was constant throughout the 1990s, and we would like to know this prevalence rate. We use as our source of data a cross-sectional survey of the 1993 Canadian population, where a question was asked about the existence of chronic back pain in that year. In this case, we are interested in estimating the prevalence rate, but our population of interest is wider than the finite population targeted by the survey, since we are assuming that the 1993 prevalence rate applied throughout the 1990s. Even though we are estimating a quantity for a target population that is larger than that targeted by the survey producer, we say that the study is descriptive because we are estimating a quantity that refers to a finite target population.

(c) An analytic study where the researcher's target population is infinite and the values of the characteristics of the units in the finite population from which the sample is selected are considered to be outcomes of random variables from a statistical model. This is the case we focus on in this chapter.

Example 2.4. Suppose we want to know whether obesity, age, and gender are important risk factors for a senior needing to leave his/her home to go to reside in a long-term care facility. In particular, we are interested in the impact of age, after controlling for the other variables. Our data source is a longitudinal survey of seniors in the Canadian province of Ontario, where the sample was chosen from seniors living in households in 1992 and sampled individuals were followed for 6 years. In this case, our objectives are primarily analytical. We would like to study statistical models that can be used to explain the relationships among the variables of interest. Our target population is wider than the particular finite population targeted by the survey; rather, it is the conceptually infinite population, represented by the statistical model, from which the values of the characteristics of interest in the finite population were generated.

2.1. Units of analysis

For either a descriptive or an analytic study, the structure of the units of analysis is not necessarily the same as the structure of the units selected by the sampling plan. This would be the case when the units of analysis are derived from the units that were actually sampled. As an example, the survey population units could be persons, whereas the units of analysis could be households. Another example would be where the units observed in the sample are persons in a longitudinal employment survey, and the units of analysis are job spells held by the people surveyed.

The relationship between the units of a population and the units of analysis is important. The researcher must be aware of the differences between the two types of units. Survey weights for the units to be analyzed may not be provided and would need to be constructed; for example, this would be the situation when the units of analysis are households, the units observed in the sample are persons, and the data files contain just person-level weights. In the case of a longitudinal study, when the units of analysis are spells, such as spells of unemployment, there may be multiple units of analysis for the same person, so that the estimation of variances of survey estimates must account for the fact that some units of analysis refer to the same individual and may be correlated within the individual.

2.2. Weighting and estimation

In Chapters 8 and 25, weighting and estimation approaches for estimating finite population characteristics from cross-sectional surveys are discussed in detail. Here, we review some of the main approaches for making inferences about survey characteristics in the finite population targeted by the survey producer.

In finite population sampling, the finite survey population consists of N units. We denote by $\mathbf{Y}_i = (Y_{1i}, \ldots, Y_{qi})'$, for $i = 1, \ldots, N$, the values for q survey variables measured on the ith unit. (It may be possible for the number of variables to vary among units, but for the sake of simplicity of notation, we consider here the case where the same number of survey variables of interest is associated with each population unit.) A sample is selected from the finite population and we suppose that n units are observed in the sample. We denote survey observations by $\mathbf{y}_1, \ldots, \mathbf{y}_n$, where each of the $\mathbf{y}_i$s corresponds to one of the $\mathbf{Y}_j$s in the population. For the ith sampled unit, $\mathbf{y}_i = (y_{1i}, \ldots, y_{qi})'$, we have an associated survey weight w_i. (In some survey files, there may be more than one weight variable. Each weight variable could correspond to a different population targeted by the survey producer, or different weight variables could be used with different subsamples to represent the same population, such as is the case for weight variables associated with different cycles from a longitudinal survey with different nonrespondents in each cycle.) A survey-weighted estimate for the population total $Y_1 = \sum_{i=1}^{N} Y_{1i}$, say, is $\hat{Y}_1 = \sum_{i=1}^{n} w_i y_{1i}$.[3] Similarly, a survey-weighted estimate for a ratio $R = Y_1/Y_2$ is $\hat{R} = \hat{Y}_1/\hat{Y}_2 = \sum_{i=1}^{n} w_i y_{1i} / \sum_{i=1}^{n} w_i y_{2i}$. For example, a population

[3] An alternative way to express this is to let $I_i = 1$, for the n observed units in the finite population, and $I_i = 0$ for the N - n units of the population not observed, so that $\hat{Y}_1$ is expressed as $\sum_{i=1}^{N} I_i w_i Y_{1i}$, where the w_i's are associated with the finite population units, rather than the sampled units. We use this formulation throughout much of this chapter.

mean or proportion $\overline{Y}_1 = Y_1/N$ may be estimated using the survey-weighted estimator $\hat{\overline{Y}}_1 = \hat{Y}_1/\hat{N} = \sum_{i=1}^{n} w_i y_{1i} / \sum_{i=1}^{n} w_i$.

Normally, each survey weight variable is constructed in such a way that the survey-weighted estimator of a finite population total is approximately unbiased under the probability randomization distribution induced by the sampling plan. To account for differences between the survey population and the finite population targeted by the producer, it is common to adjust the survey weights, for example, by using poststratification, a special case of calibration (see Chapter 25). Weight adjustment is also one of the means of accounting for nonresponse (see Chapter 8). Adjustment of the survey weights can improve the accuracy of the weighted estimates for finite population characteristics, such as population means, totals, and ratios. The goal is to reduce the *sampling errors*, which are the differences between the estimates based on an observed sample and the true values for characteristics in the finite population targeted by the survey producer. In design-based estimation, the properties of the sampling errors are assessed with respect to the sampling distribution resulting from the probabilities used in the sampling plan and the assumed nonresponse mechanism.

3. Statistical inferences

In a frequency-based framework for statistical inferences, there is interest not only in what is observed but also in what could have been observed had different samples been selected. The properties of estimators are studied in terms of expectations, variances, or other measures related to the distribution of the random variables generating the sample observations, as described below. We use this distribution to perform tests of statistical hypotheses and to construct confidence intervals.

Of interest is the distribution of estimates under hypothetical random repetitions. This distribution depends on whether or not a statistical model is presumed to have generated the values of the characteristics of interest in the finite population. As well, the distribution of the estimates may be affected by the sample design. Hence, the inferences depend on the assumptions made about the randomization mechanism used to create the hypothetical repetitions.

Suppose that for a scalar parameter, these hypothetical repetitions yield estimates $\hat{\theta}_1, \hat{\theta}_2 \ldots$. Assuming the expected value of the $\hat{\theta}_i$s exists and is equal to $\mu_{\hat{\theta}}$, the limit of the average value of $\hat{\theta}_1, \hat{\theta}_2 \ldots, \hat{\theta}_K$ for large K, would converge to $\mu_{\hat{\theta}}$. When the sampling variance $V_{\hat{\theta}}$ exists, it would be the limit of the average value of $(\hat{\theta}_i - \mu_{\hat{\theta}})^2$, over $i = 1, \ldots, K$, for large K. Extension to the multiparameter case is straightforward.

We now distinguish between two common randomization mechanisms that researchers assume to have generated the hypothetical repetitions for determining the distribution of the estimates, and we introduce a third mechanism that can incorporate both. First, for design-based randomization, the sampling distribution for $\hat{\theta}$ is based on the probability randomization distribution resulting from the plan for sampling from a finite population. The values of the characteristics of interest are assumed to be fixed quantities. In cases where the finite target population is different from the survey population, it may be necessary to make additional assumptions about the relationship of the target population to the actual survey population from which the sample was drawn. As

an example, the finite target population may be for a different reference period than the survey population and reweighting on age and sex is applied to help account for the differences.

Secondly, when the statistical inferences are based on model-based randomization, it is assumed that the observed units can be considered to be a realization of random variables that follow the distribution of a statistical model. This model may need to include factors that explain the impact of the sampling plan. For example, if a stage of sampling is based on choosing clusters, a model-based approach might incorporate a random effects component in the model to account for the clustering.

A third randomization mechanism that can incorporate both design- and model-based randomization is known as model-design-based randomization. For this mechanism, the randomization distribution for the values of the characteristics of interest for the observed units is considered to be the realization of random variables arising from a three-phase process as follows:

- In the first phase, values of the characteristics of a finite target population are generated, based on random variables of a statistical model.
- The second phase augments the finite population variables with design variables, such as stratification and clustering identifiers. The values of these design variables can depend on the outcomes of the random variables in the first phase and may be random.
- In the third phase, a probability sample is selected from the finite population using the design variables.

This three-phase framework is similar to that given in Molina et al. (2001).

We note that here we are presuming that the values of the survey variables in the finite population from which the survey sample was drawn are generated before the survey frame is constructed. An alternative formulation could be that a statistical model has generated the values of the survey variables only after having selected the units to be included in the sample. For example, in DuMouchel and Duncan (1983) it is assumed that the coefficients of the linear regression model that is used to generate the values of the survey variable can depend on the particular stratum within which each of the selected units in the sample falls. However, for most researchers, the purpose of the analysis is to understand the behavior of the units in the finite target population from which the sample was drawn, and if any survey design variables are relevant to the analysis, such variables should be included in the statistical model.

We illustrate the model-design-based randomization in Fig. 1. As shown in the figure, for a given superpopulation model, several finite populations could have been generated. For each of these finite populations, we have an associated finite population quantity given by θ_N. Then for the actual finite population realized, we have a sampling plan that could give rise to a number of different samples. For the sample actually selected, we form an estimate $\hat{\theta}$, which is the one based on the realized sample. The figure illustrates that in the model-design-based framework, we consider not only the possible samples from the actual finite population from which the survey sample was selected but also the samples that could have been generated from other possible finite populations.

The model-design-based randomization framework can also include additional randomization phases that account for nonresponse and measurement error, but we do not

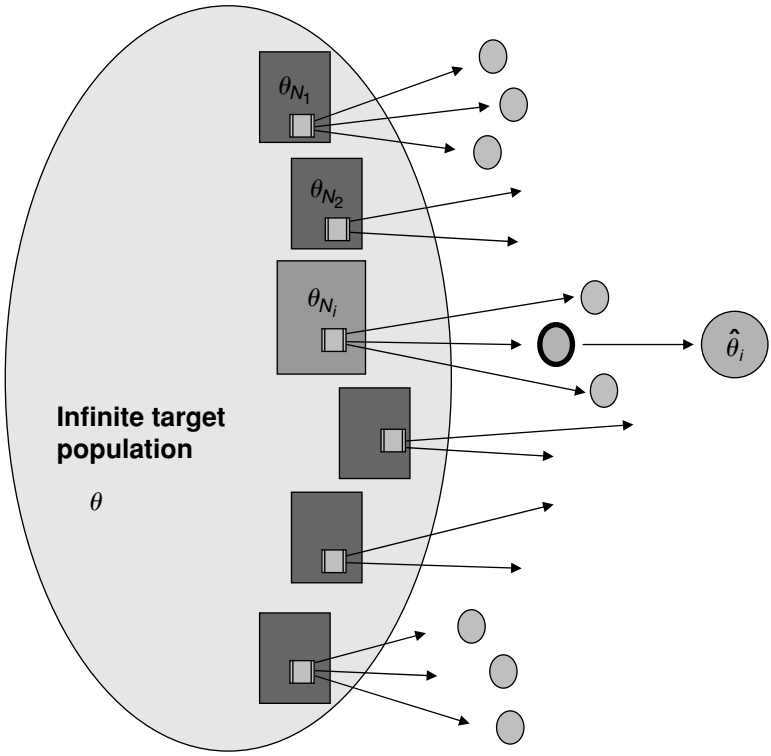


Fig. 1. Model-design randomization.

discuss these in this chapter. The methods for small area estimation (Chapter 32) and for model-based prediction of finite population totals (Chapter 23) can be included in this framework when the target population of interest is finite, but where there is a model assumed to have generated the values of the finite target population.

3.1. Informative sampling and ignorability

In the model-design-based randomization framework, two concepts encountered in the literature are informativeness and ignorability. See Pfeffermann (1993), Binder and Roberts (2001), and Chambers (2003) for some discussion of these.

In this framework, the observed sample is considered to be the result of a three-phase process as described previously, where the values of the characteristics of the finite population are generated according to the first-phase model, design variables are added in the second phase, possibly conditional on the outcomes of the random process in the first phase, and at the third phase, the sample is drawn according to the survey design. When the distribution of the sample observations resulting from the three phases is the same as the distribution that would have arisen had the observations been generated directly from the first-phase model, the sampling is said to be *noninformative*. Otherwise, the sampling is informative. One-stage simple random sampling designs are noninformative. For more complex sampling plans, whether or not the sampling is informative

will depend on the statistical distribution for the sample resulting from the three-phase process. See Chapter 39 for a more rigorous description of informative sampling.

The specification of the first-phase model can determine whether or not the sampling is informative. For example, a first-phase model that does not include certain design variables could be informative, whereas a first-phase model that includes these design variables may yield a noninformative sampling plan. Since it is the parameters of the first-phase model that are of interest to the researcher, he must give careful consideration to the appropriateness of including such design variables in the model. For some studies, including such variables could be appropriate, whereas for other studies it may not be. The following example is taken from Korn and Graubard (1999, Section 4.5). Suppose the researcher is interested in understanding the relationship between a mother's smoking behavior and gestational age of the newborn, and the sampling plan involves birthweight of the newborn. In this case, including birthweight in the model would be inappropriate for understanding the relationship between smoking and gestational age. By not including birthweight in the model, the sampling plan might be informative. On the other hand, if the sampling plan involves mother's age, then including mother's age in the model might result in having a noninformative sampling plan, but the interpretation of the other model parameters would change. The appropriateness of adding mother's age would depend on the nature of the relationship between smoking behavior and gestational age that the researcher is studying. The change to the interpretation of the model parameters resulting from including design variables in the model is also discussed in Chapter 39. It should be noted that even if the researcher chooses to include such design variables in the model, there might still be some model misspecification so that the first-phase model would still not adequately explain the sampling distribution of the observed sampled values.

If a model-based method of inference is valid under a model-design-based randomization process, the sampling is said to be *ignorable* for that analysis. Otherwise, it is *nonignorable*. For example, when fitting a linear model, suppose the model residuals are correlated within sampled clusters in a cluster sample where more than one unit is selected from each cluster. In this case, the sample design is nonignorable if the intracluster correlation is not properly taken into account in the model and in the estimation method. On the other hand, if only one unit is selected from each cluster, the intracluster correlation would not need to be taken into account in the model describing the sample observations.[4] Noninformative sampling is always ignorable. Some research has been done on diagnostics for ignorability (see Section 4.3).

4. General theory for fitting models

In this section, we give some of the technical details for the properties of design- and model-based estimators. The reader who is interested in skipping the mathematical derivations can find a summary of the main results at the end of the section. We leave out the assumptions on the rates of convergence that would be required for some of the asymptotic results to be valid. To simplify the presentation, we focus on the case of

[4] Note that if, as in Chapter 23, we were interested instead in predicting the values of the unsampled units in the cluster, the intracluster correlation would need to be taken into account.

inference for the regression coefficients of a linear regression model. However, these results can be extended to estimating parameters for many models that use linear estimating equations (see Chapter 26), such as the parameters in generalized linear models discussed by Nelder and Wedderburn (1972). Also, the framework discussed for the regression case can be modified to include the quasi-likelihood function approach to parameter estimation, (Wedderburn, 1974), the generalized estimating equation approach (Liang and Zeger, 1986) and, more generally, M-estimation for parameters of models with independently distributed observations. This section includes results first discussed by Binder and Roberts (2003).

Suppose that we are interested in estimating the parameters of a model that we assume has generated the values of the variables of interest in a finite population from which we have selected a sample. As an example, suppose that the finite population contains a dependent variable $\mathbf{Y} = (Y_1, \ldots, Y_N)'$ and p-dimensional vector-valued explanatory variables given by $\mathbf{x}_1, \ldots, \mathbf{x}_N$. We assume that population values for $\mathbf{Y}$ were generated by the model $Y_i = \mathbf{x}'_i\boldsymbol{\theta} + \varepsilon_i$, where the ε_is are independent $N(0, \sigma^2)$ random variables. If we could observe the complete finite population, the *population-based maximum likelihood estimator* for the unknown parameter $\boldsymbol{\theta}$ would be $\boldsymbol{\theta}_N$, the solution to the estimating equation

$$\mathbf{U}_N(\boldsymbol{\theta}_N) = \sum_{i=1}^{N} \mathbf{x}_i(Y_i - \mathbf{x}'_i\boldsymbol{\theta}_N) = \mathbf{0}. \tag{1}$$

The quantities $\boldsymbol{\theta}_N$ are the ordinary least-squares estimators for the regression coefficients based on the complete finite population. We say that $\boldsymbol{\theta}_N$ are the finite population quantities associated with the model parameters $\boldsymbol{\theta}$. For an arbitrary value of $\tilde{\boldsymbol{\theta}}$, we define $\mathbf{u}_i(\tilde{\boldsymbol{\theta}})$ as $\mathbf{u}_i(\tilde{\boldsymbol{\theta}}) = \mathbf{x}_i(Y_i - \mathbf{x}'_i\tilde{\boldsymbol{\theta}})$, so that $\boldsymbol{\theta}_N$, the finite population quantities associated with the model parameters $\boldsymbol{\theta}$, may be written as the solution to

$$\mathbf{U}_N(\boldsymbol{\theta}_N) = \sum_{i=1}^{N} \mathbf{u}_i(\boldsymbol{\theta}_N) = \mathbf{0}. \tag{2}$$

Suppose we have selected a sample from the finite population and we let $I_i = 1$, for $i = 1, \ldots, N$, if the ith unit is in the observed sample, and $I_i = 0$ otherwise. We denote by $\hat{\boldsymbol{\theta}}$ a model-based estimator for the unknown parameters $\boldsymbol{\theta}$, given by the solution to the estimating equation

$$\hat{\mathbf{U}}(\hat{\boldsymbol{\theta}}) = \sum_{i=1}^{N} I_i\mathbf{x}_i(Y_i - \mathbf{x}'_i\hat{\boldsymbol{\theta}}) = \sum_{i=1}^{N} I_i\mathbf{u}_i(\hat{\boldsymbol{\theta}}) = \mathbf{0}. \tag{3}$$

This model-based estimator is the *sample-based maximum likelihood estimator*. It is similar to expression (1), except that it is based on only the sampled observations. The estimator $\hat{\boldsymbol{\theta}}$ is the ordinary least-squares estimator for the regression coefficients based on the sample. This estimator is appropriate when it can be assumed that the sampling is ignorable. If, however, the sampling is not ignorable, it may be possible to modify the model assumptions for the distribution of the observed sample by accounting for how the random variables for the observed sample are affected by the sample design, and then to use estimating equations appropriate for the modified model. See Chapter 39 for some examples.

Suppose that when the ith unit is in the observed sample, the survey weight is given by w_i. Using the subscript p to denote the randomization due to the sampling process for selecting units from the finite population, a design-based weighted estimator for the unknown parameter $\boldsymbol{\theta}$ is denoted by $\hat{\boldsymbol{\theta}}_p$, the solution to the estimating equation

$$\hat{\mathbf{U}}_p(\hat{\boldsymbol{\theta}}_p) = \sum_{i=1}^{N} I_i w_i \mathbf{x}_i (Y_i - \mathbf{x}_i' \hat{\boldsymbol{\theta}}_p) = \sum_{i=1}^{N} I_i w_i \mathbf{u}_i(\hat{\boldsymbol{\theta}}_p) = \mathbf{0}. \tag{4}$$

Note that, by convention, $I_i w_i = 0$ whenever the ith unit is not in the observed sample. We are assuming that the survey weight variable is constructed in such a way that the survey-weighted estimator of a finite population total is approximately unbiased under the probability randomization distribution induced by the sampling plan. The estimator $\hat{\boldsymbol{\theta}}_p$ in (4) is known as the *pseudo maximum likelihood estimator* of the regression coefficients. It is the survey-weighted least-squares estimator for the regression coefficients based on the sample.

We consider the properties of our random variables $\boldsymbol{\theta}_N$, $\hat{\boldsymbol{\theta}}_p$, and $\hat{\boldsymbol{\theta}}$ under the model-design (ξp) randomization described in Section 3. By taking a linear expansion of $\mathbf{U}_N(\boldsymbol{\theta}_N)$ around $\boldsymbol{\theta}_N = \boldsymbol{\theta}$, we have from (1) and (2) that

$$\boldsymbol{\theta}_N - \boldsymbol{\theta} = \mathbf{S}_{xx}^{-1} \sum_{i=1}^{N} \mathbf{x}_i (Y_i - \mathbf{x}_i' \boldsymbol{\theta}) = \mathbf{S}_{xx}^{-1} \mathbf{U}_N(\boldsymbol{\theta}), \tag{5}$$

where

$$\mathbf{S}_{xx} = \sum_{i=1}^{N} \mathbf{x}_i \mathbf{x}_i'. \tag{6}$$

As well, since $\boldsymbol{\theta}_N$ is fixed for a given finite population, we see that the model-design expectation of $\boldsymbol{\theta}_N$ is the model parameter $\boldsymbol{\theta}$.

Now, considering the model variance for $\boldsymbol{\theta}_N$, we have from (5) that

$$\mathbf{V}_\xi[\boldsymbol{\theta}_N] = \mathbf{S}_{xx}^{-1} \mathbf{V}_\xi[\mathbf{U}_N(\boldsymbol{\theta})] \mathbf{S}_{xx}^{-1}. \tag{7}$$

Under the assumed linear regression model,

$$\mathbf{V}_\xi[\mathbf{U}_N(\boldsymbol{\theta})] = \sigma^2 \mathbf{S}_{xx}, \tag{8}$$

so that we obtain the familiar result that

$$\mathbf{V}_\xi[\boldsymbol{\theta}_N] = \sigma^2 \mathbf{S}_{xx}^{-1}. \tag{9}$$

It should be noted that the model-based variance given in expression (7) is correct whether or not the assumed regression model is valid.

4.1. Properties of the design-based estimator

We now consider the model-design properties of $\hat{\boldsymbol{\theta}}_p$. From expression (4), we have, using a linearization expansion of $\hat{\mathbf{U}}_p(\boldsymbol{\theta}_p)$ around $\hat{\boldsymbol{\theta}}_p = \boldsymbol{\theta}_N$, that

$$\hat{\boldsymbol{\theta}}_p - \boldsymbol{\theta}_N = \hat{\mathbf{S}}_{xx}^{-1} \hat{\mathbf{U}}_p(\boldsymbol{\theta}_N), \tag{10}$$

where

$$\hat{\mathbf{S}}_{xx} = \sum_{i=1}^{N} I_i w_i \mathbf{x}_i \mathbf{x}_i'. \tag{11}$$

For large sample sizes (or, in the case of a multistage survey design, for a large number of primary sampling units), $\mathbf{E}_p[\hat{\mathbf{S}}_{xx}] \approx \mathbf{S}_{xx}$. We assume, therefore, that $\hat{\mathbf{S}}_{xx} \to \mathbf{S}_{xx}$ in probability, so that

$$\hat{\boldsymbol{\theta}}_p - \boldsymbol{\theta}_N \to \mathbf{S}_{xx}^{-1} \hat{\mathbf{U}}_p(\boldsymbol{\theta}_N) \tag{12}$$

in probability. (See Chapter 40 for technical details of asymptotics in finite population sampling). Since

$$\mathbf{E}_p[\hat{\mathbf{U}}_p(\boldsymbol{\theta}_N)] \approx \hat{\mathbf{U}}_p(\boldsymbol{\theta}_N) = \mathbf{0}, \tag{13}$$

we have from (12) that $\hat{\boldsymbol{\theta}}_p$ is asymptotically design unbiased for $\boldsymbol{\theta}_N$. Also, since $\boldsymbol{\theta}_N$ is model unbiased for $\boldsymbol{\theta}$, we have that $\hat{\boldsymbol{\theta}}_p$ is asymptotically model-design unbiased for $\boldsymbol{\theta}$.

The choice of whether to use the design-based survey-weighted estimator or the model-based ordinary least-squares estimator for the regression coefficients should be made on the basis of efficiency or robustness. We say that a method is robust when the method gives valid inferences not only under the ideal conditions of the assumed model being correct but also when there is a departure from these assumptions. The ordinary least-squares estimator will have smaller model-design-based variances than the survey-weighted estimator when the sampling is ignorable since the ordinary least-squares estimator is the minimum variance linear unbiased estimator under the model. However, using model-based estimates and model-based variances could lead to inappropriate inferences under the model-design-based framework when the model assumptions are violated for the sample. On the other hand, as we discuss below, inferences based on design-based estimators can be valid under the model-design-based framework, even when some of the model assumptions are violated.

We now consider the model-design-based variance of $\hat{\boldsymbol{\theta}}_p$. We use the decomposition

$$\mathbf{V}_{\xi p}[\hat{\boldsymbol{\theta}}_p] = \mathbf{E}_\xi \mathbf{V}_p[\hat{\boldsymbol{\theta}}_p] + \mathbf{V}_\xi \mathbf{E}_p[\hat{\boldsymbol{\theta}}_p]. \tag{14}$$

From expression (12), we see that, asymptotically,

$$\mathbf{V}_p[\hat{\boldsymbol{\theta}}_p] = \mathbf{S}_{xx}^{-1} \mathbf{V}_p[\hat{\mathbf{U}}_p(\boldsymbol{\theta}_N)] \mathbf{S}_{xx}^{-1}. \tag{15}$$

Therefore, from expression (12) and (14),

$$\mathbf{V}_{\xi p}[\hat{\boldsymbol{\theta}}_p] \to \mathbf{S}_{xx}^{-1} \mathbf{E}_\xi \mathbf{V}_p[\hat{\mathbf{U}}_p(\boldsymbol{\theta}_N)] \mathbf{S}_{xx}^{-1} + \mathbf{V}_\xi[\boldsymbol{\theta}_N] \tag{16}$$

for large samples. Since $\boldsymbol{\theta}_N \to \boldsymbol{\theta}$, it follows that

$$\mathbf{V}_{\xi p}[\hat{\boldsymbol{\theta}}_p] \to \mathbf{E}_\xi \mathbf{V}_p \left[\mathbf{S}_{xx}^{-1} \hat{\mathbf{U}}_p(\boldsymbol{\theta}) \right] + \mathbf{V}_\xi[\boldsymbol{\theta}_N]. \tag{17}$$

For many sample designs and models, $\mathbf{E}_\xi \mathbf{V}_p[\mathbf{S}_{xx}^{-1} \hat{\mathbf{U}}_p(\boldsymbol{\theta})]$ is $O(n^{-1})$. Also since $\mathbf{V}_\xi[\boldsymbol{\theta}_N]$ is $O(N^{-1})$, the magnitude of the term $\mathbf{V}_\xi[\boldsymbol{\theta}_N]$, compared with the magnitude of $\mathbf{V}_{\xi p}[\hat{\boldsymbol{\theta}}_p]$,

is negligible when the sampling fraction, n/N, is small. We note that even if there are some units in the finite population that are selected with certainty, if the overall sampling fraction is small, $\mathbf{E}_\xi \mathbf{V}_p[\mathbf{S}_{xx}^{-1}\hat{\mathbf{U}}_p(\boldsymbol{\theta})]$ is $O(n^{-1})$. Therefore, for small sampling fractions and large sample sizes, the design-based variance of $\hat{\boldsymbol{\theta}}_p$ is model unbiased for the design-model-based variance of $\hat{\boldsymbol{\theta}}_p$. This is an important result for the analysis of survey data from complex surveys, since the design-based variance is derived from only the probability randomization distribution of the sampling plan, without explicit reference to the first-phase model presumed to have generated the finite population values.

It is important to note also that the model-design-based variance of $\hat{\boldsymbol{\theta}}_p$ in (17) is valid even if the assumed regression model is incorrect, providing the model-expected value of $\mathbf{U}_N(\boldsymbol{\theta})$ is equal to zero; that is, providing the model is linear with no missing explanatory variables. In practice, we would need to estimate the model-design-based variance in (17). If there is an asymptotically design-unbiased estimator of $\mathbf{V}_{\xi p}[\hat{\boldsymbol{\theta}}_p]$, then this estimator will be asymptotically model-design unbiased. To estimate the design-based variance $\mathbf{V}_p[\mathbf{S}_{xx}^{-1}\hat{\mathbf{U}}_p(\boldsymbol{\theta})]$ in (17), we note that $\hat{\mathbf{U}}_p(\boldsymbol{\theta})$ is an estimator of a total, so that design-based methods for estimating the variance of a total for a finite population may be used. Since $\boldsymbol{\theta}$ and $\mathbf{S}_{xx}$ are unknown, we would substitute the estimates using $\hat{\boldsymbol{\theta}}_p$ and $\hat{\mathbf{S}}_{xx}$, respectively, to compute the estimated design-based variance. (We briefly discuss estimating the design-based variance in Section 6.)

We see, therefore, that when the sampling fraction is small and the sample size is large, the design-based variance, as an estimate of the model-design-based variance of $\hat{\boldsymbol{\theta}}_p$, is robust to certain departures from the variance assumptions under the model, provided that $\hat{\boldsymbol{\theta}}_p$ is asymptotically design unbiased for $\boldsymbol{\theta}_N$, and that $\boldsymbol{\theta}_N \to \boldsymbol{\theta}$.

However, when the sampling fraction is not small, a design-based estimator of the variance would not be completely appropriate, since the second term of $\mathbf{V}_{\xi p}[\hat{\boldsymbol{\theta}}_p]$ in (14) would not be properly accounted for. Korn and Graubard (1998a) discuss this situation in some detail. We discuss the estimation of $\mathbf{V}_\xi[\boldsymbol{\theta}_N]$ further in Section 5.1.

A question that is often asked is whether replacing the design-based variance $\mathbf{V}_p[\mathbf{S}_{xx}^{-1}\hat{\mathbf{U}}_p(\boldsymbol{\theta})]$ in (17) by a model-based variance $\mathbf{V}_\xi[\mathbf{S}_{xx}^{-1}\hat{\mathbf{U}}_p(\boldsymbol{\theta})]$ yields an asymptotically correct model-design-based variance of $\hat{\boldsymbol{\theta}}_p$. If this were the case, it would be possible to estimate the model variance, using only the survey weights and any design information used in the first-phase model, without needing all the design information that would be normally required for estimating the design-based variance. Now if the sampling is ignorable, then $\mathbf{V}_{\xi p}[\mathbf{S}_{xx}^{-1}\hat{\mathbf{U}}_p(\boldsymbol{\theta})] = \mathbf{V}_\xi[\mathbf{S}_{xx}^{-1}\hat{\mathbf{U}}_p(\boldsymbol{\theta})]$. Since when the sampling fraction is small and the sample size is large $\mathbf{V}_p[\mathbf{S}_{xx}^{-1}\hat{\mathbf{U}}_p(\boldsymbol{\theta})]$ is model unbiased for $\mathbf{V}_{\xi p}[\mathbf{S}_{xx}^{-1}\hat{\mathbf{U}}_p(\boldsymbol{\theta})]$, the model-based variance and the design-based variance are asymptotically equal. Providing we have consistent estimates of the variances, the two estimated variances should be close when the sampling is ignorable. Therefore, comparing the two variances is a possible diagnostic tool for whether or not the sampling is ignorable (see Section 4.3 for more discussion of tools for diagnosing nonignorability).

4.2. Properties of the model-based estimator

We now consider the model-design properties of the model-based estimator $\hat{\boldsymbol{\theta}}$ for estimating the parameter $\boldsymbol{\theta}$. By taking a linear expansion of $\hat{\mathbf{U}}(\hat{\boldsymbol{\theta}})$ around $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}$, we have

from expression (3) that

$$\hat{\boldsymbol{\theta}} - \boldsymbol{\theta} = \left[\sum_{i=1}^{N} I_i \mathbf{x}_i \mathbf{x}_i'\right]^{-1} \sum_{i=1}^{N} I_i \mathbf{u}_i(\boldsymbol{\theta}). \tag{18}$$

We denote $E_p[I_i]$ by π_i, the probability that the ith unit is included in the sample. Now, we assume that for large sample sizes from a given finite population that

$$\sum_{i=1}^{N} I_i \mathbf{x}_i \mathbf{x}_i' \to \mathbf{E}_p\left[\sum_{i=1}^{N} I_i \mathbf{x}_i \mathbf{x}_i'\right] = \sum_{i=1}^{N} \pi_i \mathbf{x}_i \mathbf{x}_i', \tag{19}$$

which we denote by $\mathbf{S}_{xx}^{(\pi)}$. Therefore, by taking design-based expectations in expression (18), it follows that, asymptotically,

$$\mathbf{E}_p[\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}] \to \left[\mathbf{S}_{xx}^{(\pi)}\right]^{-1} \sum_{i=1}^{N} \pi_i \mathbf{u}_i(\boldsymbol{\theta}), \tag{20}$$

in probability. Note that the π_is can vary under hypothetical repetitions of the finite population generated by the model. We assume that as $n, N \to \infty$, the limit of $\mathbf{S}_{xx}^{(\pi)}$ exists and is equal to $\mathbf{E}_\xi[\mathbf{S}_{xx}^{(\pi)}]$. Taking the model expectation of expression (20), it follows that, asymptotically,

$$\mathbf{E}_{\xi p}[\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}] = \left[\mathbf{E}_\xi\left[\mathbf{S}_{xx}^{(\pi)}\right]\right]^{-1} \sum_{i=1}^{N} \mathbf{E}_\xi\left[\pi_i \mathbf{u}_i(\boldsymbol{\theta})\right]. \tag{21}$$

We know that if the sampling is ignorable, $\hat{\boldsymbol{\theta}}$ is model-design unbiased for $\boldsymbol{\theta}$. However, in general, under nonignorable designs, the model-based estimate of $\boldsymbol{\theta}$ may be biased and inconsistent.

Using methods similar to the case of the design-based estimator, it is possible to derive the expression for the model-based variance of $\hat{\boldsymbol{\theta}}$, but we leave out the details here. Again if the sampling is ignorable, the model variance $\mathbf{V}_\xi[\hat{\boldsymbol{\theta}}]$ will be design unbiased for $\mathbf{V}_{\xi p}[\hat{\boldsymbol{\theta}}]$.

When under a specific model, the sampling is nonignorable, it may be possible, as mentioned in Section 3.1, to modify the model to incorporate the effect of the sampling mechanism; see, for example, Chambers (1986). An alternative approach that explicitly accounts for the conditional distribution given the sampling mechanism is given in Chapter 39.

4.3. Comparing model- and design-based estimators

When the sample design is not ignorable, we know that both $\hat{\boldsymbol{\theta}}_p$ and $\hat{\boldsymbol{\theta}}$ are model-design unbiased for estimating $\boldsymbol{\theta}$. Also, for ignorable sample designs and for small sampling fractions, the model-based variance and the design-based variance for $\mathbf{S}_{xx}^{-1}\hat{\mathbf{U}}_p(\boldsymbol{\theta})$ (and hence for $\hat{\boldsymbol{\theta}}_p$) are asymptotically equal. Therefore, an indicator for checking on the ignorability of the sample design is to compare the values of $\hat{\boldsymbol{\theta}}_p$ and $\hat{\boldsymbol{\theta}}$ and to compare an estimate of $\mathbf{V}_\xi[\hat{\boldsymbol{\theta}}_p]$ with an estimate of $\mathbf{V}_p[\hat{\boldsymbol{\theta}}_p]$. Another possibility, but one that is often not as simple to compute, is to compare an estimate of $\mathbf{V}_\xi[\hat{\boldsymbol{\theta}}]$ with an estimate of $\mathbf{V}_p[\hat{\boldsymbol{\theta}}]$, the model and design-based variances of the model-based estimates of the parameters.

Table 1
Properties of design- and model-based estimators under model-design-based randomization (for small sampling fractions and large sample sizes)

	Assumed first-phase model is valid and sampling is ignorable	Assumed first-phase model is misspecified or the sampling is nonignorable
Model-based estimator	Asymptotically unbiased Efficient Valid variance estimates Valid inferences May be best	May be inconsistent Variance estimates may be invalid Inferences may be invalid
Design-based estimator	Asymptotically unbiased May be inefficient Valid variance estimates Valid inferences	If the mean of the estimating equation is zero under the model: Asymptotically unbiased Valid variance estimates Valid inferences

For a single parameter θ, the factor, $\hat{V}_p[\hat{\theta}_p]/\hat{V}_\xi[\hat{\theta}_p]$, is the estimate of the inflation of the variance due to the sample design. Binder et al. (2005) have conducted some simulations to study this. Other measures of ignorability based on only the point estimates have also been proposed in the literature; for example, Pfeffermann (1993) and Asparouhov (2004). DuMouchel and Duncan (1983) and Fuller (1984) have suggested statistical tests for testing whether $\hat{\boldsymbol{\theta}}_p$ and $\hat{\boldsymbol{\theta}}$ are estimating the same quantities in the case of a linear regression model. More generally, it is also possible to compute $[\hat{\mathbf{U}}_p(\hat{\boldsymbol{\theta}})]$ and a model-based estimate of its model variance and to test for whether $\mathbf{E}_\xi[\hat{\mathbf{U}}_p(\hat{\boldsymbol{\theta}})] = \mathbf{0}$. Some further discussion is given in Chapter 39.

To summarize the properties of the design- and the model-based estimators, Table 1 displays the advantages and disadvantages of model-based versus design-based estimation.

5. Cases where design-based methods can be problematic

There are situations where the general theory given to support design-based methods of inference for model parameters can be problematic. We consider some of these here.

5.1. Nonnegligible sampling fractions

When the sampling fraction is small and the sample size is large, the term, $\mathbf{V}_\xi[\boldsymbol{\theta}_N]$ in expression (16), can be ignored for the model-design-based variance of $\hat{\boldsymbol{\theta}}_p$. However, when the sampling fraction is not negligible, this term must be included. Since this term depends on the model, it would seem to be necessary to use model assumptions to estimate this. For example, if the coefficients of a linear model have been estimated by $\hat{\boldsymbol{\theta}}_p$, the survey-weighted least-squares estimate, we would need to estimate $\mathbf{V}_\xi[\boldsymbol{\theta}_N] = \mathbf{V}_\xi[(\mathbf{X}_N'\mathbf{X}_N)^{-1}\mathbf{X}_N'\mathbf{y}_N]$, which would require knowledge of the model variance structure. However, for many sample designs, it is possible to estimate the model-design-based variance $\mathbf{V}_{\xi p}[\hat{\boldsymbol{\theta}}_p]$ even without full model details. It turns out that using a variance estimator that assumes that the sampling is with replacement at the

first stage of selection will give an estimate of the correct model-design-based variance regardless of the sampling fraction.

We now demonstrate why the with-replacement variance estimator can be appropriate by considering the example of a single-stage sample design for the case of fitting a linear regression model with independent and identically distributed normal errors. Denoting $E_p[I_i I_j w_i w_j]$ by d_{ij}, and taking the simple case where $E_p[I_i w_i] = 1$, the model-design-based variance of $\hat{\mathbf{U}}_p(\tilde{\boldsymbol{\theta}})$ is

$$\mathbf{V}_{\xi p}[\hat{\mathbf{U}}_p(\tilde{\boldsymbol{\theta}})] = \sum_{i=1}^{N}\sum_{j=1}^{N} \mathbf{E}_\xi[d_{ij}\mathbf{u}_i(\tilde{\boldsymbol{\theta}})\mathbf{u}_j'(\tilde{\boldsymbol{\theta}})], \tag{22}$$

since $\mathbf{V}_\xi \mathbf{E}_p[\hat{\mathbf{U}}_p(\tilde{\boldsymbol{\theta}})] = \mathbf{0}$. A design-based estimator for the variance of $\hat{\mathbf{U}}_p(\tilde{\boldsymbol{\theta}})$, assuming a with-replacement sampling plan for a single-stage design, is given by

$$\hat{\mathbf{V}}_{wr}[\hat{\mathbf{U}}_p(\tilde{\boldsymbol{\theta}})] = \frac{n}{n-1}\left[\sum_{i=1}^{N} I_i w_i^2 \mathbf{u}_i(\tilde{\boldsymbol{\theta}})\mathbf{u}_i'(\tilde{\boldsymbol{\theta}}) - \frac{1}{n}\sum_{i=1}^{N}\sum_{j=1}^{N} I_i I_j w_i w_j \mathbf{u}_i(\tilde{\boldsymbol{\theta}})\mathbf{u}_j'(\tilde{\boldsymbol{\theta}})\right]. \tag{23}$$

It can be shown that if $\mathbf{E}_\xi[d_{ij}\mathbf{u}_i(\tilde{\boldsymbol{\theta}})\mathbf{u}_j'(\tilde{\boldsymbol{\theta}})] = \mathbf{0}$ for $i \neq j$ then the model-design-based expectation of the with-replacement variance estimator in (23) is equal to the model-design-based variance in (22) when $\tilde{\boldsymbol{\theta}}$ is equal to the true value of the model parameter $\boldsymbol{\theta}$.

This result can also be extended to the case where there is some nonzero model correlation between units, provided that this correlation is not too large. We do not give the details here.

In practice, to estimate the variance, an estimator for $\boldsymbol{\theta}$ in expression (23) is required, but the effect on the estimate of substituting $\hat{\boldsymbol{\theta}}_p$ for $\boldsymbol{\theta}$ in (23) is asymptotically negligible.

5.2. *Small sample sizes or rare characteristics*

For small-scale surveys, or for studies of domains where the sample sizes are small, a model-based approach may be preferred to a design-based approach, especially if the design-based approach leads to much higher variances; see, for example, Kalton (1983) and Little (2004). Small sample sizes can also occur in the case of event history analysis when the number of observed events is small, even for a large-scale survey. As mentioned in Section 4, the use of design-based (weighted) estimates rather than model-based (unweighted) estimates can lead to less efficient estimates when the model is correct. When the sample size is small, this increased variance may be unacceptable. If a model-based approach is taken, consideration should be given to incorporating into the model features of the sample design so that the sample is ignorable, as discussed in Section 3.1.

Problems with typical design-based methods can also arise when the researcher is fitting models for rare characteristics, even when the survey sample size is large. As we mention in Section 6 below, for estimating variances using jackknife or bootstrap replicates, alternatives, such as the linearized jackknife or the linearized estimating function bootstrap (see Binder et al., 2004; Yung and Rao, 1996), might be considered for estimating design-based variances in such cases. Korn and Graubard (1998b)

discussed alternatives for confidence interval estimation of estimating population proportions when the sample contains only a small number of positive counts.

5.3. Models that include random effects

Many models of interest to researchers contain random effects. Mixed effects models, of which hierarchical or multilevel models are specific cases, are models that include both fixed effects parameters and parameters for variance components corresponding to random effects. We consider two issues here: first, that of finding appropriate point estimates for both the fixed effects parameters and the random effects parameters, and second, that of problems associated with making appropriate inferences about these parameters. We consider both these issues in the context of the model-design-based randomization framework.

In Section 4, we described large-sample approaches that are appropriate when the estimating equations provide consistent estimates of the unknown parameters. However, for hierarchical or multilevel models, for example, using unrestricted maximum likelihood estimators to define the estimating equations for point estimates of the parameters of the variance components can lead to biased and inconsistent estimators when the number of units sampled at the first level of the hierarchy[5] is small even when the sample design is ignorable. Careful consideration must be given to deciding which estimators are appropriate for each of the following three cases: (a) when the complete finite population is observed, (b) when a sample of the finite population is observed and the sample design is ignorable, and (c) when a sample of the finite population is observed, and the sample design is not ignorable. Cases (a) and (b) are covered in the nonsurvey-based statistical literature using model-based approaches. For case (c), even the meaning of a "design-based approach" is not necessarily clear. What has been proposed in the literature for this case is to first determine appropriate model-based estimators for case (a), and by expressing these estimators in terms of estimating equations, to develop weighted estimating equations as a basis for defining the estimators. However, finding appropriate weights using this approach can be problematic. Weighting at each level of the model hierarchy is required and some weight rescaling within levels is generally recommended. Acceptable rescaling methods are still being researched.

Some approaches considered for point estimation of the model parameters are iterative generalized least-squares *multilevel pseudo maximum likelihood* and a model-dependent approach; see, for example, Pfeffermann et al. (1998b), Kovačević and Rai (2003), Asparouhov (2006), Pfeffermann et al. (2006), Pfeffermann and Sverchkov (2007); however, such approaches can be complex to implement and may not give the desired results. Korn and Graubard (2003), for example, have suggested that survey-weighted estimators for estimating the variance components, even when using rescaled weights, could be badly biased.

Although estimating the variance parameters of random effects components can be problematic, the estimation of the fixed effects (such as the fixed regression coefficients in a mixed effects model) is more tractable. To see this, we partition the model parameter

[5] In the hierarchical modeling literature, the first level of the hierarchy is what the sampling literature would refer to as the final stage of sampling. For example, if the model includes students within schools, the first level of the hierarchy is the sample of students.

$\boldsymbol{\theta}$ into two components, one for the fixed effects parameters ($\boldsymbol{\theta}_0$) and one for the random effects parameters ($\boldsymbol{\eta}$); that is, $\boldsymbol{\theta} = (\boldsymbol{\theta}_0', \boldsymbol{\eta}')'$. For many models we can partition the estimating functions into two components $\mathbf{U}_p(\boldsymbol{\theta}_0, \boldsymbol{\eta}) = [\mathbf{U}_p^1(\boldsymbol{\theta}_0, \boldsymbol{\eta})', \mathbf{U}_p^2(\boldsymbol{\theta}_0, \boldsymbol{\eta})']'$, where the first component $\mathbf{U}_p^1(\boldsymbol{\theta}_0, \boldsymbol{\eta})$ has model-design-based expectation equal to zero when $\boldsymbol{\theta}_0$ is the true value of the first component of $\boldsymbol{\theta}$, for any arbitrary value of $\boldsymbol{\eta}$. We assume that the dimensionality of $\mathbf{U}_p^1(\boldsymbol{\theta}_0, \boldsymbol{\eta})$ is the same as the dimensionality of $\boldsymbol{\theta}_0$, and that $\boldsymbol{\theta}_0$ can be estimated by $\hat{\boldsymbol{\theta}}_0$ by setting $\mathbf{U}_p^1(\hat{\boldsymbol{\theta}}_0, \boldsymbol{\eta})$ to zero. In general, the estimate of $\boldsymbol{\theta}_0$ will depend on $\boldsymbol{\eta}$. In the case where $\boldsymbol{\theta}_0$ corresponds to the regression coefficients of linear terms, $\hat{\boldsymbol{\theta}}_0$ would be asymptotically model-design unbiased for estimating $\boldsymbol{\theta}_0$, the fixed effects parameters, and the variance arising from (15) would be asymptotically valid under the same assumptions made for the previously discussed design-based estimator in (4). For example, for a linear regression model with error terms originating from a random effects model, the design-based estimates of the regression coefficients can be model-design unbiased even when the estimates of the parameters of the error model are biased. Generally, design-based methods may be used to estimate the variance of these estimated regression coefficients.

In addition to the problem of finding suitable point estimates for the parameters of the model, there is the issue of making inferences, such as constructing confidence intervals for the estimates, especially for the estimates of the parameters of a random effects model. Research in the area of estimating variances of the estimates of both the fixed effects parameters and the random effects parameters is quite recent. Pfeffermann et al. (1998b) considered a robust design-based sandwich estimator, a variant of the Taylor linearized method. Asparouhov (2004, 2006) advocates the use of a similar robust sandwich estimator. Rabe-Hesketh and Skrondal (2006) also used a sandwich (model-based) estimator in a generalized linear model. Korn and Graubard (2003) suggested variance estimation based on resampling clusters (level-2 units), in particular the delete-one PSU jackknife method. Grilli and Pratesi (2004) described a possible two-stage bootstrap, but used only the cluster bootstrap for variance estimation when fitting multilevel ordinal and binary models. Stapleton (2002) also discussed these issues for structural equation models. Multilevel models are also discussed in Chapter 39.

5.4. Population-based case-control studies

For some population-based surveys, an analytic goal is to compare a sample of cases and a sample of controls to estimate the strength of association of certain causal factors for an outcome being studied. Such studies may have very different sampling fractions for the case and control groups. Although performing a standard model-design-based analysis is viable here, the variances of the estimates using design-based methods can be unacceptably large due to the difference in the weights between the two groups. Scott (2006) used weight rescaling to improve the efficiency of the estimates, still allowing for informative sampling. (See also Chapter 38). We discuss more general cases of integrating population-based surveys in Section 7.

5.5. Event history analysis

As mentioned in Section 2.2, survey weight variables are constructed to ensure that estimates for a finite target population are approximately design unbiased for the population

characteristics. However, in a longitudinal survey, there may be more than one survey weight variable on the data file because a different number of units responds to each cycle of the survey. Generally, each of these weight variables is constructed by the survey producer so that the use of the nonzero weights of a particular weight variable to compute a weighted estimate would yield approximately unbiased estimates of population totals for the survey producer's target population of the longitudinal survey at a particular point in time.

When a survival distribution is assumed to have been generated by a statistical model, it is necessary to consider which survey weight is appropriate when using a design-based approach for fitting this model. As Lawless (2003) pointed out, the situation is complex because spells can originate and end at arbitrary times between follow-up interviews or data collection points. This could create a need for time-varying weights. For example, for analyzing spell data, the appropriate weight could depend on the survey cycles in which the observed start points or observed end points of the spells occur. This is an area where further research is needed.

We also note that for fitting a proportional hazards model to complex survey data, Binder (1992) and Lin (2000) gave details on how to perform a linearization that can be used to compute the design-based variance.

6. Estimation of design-based variances

For large sample sizes (or, in the case of a multistage survey design, for a large number of primary sampling units), confidence intervals for quantities of interest are usually constructed and tests of hypotheses are conducted assuming approximate normality of a pivotal quantity, such as a t-statistic or a z -score. Confidence intervals may also be obtained more directly by inverting the confidence interval based on the estimating function itself, using the estimated variance of the estimating function; see, for example, Binder and Patak (1994). What is required to implement these procedures are estimates of the variances of estimators. Variance estimation techniques for estimates of finite population totals using data from complex surveys are well established. Techniques for more complex estimates, such as design-based variance estimates for ratios and for model parameters, are summarized in Lohr (1999, Chapter 8). Also see Chapter 2.

As we have discussed in Section 4.1, the asymptotic design-based variances for estimates of model parameters can be derived from the design-based variances of estimating equations evaluated at the estimated parameter values. This approach is the basis for methods that use linearization techniques for estimating variances. Linearization techniques require that a separate formula be developed for each complex estimator. Approaches for deriving the appropriate formulae have been discussed by Binder (1983), Binder (1996), and Demnati and Rao (2004).

To avoid some of the complexities of linearization techniques, various approaches for estimating the variance using resampling or replication techniques have been developed. For a more complete discussion of these, see Chapter 28. The survey bootstrap, random groups, balanced repeated replication, and the survey jackknife are all examples of commonly-used replication methods to obtain design-based variance estimates. As is the practice for most linearization techniques, many of these replication methods make

simplifying approximations to the true survey design by assuming with-replacement sampling of the primary sampling units within each stratum.

In a common approach to estimating variances using the survey bootstrap, the weighted estimating equations must be solved using the weight variables associated with each of the bootstrap replicates in turn. As Rao (2005) pointed out, a possible difficulty with this implementation of the survey bootstrap is that the solution to the estimating equations may not exist for some bootstrap replicates. Binder, Kovačević, and Roberts (2004) considered various estimating function bootstrap methods that avoid this difficulty, including the linearized estimating function bootstrap. Yung and Rao (1996) suggested a linearized jackknife procedure, and Rao and Tausi (2004) suggested alternative estimating function jackknife variance estimators. The linearized estimating function bootstrap and the linearized estimating function jackknife are equivalent to a Taylor linearization sandwich variance estimator based on expression (15), using the jackknife or bootstrap variance estimator to estimate the variance in the centre of the "sandwich."

7. Integrating data from more than one survey

Comparable variables are frequently available from more than one survey source, leading researchers to ask whether and how the data from the different sources could be integrated into a single analysis. These questions are most frequently asked when the sample sizes for the problem under study are small in each of the survey sources. This topic is discussed in Korn and Graubard (1999, Chapter 8).

We discuss here the case where a statistical model can be assumed to have generated the values of the characteristics of interest for the survey provider's finite target populations for each of the surveys being integrated. This statistical model may contain parameters that are specific to each finite population. We also confine ourselves to the situation where the samples of the different survey sources are independently selected. We give two broad choices for integrating the data – a pooling approach and a separate approach. It should be noted that only under very specific conditions would the two approaches give the same point estimates.

7.1. Pooling approach to integration

For the pooling approach to integration, the researcher considers each survey target population to be a superstratum of the larger finite population defined as the union of all the finite target populations for the individual surveys. The data from the different surveys are then pooled together and treated as if they were from a single survey from this larger population. It is then generally straightforward to allow for and to test for inequalities in parameters among the different finite populations making up the larger population. For example, in the case where an assumed model describes a linear relationship between a dependent variable and explanatory variables, the modeling process could begin with distinct intercept and slope parameters for the different finite populations; statistical tests could then be carried out for assessing whether common intercepts or slope parameters would be sufficient.

In some situations, such as when the sample sizes or the survey designs are very different among the surveys being combined, consideration could be given to whether more

efficient estimates of the parameters, could be obtained through rescaling the weights of the different surveys. However, a rescaling that is most efficient for one parameter estimate is unlikely to be most efficient for another. As well, an estimate based on rescaled weights may not be estimating a meaningful quantity when a parameter is erroneously assumed to be constant over the different finite populations. A situation where rescaling does improve efficiency is mentioned in Section 5.4 in the context of population-based case-control studies.

7.2. Separate approach to integration

For the separate approach to integrating data from more than one survey, the parameters of interest are first estimated from each data source separately and then the estimates are combined through averaging. As for the pooling approach, it is advisable to check on the assumption of equality of the parameters across the different finite populations, through statistical testing, even though the power of the statistical tests may not be high if the sample sizes from the different survey sources are small.

A weighted average of the estimates of an individual parameter could give a more efficient estimate than a simple average. In fact, the optimal weighted average is achieved by using a weight for each estimate that is inversely proportional to its variance. However, there are problems with attempting to obtain optimal weighted estimates. One is that, in practice, the variances need to be estimated, and these estimates could be quite inaccurate when the sample sizes from each of the surveys are small. Another problem is that variances could be quite different for different parameters, so that a weighting that would give an optimal estimate of one parameter would not necessarily be optimal for another parameter. When implementing a separate approach, a possible tactic would be to determine some average or percentile of variance estimates of a large number of parameters in each survey to produce a set of common weighting factors to be used for combining estimates of all parameters.

8. Some final remarks

Issues associated with making inferences for model parameters for data obtained from a complex survey need to be discussed in the context of the population targeted by the survey producer and the relationship between this population and the parametric model. Typical complex survey designs often lead to nonignorable samples for many models assumed by researchers. When the sample is nonignorable, design- (weighted) and model-based (unweighted) point estimates may or may not be similar. If they are not similar, it may be possible to modify the model for the observations in the sample to account for the effect of the sampling mechanism on the distribution of the survey data. However, even if the point estimates are similar, a modified model might be more appropriate if it is possible that the observed sample is informative. A tool to indicate if the sample is nonignorable is a comparison of the design- and the model-based point estimates, as well as a comparison of an estimate of the design-based variance of an estimate with an estimate of its model-based equivalent.

When modifying a model to account for the impact of the survey design, such as by adding survey design variables to the model, the original model parameters may change

to having an unintended interpretation. This could be the case if the researcher is not interested in the impact of the design variables on the characteristics being studied. This point is also discussed in Chapter 39.

When the sample size is small, a model-based approach may be preferred to a design-based approach, especially if the design-based approach leads to much higher variances; see Kalton (1983). The researcher should attempt to use models that would yield non-informative samples, if possible.

We have not discussed computer software that is appropriate for analyzing survey data, as this is covered in Chapter 13. However, it is worth noting that many researchers use software that has not been developed to fully account for the survey design, even though the software can compute survey-weighted point estimates. Many users rescale the weight variable when using such software knowing that by using normalized survey weights (rescaling the weights to sum to the sample size) the survey-weighted point estimates can be obtained, and also expecting that the variance estimates computed by the software would be correct if the sampling is ignorable. However, in general, such software does not provide correct design-based variance estimates, even when the sample is ignorable. Therefore, the inferences using conventional software with normalized weights are not, in general, valid, even when the sampling is ignorable. However, for a standard linear regression model, the with-replacement variance estimator given in (23) is equivalent to the Huber–White robust estimator that is available in some commercial software, and this estimator would be suitable when the sampling is ignorable.

It should be noted that, when the sampling is ignorable, the model-based approach could lead to more powerful tests of hypotheses and shorter confidence intervals for estimated parameters than the design-based approach.

Of course, the first-phase model assumed to have generated the finite population values may have been incorrectly specified. A design-based approach may be robust to departures from the assumed correlation structure for the model errors. However, when the model means are misspecified, both the design- and model-based approaches may be misleading. This was also pointed out recently by Kott (2007) in the context of regression analysis. When the model means are correctly specified, the design-based approach can adjust for misspecification of the model variance structure.

There seems to be a misconception that the use of finite population methods is wrong when you want to analyze a phenomenon believed to hold beyond the finite population under study. However, as we have seen, these finite population methods are, in most common cases, also appropriate for inferences beyond that finite population.

Finally, we mention some of the areas where further research is needed:

- Fitting and making inferences for mixed effects models
- Fitting models to data from more than one survey
- Estimating model-design-based variances when the sampling fractions are large
- Improving resampling methods for estimating design-based variances of model parameters estimates.

Sample Surveys: Inference and Analysis, Vol. 29B
ISSN: 0169-7161
© 2009 Elsevier B.V. All rights reserved
DOI: 10.1016/S0169-7161(09)00225-9

Calibration Weighting: Combining Probability Samples and Linear Prediction Models

Phillip S. Kott

1. Introduction

Suppose we wanted to estimate totals for a number of target variables based on data from a probability sample. If we knew the selection probability, π_k, for each sample element k in the sample S, then we could estimate any population total, $T_y = \sum_{i \in U} y_k = \sum_U y_k$, where U denotes the population, with the expansion estimator $t_y^E = \sum_S y_k/\pi_k = \sum_U y_k I_k/\pi_k$, where $I_k = 1$ when $k \in S$ and 0 otherwise. Treating the I_k as random variables, it is easy to see that t_y^E is an unbiased estimator for T_y. We call properties arising when the I_k are treated as random variables randomization-based. Although the term "design-based" is more commonly used, it is a misnomer because there are nonprobability sampling designs.

We can also write $t_y^E = \sum_U d_k y_k = \sum_S d_k y_k$, where $d_k = I_k/\pi_k$ is called "the sampling weight of element k." These weights can be used for estimating the population total of any survey variable, that is, any variable whose values are collected from the sampled elements.

Deville and Särndal (1992) coined the term "calibration estimator" to describe an estimator of the form $t_y^{\mathrm{CAL}} = \sum_S w_k y_k$, where $\sum_S w_k \mathbf{x}_k = \sum_U \mathbf{x}_k = T_\mathbf{x}$ for some row vector of P benchmark variables, $\mathbf{x}_k = (x_{1k}, \ldots, x_{Pk})$, about which $T_\mathbf{x}$ is known. Benchmark variables are often called "auxiliary variables" or "control variables." In the latter case, $\mathbf{x}_k$ is usually known for all $k \in U$.

There is generally a continuum of sets $\{w_k | k \in S\}$, that satisfy the (vector) calibration equation:

$$\sum_{k \in S} w_k \mathbf{x}_k = T_\mathbf{x}, \tag{1}$$

which can also be rendered as P univariate "calibration equations:"

$$\sum_{k \in S} w_k x_{pk} = \sum_{k \in U} x_{pk} \quad \text{for} \quad p = 1, \ldots, P.$$

To choose among the sets, Deville and Särndal required that calibration weights minimize a nonnegative distance (or loss) function subject to the w_k satisfying Eq. (1). Such

a function measures the distance between the vectors $(w_1, \ldots, w_n)'$ and $(d_1, \ldots, d_n)'$ in some sense and has its unconstrained minimum at 0 when each $w_k = d_k$. One popular example of a distance function is

$$L(w_1, \ldots, w_n) = \sum_{k \in S} \frac{(w_k - d_k)^2}{c_k d_k}. \tag{2}$$

As with the expansion estimator, the same set of calibration weights can be used no matter what the variable of interest, y_k. When the particular y_k is a linear combination of the components of $\mathbf{x}_k$ for all $k \in U$, say $\mathbf{x}_k\boldsymbol{\beta}$, then t_y^{CAL} equals T_y exactly ($\boldsymbol{\beta}$ is a column vector, as will be all vectors in this chapter not specifically described as row vectors). That is a great strength of calibration weighting and the reason why the calibration estimator is often much more efficient (has a smaller mean squared error) than the expansion estimator.

Another strength of calibration weighting is that $\{w_k | k \in S\}$ and $\{d_k | k \in S\}$ must be close because their difference is in some sense minimized. As a result, with a sufficiently large sample, t_y^{CAL} is close to randomization unbiased no matter what the y-variable as long as reasonable regularity conditions are met.

Because t_y^{CAL} estimates T_y perfectly when $y_k = \mathbf{x}_k\boldsymbol{\beta}$ exactly, it is reasonable to expect t_y^{CAL} to be a good estimator when y_k and $\mathbf{x}_k\boldsymbol{\beta}$ are close. This can be formalized by assuming the y_k are random variables satisfying the linear prediction model:

$$y_k = \mathbf{x}_k\boldsymbol{\beta} + \varepsilon_k, \tag{3}$$

where $E\left(\varepsilon_k | \{\mathbf{x}_g, I_g; g \in U\}\right) = 0$ for all $k \in U$. Under this model, it is easy to see that t_y^{CAL} is an unbiased estimator for T_y in the sense that $E_\varepsilon\left(t_y^{\mathrm{CAL}} - T_y | \{\mathbf{x}_g, I_g; g \in U\}\right) = 0$. (Strictly speaking, t_y^{CAL} is an unbiased predictor for T_y, because the latter is a random variable under the model.) The subscript ε refers to treating the ε_k as random variables. In this context, the I_k are treated as fixed constants.

One practical problem with (prediction) model-based analysis is that we are usually interested in estimating totals for a number of survey variables at the same time. It is often unreasonable to assume that different variables satisfy the same linear model.

This problem can be made to all but disappear. Suppose we had postulated separate models for J different survey variables, $y_{1k}, \ldots, y_{Jk}$:

$$y_{jk} = \mathbf{x}_{jk}\boldsymbol{\beta}_j + \varepsilon_{jk},$$

where $\mathbf{x}_{jk}$ is a P_j-component row vector, and $E\left(\varepsilon_{jk} | \{\mathbf{x}_{1g}, \ldots, \mathbf{x}_{Jg}, I_g; g \in U\}\right) = 0$ for all $k \in U$. It is obvious that the model in Eq. (3) still holds with $\mathbf{x}_k$ now equal to $(\mathbf{x}_{1k}, \ldots, \mathbf{x}_{Jk})$. Duplicated and singular components of $\mathbf{x}_k$ can be pruned with no practical effect on the model (a singular component is a linear combination of other components).

A simple example is the following. Suppose y_{1k} is the current planted corn acres for farm k, and y_{2k} the farm's current planted wheat acres. Several years ago, all the farms in the population provided their annual corn and wheat acres to the Census of Agriculture. Denoting these previous values for farm k as x_{1k} and x_{2k}, respectively, the combined linear model inherent in calibration takes the form:

$$y_{jk} = \begin{pmatrix} 1 & x_{1k} & x_{2k} \end{pmatrix} \begin{pmatrix} \beta_{0j} \\ \beta_{1j} \\ \beta_{2j} \end{pmatrix} + \varepsilon_{jk}$$

for $j = 1$ or 2. Notice that the $\mathbf{x}_k$-vector is common to both the model for corn and wheat. The $\boldsymbol{\beta}$-vector is not. The common $\mathbf{x}_k$-vector allows the creation of a common set of calibration weights for each survey variable.

Calibration has its drawbacks. In the simple farm example, it may be reasonable to assume that β_{12} and β_{21} are zero, in other words, that corn in the census year has no effect on the current amount of planted wheat, and that census-year wheat has no effect on current-year corn. Explicitly assuming $\beta_{12} = \beta_{21} = 0$ would likely increase the efficiency of the resulting estimator. Unfortunately, calibration does not allow us to do that. It is the price we pay for developing a single set of weights for all survey variables.

Poststratification is a form of calibration that preceded Deville and Särndal by decades. It is most often used to adjust for unit nonresponse in the sample or coverage errors in the sampling frame, but in the discussion to immediately follow we assume a perfect frame and complete response.

Suppose the components of $\mathbf{x}_k$ are binary classification variables such that $x_{pk} = 1$ when k is in group (poststratum) p and 0 otherwise. In a human population, for example, we can have $x_{1k} = 1$ and $x_{2k} = 0$ when k is male, and $x_{1k} = 0$ and $x_{2k} = 1$ when k is female.

When each k is in one and only one of the P groups, and the group population sizes, the N_p, are known, a poststratified or group-mean-model estimator performs a simple ratio adjustment in each group, setting $w_k = \left(N_p / \sum_{j \in S} d_j x_{pj}\right) d_k$ when k is in both the sample and in group p. It is easy to see that the calibration equation $\sum_S w_j x_{pj} = N_p$ holds for all p. Moreover, $E_I\left(N_P / \sum_S d_j x_{pj}\right) \approx 1$ for a sufficiently large sample under mild conditions because $E_I\left(\sum_S d_j x_{pj}\right) = N_p$. Thus, $w_k \approx d_k$. The subscript I denotes that the expectation treats the I_k as random variables.

Building on the example aforementioned, suppose $x_{3k} = 1$ when individual k is of African origin, and $x_{3k} = 0$ otherwise. Iterative proportional fitting or raking essentially performs a ratio adjustment for one group at a time, treating the results of the previous ratio adjustment as the $\{d_k\}$. The method recycles through the groups as necessary (in practice four or fewer times) until a set of calibration weights is effectively found; that is, the final weights satisfy the calibration equation within roundoff error. On rare occasions, raking will fail to find a set of final calibration weights.

Deming and Stephan (1940) called raking "a least squares adjustment," but it is not. Nevertheless, it turns out that the calibration weights most often used in practice have the linear form: $w_k = d_k(1 + c_k \mathbf{x}_k \mathbf{g})$ for some vector $\mathbf{g}$ and set of constants $\{c_k | k \in S\}$. These weights result from minimizing the "least-squares" distance function in Eq. (2) subject to the calibration Eq. (1). Deville and Särndal observed that raking weights have a different form: $w_k = d_k \exp(\mathbf{x}_k \mathbf{g})$. Nevertheless, when the $\mathbf{x}_k \mathbf{g}$ are small, these weights are very close to the linear calibration weights, $w_k = d_k(1 + \mathbf{x}_k \mathbf{g})$.

Section 2 develops the asymptotics needed for this chapter. The general framework follows Isaki and Fuller (1982), but with a stronger focus on the relative mean squared error of a calibration estimator. Section 3 discusses the randomization and model-based properties of the so-called generalized regression (GREG) estimator (see, for example, Särndal et al., 1989), which translates into the linear-calibration estimator with $w_k = d_k(1 + c_k \mathbf{x}_k \mathbf{g})$.

In Section 4, we follow Estevao and Särndal (2000) and move away from Deville and Särndal's distance-function-based definition of calibration weighting. Linear

calibration weights have the form: $w_k = d_k(1 + \mathbf{h}_k\mathbf{g})$, where $\mathbf{h}_k$ is a row vector with the same dimension as $\mathbf{x}_k$. In addition to the calibration equation, the sampling design and population values must be such that $\mathbf{h}_k\mathbf{g}$ tends to zero as the sample size grows arbitrarily large.

The purely randomization-based "optimal" calibration estimator proposed by Rao (1994; see also Tillé, 1999) can be put in the form of a linear-calibration estimator. Montanari and Ranalli (2002) introduced the useful concept of a design-balanced variable when treating a randomization-optimal estimator as a calibration estimator.

Section 5 follows Särndal et al. (1989) and Kott (1990, 2005) in developing estimators for the model variance and randomization mean squared error of a linear-calibration estimator simultaneously. It then addresses ways to reduce the model bias of this variance-estimation strategy even further.

Section 6 discusses potentially nonlinear calibration weights of the form: $w_k = d_k f(\mathbf{h}_k\mathbf{g})$. In some applications, however, $f(.)$ may be a truncated linear function, truncated to prevent the calibration weights from being too large or too small. Huang and Fuller (1978) provided an early example of this.

Folsom and Singh (2000) showed how calibration weighting could be used to compensate for unit nonresponse or coverage errors. In Section 7 this quasi-randomization framework, $\rho(\mathbf{h}_k\boldsymbol{\gamma}) = 1/f(\mathbf{h}_k\boldsymbol{\gamma})$ is the probability of element k being covered by the frame or responding to the survey, where $\rho(.)$ is known but the governing parameter $\boldsymbol{\gamma}$ is not. Folsom and Singh also introduced a flexible form for $f(.)$ (and thus $\rho(.)$) called the "general exponential model."

In Folsom and Singh, $\mathbf{h}_k = \mathbf{x}_k$. That was not the case in Lundström and Särndal (1999), but like in Fuller et al. (1994) before it, only linear $f(.)$ were treated. We will follow Kott (2006) and allow a not-necessarily-linear $f(.)$ to be a function of variables other than the benchmark variables.

Section 8 concludes with a brief discussion of other approaches and other issues, many of that are treated elsewhere in this volume.

2. Randomization consistency and other asymptotic properties

The estimator t based on a sample of n elements is said to be a consistent estimator for a finite value, T, when $p\lim_{n\to\infty}(t) = T$. Fuller (1976, Chapter 5) showed that a sufficient condition for consistency is $\lim_{n\to\infty}\{E\left[(t-T)^2\right]\} = 0$. This means that both the bias and the mean squared error of t vanish as the sample size grows arbitrarily large.

The definition of consistency has to be modified when T is a finite-population total. For one thing, the population size, N, needs to grow along with the expected sample size, n. Because T itself will usually also be growing, an estimator t is said to be randomization consistent when its relative error, $(t-T)/T$, has a probability limit of 0 as n grows arbitrarily large (and N along with it). A sufficient condition for randomization consistency is that the relative mean squared error of t, $E[(t-T)^2]/T^2 = E\{[(t-T)/T]^2\}$, has an asymptotic limit of zero.

For convenience, we focus on a single survey variable and assume that all $y_k \geq 0$ and $z_{ak} \geq 0$, where $\mathbf{z}_k = (z_{1k}, \ldots, z_{Qk})$ is a vector of values associated with element k, and $Q \geq P$. Moreover, we assume the sampling design and population are such that as

the population size, N, and expected sample size, n, grow arbitrarily large,

$$0 < L_y \leq \sum_{k\in U} y_k^{\delta}/N \leq B_y < \infty, \qquad \delta = 1, \ldots, 4; \tag{4}$$

$$0 < L_{za} \leq \sum_{k\in U} z_{ak}^{\delta}/N \leq B_{za} < \infty, \quad \delta = 1, \ldots, 4; \text{ for all } a, \tag{5}$$

where $(n/N)\pi_k^{-1}$ is one of the components of $\mathbf{z}_k$. Unlike Isaki and Fuller, we allow the possibility that N grows at an asymptotically faster rate than n. Notice that our framework also allows the realized sample size, n_S, for a particular sample S to be random.

Under the regularity conditions mentioned earlier, it is not hard to show that for sampling designs where $E\left(I_k I_j\right) = \pi_{kj} \leq \pi_k \pi_j$ when $k \neq j$,

$$T_y = O(N), \text{ and}$$

$$\mathrm{Var}_I\left(t_y^E\right) = \sum_{k\in U}\sum_{j\in U}\left(\pi_{kj} - \pi_k\pi_j\right)\frac{y_k}{\pi_k}\frac{y_j}{\pi_j} \leq \sum_{k\in U}\left(\frac{1}{\pi_k} - 1\right)y_k^2 = O\left(N^2/n\right).$$

(by definition, $\pi_{kk} = \pi_k$). The last step makes use of Schwartz's inequality (i.e., $\sum y_k^2/\pi_k \leq \sqrt{\sum y_k^4 \sum 1/\pi_k^2}$). Because the expansion estimator is randomization unbiased, its relative randomization mean squared error is the same as its relative randomization variance, which is $O(1/n)$. Thus, t_y^E is randomization consistent with a relative error of $O_P(1/n^{1/2})$.

The joint selection probabilities in many element sampling plans satisfy $\pi_{kj} \leq \pi_k\pi_j$ whenever $k \neq j$. Simple random sampling, stratified simple random sampling, and Poisson sampling are among them. Asok and Sukhatme (1976) showed that $\pi_{kj} = \frac{n-1}{n}\pi_k\pi_j\,[1 + O(n/N)]$ under Sampford sampling and Goodman–Kish sampling (systematic unequal probability sampling from a randomly order list). Consequently, both sampling plans are in this class when n is sufficiently large and $N \geq O(n^{3/2})$ (stating the last inequality more formally, $\lim_{n\to\infty} n^{3/2}/N = C$, where C is finite and possibly 0).

In many multistage sampling plans, when elements k and j are in the same primary sampling unit (PSU), π_{kj} will usually exceed $\pi_k\pi_j$. To extend asymptotic properties to multistage samples where $\pi_{kj} \geq \pi_k\pi_j$ need hold only when k and j are in different PSUs, we first divide the population into PSUs, and assume that the number of these PSUs, N_1, grows proportionally with N. We similarly assume that the expected number of PSUs in the first-stage sample, n_1, grows proportionally with n. We add the assumption that the individual population size for each PSU i is bounded. Finally, we replace Eqs. (4) and (5) with PSU-level analogues, letting, for example, $t_{y(i)}$ be the sum of the y-values across all the elements in i. Eq. (5) can be replaced by $0 < L_{y'} \leq \sum t_{y(i)}^{\delta}/N_1 \leq B_{y'} < \infty$, where the summation is over the N_1 PSUs. The proof is left to the reader who should note that $\pi_{kj} \leq \max\{\pi_k, \pi_j\}$, which implies $\left(\pi_{kj} - \pi_k\pi_j\right)/\left(\pi_k\pi_j\right) \leq \max\left\{\pi_k^{-1}, \pi_j^{-1}\right\} - 1$.

One common sampling plan that does not lead to randomization consistent estimation is systematic sampling from an ordered list. The problem is that given any element k, the number of other elements j such that $\pi_{kj} > \pi_k\pi_j$ grows at the same rate as the (expected) sample size.

3. The GREG estimator

It is common to call the randomization-consistent regression estimator the "general(ized) regression" or "GREG estimator." For our purposes, it has the form:

$$t_y^{\text{GREG}} = t_y^E + \left(T_{\mathbf{x}} - \sum_{k\in S} d_k \mathbf{x}_k\right)\left(\sum_{k\in S} c_k d_k \mathbf{x}_k' \mathbf{x}_k\right)^{-1} \sum_{k\in S} c_k d_k \mathbf{x}_k' y_k, \tag{6}$$

where $\mathbf{x}_k$ is a row vector composed of components of $\mathbf{z}_k$ in Eq. (5), $d_k = 1/\pi_k$ for $k \in S$ (as before), c_k is also a component of $\mathbf{z}_k$, which may or may not be a function of $\mathbf{x}_k$, and $\lim_{N\to\infty} \sum_U c_k \mathbf{x}_k' \mathbf{x}_k / N = \mathbf{\Phi}$ is positive definite matrix. This last condition means that $\sum_S c_k d_k \mathbf{x}_k' \mathbf{x}_k$ will usually be invertible in practice. We will assume that it is always invertible for convenience.

Sometimes the c_k within Eq. (6) are assumed to be proportional to the inverses of $E\left(\varepsilon_k^2\right)$. We do not make that assumption here.

Let $\mathbf{b} = \left(\sum_S c_k d_k \mathbf{x}_k' \mathbf{x}_k\right)^{-1} \sum_S c_k d_k \mathbf{x}_k' y_k$, and $\mathbf{B} = \left(\sum_U c_k \mathbf{x}_k' \mathbf{x}_k\right)^{-1} \sum_U c_k \mathbf{x}_k' y_k$. The GREG estimator can be written as $t_y^{\text{GREG}} = t_y^E + \left(T_{\mathbf{x}} - \sum_S d_k \mathbf{x}_k\right)\mathbf{b}$, which is close to the idealized general difference estimator:

$$t_y^{\text{GDIF}} = t_y^E + \left(\sum_{k\in U} \mathbf{x}_k \mathbf{B} - \sum_{k\in U} d_k \mathbf{x}_k \mathbf{B}\right),$$

where $\mathbf{x}_k\mathbf{B}$ (which cannot be computed with survey data only) plays the role of the scalar x_k in the standard difference estimator. The general-difference estimator is randomization unbiased.

The GREG estimator in Eq. (6) can be rewritten in calibration form as $t_y^{\text{GREG}} = \sum_S w_k y_k$, where

$$w_k = d_k + \left(T_{\mathbf{x}} - \sum_{j\in S} d_j \mathbf{x}_j\right)\left(\sum_{j\in S} c_j d_j \mathbf{x}_j' \mathbf{x}_j\right)^{-1} c_k d_k \mathbf{x}_k' \tag{7}$$

Strictly speaking, the w_k are functions of the realized sample, S, and the c_k, but we suppress that in the notation for convenience.

The most common benchmark variables in practice are group-membership indicators. Let $u_{pk} = 1$ when element k is in group p, and 0 otherwise. Similarly, let $u_k = 1$ for all elements in the population. When the groups are exhaustive (every element is in some group) and mutually exclusive, $\mathbf{x}_k = (u_{1k}, \ldots, u_{Pk})$, and all $c_k = 1$, the group-mean-model estimator results with (as noted in the introduction) $w_k = \left(N_p / \sum_S d_j u_{pj}\right) d_k$ when k is in group p.

In this example, each calibration weight must be positive, although it is possible for some calibration weights to be to less than unity, especially when certainty selections (element k is a certainty selection when $\pi_k = 1$) are grouped with noncertainties. More generally, if the benchmark variables are not all mutually exclusive group-membership indicators, then there is no guarantee that every w_k will be nonnegative. Computer packages often cannot handle negative weights.

Many find less-than-unity calibration weights troubling. An element with such a weight does not appear to fully represent itself. This can be particularly irksome when that element has a positive value for some y-variable, whereas most of the other sampled

elements have zero values. We will return to the issue of negative and less-than-unity calibration weights in Section 3.4.

3.1. The randomization-based properties of the GREG estimator

Let us assume that the regularity conditions and sample plan are such that $t_y^E - T_y = O_P\left(N/n^{1/2}\right)$, $\sum_S d_k \mathbf{x}_k - T_\mathbf{x} = \mathbf{O}_P\left(N/n^{1/2}\right)$ and $\sum_S c_k d_k \mathbf{x}_k' \mathbf{f}_k - \sum_U c_k \mathbf{x}_k' \mathbf{f}_k = \mathbf{O}_P\left(N/n^{1/2}\right)$, where $\mathbf{f}_k$ can be $\mathbf{x}_k$ or y_k. Define $e_k = y_k - \mathbf{x}_k\mathbf{B} = y_k - \mathbf{x}_k\left(\sum_U c_j \mathbf{x}_j' \mathbf{x}_j\right)^{-1} \sum_U c_j \mathbf{x}_j' y_j$ so that $\sum_U c_k \mathbf{x}_k' e_k = 0$. This equality makes $\sum_S c_k d_k \mathbf{x}_k' e_k = \mathbf{O}_P\left(N/n^{1/2}\right)$. As a result, we can express the error of t_y^{GREG} as

$$
\begin{aligned}
t_y^{\text{GREG}} - T_y &= \sum_{k\in S} w_k y_k - \sum_{k\in U} y_k \\
&= \sum_{k\in S} w_k e_k - \sum_{k\in U} e_k \\
&= \sum_{k\in S} d_k e_k - \left(T_\mathbf{x} - \sum_{k\in S} d_k \mathbf{x}_k\right)\left(\sum_{k\in S} c_k d_k \mathbf{x}_k' \mathbf{x}_k\right)^{-1} \\
&\quad \times \sum_{k\in S} c_k d_k \mathbf{x}_k' e_k - \sum_{k\in U} e_k \\
&= \sum_{k\in S} d_k e_k - \sum_{k\in U} e_k + O_P\left(N/n\right).
\end{aligned}
\tag{8}
$$

Because $|e_k| \le y_k + |\mathbf{x}_k\mathbf{B}|$, it is not hard to see the GREG estimator is randomization consistent with a relative randomization bias and mean squared error of asymptotic order $1/n$. The randomization bias is an asymptotically insignificant contributor to the mean squared error, mse, when $p\lim_{n\to\infty}\left(n \cdot \text{mse}/N^2\right) > 0$, a mild condition violated when nearly all the e_k in the population are zero, which we assume not to be the case for convenience.

3.2. Model-based properties of the GREG estimator

Suppose the y_k are random variables that satisfy the linear model in Eq. (3). In addition, assume $E\left(\varepsilon_k|\{\mathbf{x}_g, I_g; g \in U\}\right) = E\left(\varepsilon_k\varepsilon_j|\{\mathbf{x}_g, I_g; g \in U\}\right) = 0$ for $k \neq j$, and $E\left(\varepsilon_k^2|\{\mathbf{x}_g, I_g; g \in U\}\right) = \sigma_k^2 < \infty$. The σ_k^2 need not be known. Moreover, there is no reason that I_g cannot be a function of the components of $\mathbf{z}_g$.

It is easy to see that as long as the regression weights satisfy the calibration equation, $\sum_S w_k \mathbf{x}_k = T_\mathbf{x}$, t_y^{GREG} will be model unbiased. Its model variance, as well as the model variance of any calibration estimator, is (suppressing the conditioning on $\mathbf{x}_g$ and I_g for notational convenience)

$$
\begin{aligned}
E_\varepsilon\left[\left(t_y^{\text{GREG}} - T_y\right)^2\right] &= E_\varepsilon\left[\left(\sum_{k\in S} w_k\varepsilon_k - \sum_{k\in U}\varepsilon_k\right)^2\right] \\
&= \sum_{k\in S} w_k^2\sigma_k^2 - 2\sum_{k\in S} w_k\sigma_k^2 + \sum_{k\in U}\sigma_k^2
\end{aligned}
$$

$$
\begin{aligned}
&= \sum_{k\in S} w_k^2\sigma_k^2 - \sum_{k\in S} w_k\sigma_k^2 - \left(\sum_{k\in S} w_k\sigma_k^2 - \sum_{k\in U}\sigma_k^2\right) \\
&= \sum_{k\in S} w_k^2\sigma_k^2 - \sum_{k\in S} w_k\sigma_k^2 + O_P\left(N/n^{1/2}\right),
\end{aligned}
\tag{9}
$$

under mild conditions, in particular, those calibration estimators where $w_k = d_k\left[1 + O_P\left(1/n^{1/2}\right)\right]$, and $\sum_S d_k\sigma_k^2 - \sum_U \sigma_k^2 = O_P\left(N/n^{1/2}\right)$. Notice that we are using randomization-based asymptotic results in a model-based context. We are not, however, averaging over all possible samples, which is what randomization-based theory routinely does.

If σ_k^2 has the form $\mathbf{x}_k\zeta$ for some not-necessarily-specified vector ζ, then $\sum_S w_k\sigma_k^2 = \sum_S w_k\mathbf{x}_k\zeta = \sum_U \mathbf{x}_k\zeta = \sum_U \sigma_k^2$, and the model variance of t_y^{GREG} collapses to $\sum_S\left(w_k^2 - w_k\right)\sigma_k^2$ exactly. Whether or not σ_k^2 can be expressed as $\mathbf{x}_k\zeta$, when $N \geq O(n^{3/2})$ and $\pi_k = O(n/N) < 1$, the model variance is dominated by $\sum_S w_k^2\sigma_k^2$.

For a multistage sample it makes sense to allow the possibility that ε_k and ε_j are correlated when k and j are in the same PSU, but not otherwise. Under the regularity conditions discussed previously for a multistage sample, if $\pi_{kj} \leq \pi_k\pi_j$ for j and k from different PSUs and $N \geq O(n^2)$, it is not hard to show that the model variance of the GREG estimator is dominated by $\sum_{i\in S_1} E_\varepsilon\left[\left(\sum_{k\in S_i} w_k\varepsilon_k\right)^2\right]$, where S_i is the set of sampled elements in PSU i and S_1 is the set of PSUs selected for the sample.

3.3. The anticipated variance

Let us return to the model with no correlation among the elements. The model variance of t_y^{GREG} is $O_P(N^2/n)$ under mild conditions we assume to hold. If we are willing to drop $O_P(N^2/n^{3/2})$ terms (so that $w_k \approx 1/\pi_k$ and $\sum_S d_k\sigma_k^2 - \sum_U \sigma_k^2 \approx 0$), the model variance of t_y^{GREG} can be approximated by $E_\varepsilon\left[\left(t_y^{\mathrm{GREG}} - T\right)^2\right] \approx \sum_S\left(\sigma_k^2/\pi_k^2\right)(1-\pi_k)$.

The randomization expectation of the model variance of t_y^{GREG} is then

$$
E_I\left\{E_\varepsilon\left[\left(t_y^{\mathrm{GREG}} - T\right)^2\right]\right\} \approx \sum_{k\in U}\frac{\sigma_k^2}{\pi_k}(1-\pi_k). \tag{10}
$$

The right-hand side of Eq. (10) was called the asymptotic "anticipated variance" of the GREG by Isaki and Fuller (1982), although the equation goes back considerably further in the literature and "anticipated mean squared error" would have been better. They used it to mean $E_\varepsilon\left\{E_I\left[\left(t_y^{\mathrm{GREG}} - T\right)^2\right]\right\}$, the randomization mean squared error anticipated under the model. The expectation operators can be switched because $\left(t_y^{\mathrm{GREG}} - T\right)^2$ exists and is bounded under the assumptions in Eqs. (4) and (5).

Notice that the joint selection probabilities have no effect on the asymptotic anticipated variance expressed by the right-hand side of Eq. (10). Similarly, the choice for c_k does not matter in this context.

Given an expected sample size n, one can find a set of selection probabilities minimizing the asymptotic anticipated variance of t_y^{GREG} subject to $n = E\left(\sum_U I_k\right) = \sum_U \pi_k$ by solving a Lagrangian. The solution has the form, $\pi_k = n\sigma_k\big/\sum_U \sigma_j$, provided that

this value is bounded by 1 for all k. The asymptotic-anticipated-variance-minimizing π_k are called "Brewer-selection probabilities" (see Brewer, 1963). Brewer-selection probabilities also result from minimizing the expected sample size given a target asymptotic anticipated variance.

Given a vector of survey variables of interest, each with its own target asymptotic anticipated variance, a variant of Chromy's (1987) method can be used to minimize the (expected) sample size, provided (again) that no optimal $\pi_k > 1$. A suboptimal approach called "maximal Brewer selection" simply computes a univariate Brewer selection probability for each survey variable and takes the maximum of those values. See Kott and Bailey (2000). It should be noted that, unlike Brewer selection, maximal Brewer selection is not designed to limit asymptotic anticipated variances given an expected sample size.

3.4. An example

The National Agricultural Statistics Service of the U.S. Department of Agriculture (USDA) uses Poisson sampling and maximal Brewer selection to draw state samples for the June Agricultural survey. We will focus on one example. For the June 2005 survey in Pennsylvania, 1436 names were selected from the USDA list of 25,935 potential agricultural places.

USDA used the same set of 13 variables both for determining the Poisson selection probabilities and as benchmarks for calibration weighting in Pennsylvania. The element values for these variables were constructed from previous survey information. They and their statewide population totals are displayed in Table 1.

Table 2 displays some summary statistics about the calibration weights for three alternative choices for c_k in Eq. (7): $c_k = 1$, $c_k = 1 - \pi_k$, and $c_k = (1 - \pi_k)/\pi_k$. Following a suggestion in Brewer (1994), the USDA sets each c_k to $1 - \pi_k$ rather than the more common setting of unity to limit the number of calibration weights less than 1.

Table 1
List of benchmark variables for the 2006 June Agricultural Survey in Pennsylvania

Benchmark Variable	Frame Total
Number	25,935 names
Alfalfa	707,466 acres
Barley	81,458 acres
Calculated cropland	2,102,285 acres
Storage capacity	110,467,674 bushels
Corn	1,290,289 acres
Reported cropland	4,752,217 acres
Oats	176,338 acres
Other hay	1,010,729 acres
Rye	76,455 acres
Soybeans	390,659 acres
Sorghum	16,175 acres
Winter wheat	166,200 acres

Table 2
Comparing the calibration weights of GREGs with different values for the c_k

	$c_k = 1$	$c_k = 1 - \pi_k$	$c_k = \frac{1-\pi_k}{\pi_k}$
$\frac{\sum_S \lvert w_k - d_k \rvert}{\sum_S d_k}$	0.0408	0.0433	0.0675
$\frac{\text{Max}\{w_k\}}{\text{Max}\{d_k\}}$	1.0500	1.0560	1.2416
$\text{Max}\left\{\frac{w_k}{d_k}\right\}$	1.5290	1.4166	1.3486
$\text{Min}\{w_k\}$	−1.0846	0.6132	1.0000
$\text{Min}\left\{\frac{w_k}{d_k}\right\}$	−0.0587	0.3321	0.6496
Percent of $w_k < 1$ (number)	0.6964 (10)	0.0696 (1)	0.0000
Percent of $w_k \leq 0$ (number)	0.0696 (1)	0.0000	0.0000
$\frac{\sum_S w_k^2 - \sum_S w_k}{\sum_S d_k^2 - \sum_S w_k}$	1.0748	1.0797	1.1834
$\frac{\sum_S w_k^2 x_k - \sum_S w_k x_k}{\sum_S d_k^2 x_k - \sum_S w_k x_k}$	1.0357	1.0332	1.0110

Note: x_k, calculated cropland.

Using $c_k = (1 - \pi_k)/\pi_k$ would have resulted in "randomization-optimal" calibration, as we will see in Section 4.2.

The original sampling weights (the d_k) ranged from 1 to 250. There were 44 certainty selections. The first row of the table provides a measure of the average absolute change in the weights from d_k to w_k. Even though there were 13 benchmark variables subject to calibration and a random sample size, the average change was less than 7% no matter which method was used, with the average change attaining its minimum, close to 4%, when c_k equaled 1. The other values for c_k did a better job keeping the w_k above both 0 and 1.

The maximum calibration weight occurred when c_k equaled $(1 - \pi_k)/\pi_k$, which was still less than 25% higher than the largest original weight. The largest upward calibration adjustment (w_k/d_k), slightly less than 1.53, occurred when c_k equaled 1. It appears that setting $c_k = 1$ did a better job controlling the number of elements with larger-than-average calibration weights, whereas the setting $c_k = (1 - \pi_k)/\pi_k$ did a better job controlling the number of elements with larger calibration adjustments.

The last two rows of the table provide a relative measure of the model variances under the assumption that the σ_k^2 were equal or, alternatively, that they were proportional to calculated cropland – the sum of the control acres for all survey crops except hay, a benchmark variable that serves here as an omnibus size measure (recall that both unity and calculated cropland were calibration variables). Each of the relative-model-variance measures are asymptotically unity and thus independent of the choice for c_k. In the finite world of the June Agricultural Survey in Pennsylvania, however, setting the c_k equal to 1 minimized the model variance among the three choices when the σ_k^2 were assumed equal, whereas setting the c_k equal to $(1 - \pi_k)/\pi_k$ was the best of the three when the σ_k^2 were assumed proportional to calculated cropland.

4. Redefining calibration weights

4.1. Instrumental variables

In their original definition of calibration weights, Deville and Särndal (1992) required that the set of calibration weights, $\{w_k|k \in S\}$ minimize some distance function between the members of the set and the original sampling weights, the d_k, subject to satisfying the calibration equation. As a result, the calibration estimator, $t_y^{\mathrm{CAL}} = \sum_S w_k y_k$, was both unbiased under the model in Eq. (3) and usually randomization consistent.

We can remove the limitation that the calibration weights minimize a distance function, and require only that the w_k need satisfy the calibration equation and be of the functional form:

$$w_k = d_k(1 + \mathbf{h}_k\mathbf{g}), \tag{11}$$

where $\mathbf{h}_k = (h_{1k}, \ldots h_{Pk})$ is a row vector such that $\sum_S d_k\mathbf{h}_k'\mathbf{x}_k$ is invertible, and $\mathbf{g}$ is a column vector of the same dimension $\mathbf{h}_k$. This is a generalization of the GREG where $\mathbf{h}_k$ effectively replaces $c_k\mathbf{x}_k$.

It is not hard to see that $\mathbf{g} = \left(\sum_S d_k\mathbf{x}_k'\mathbf{h}_k\right)^{-1}\left(T_\mathbf{x} - \sum_S d_k\mathbf{x}_k\right)'$. Moreover, if the $|h_{pk}|$ are components of $\mathbf{z}_k$ in Eq. (4), the regularity conditions hold, and $\sum_S d_k\mathbf{h}_k'\mathbf{x}_k/N$ is invertible both for the realized N and in the probability limit, then $t_y^{\mathrm{CAL}} = \sum_S w_k y_k = \sum_S d_k y_k + \left(T_\mathbf{x} - \sum_S d_k\mathbf{x}_k\right)\left(\sum_S d_k\mathbf{h}_k'\mathbf{x}_k\right)^{-1}\sum_S d_k\mathbf{h}_k' y_k$ is randomization consistent whenever t_y^E is.

This suggests an alternative definition of calibration weights: a set of weights, $\{w_k|k \in S\}$, such that

(1) The w_k satisfy the calibration equation $\sum_S w_k\mathbf{x}_k = \sum_U \mathbf{x}_k$, and,
(2) $t_y^{\mathrm{CAL}} = \sum_S w_k y_k$ is randomization consistent whenever t_y^E is under mild conditions.

That is the definition we will use.

Calibrations estimators with weights satisfying Eq. (11) will be called "linear calibration estimators." The components of $\mathbf{h}_k$ that are not linear combinations of the benchmark variables (i.e., the components of $\mathbf{x}_k$) are sometimes called "instrumental variables" (for example, see Brewer, 1995).

When each h_{pk} is a function of the $\mathbf{x}_g$ ($g \in U$), then the linear calibration estimator is unbiased under the prediction model in Eq. (3) as long as $E\left(\varepsilon_k|\{\mathbf{x}_g, I_g; g \in U\}\right) = 0$ for all $k \in U$, which is what we assumed before. Otherwise, we may need to further assume $E\left(\varepsilon_k|\{\mathbf{h}_g; g \in U\}\right) = 0$ to establish the model unbiasedness of the calibration estimator.

A linear calibration estimator can be put in projection form, $t_y^{\mathrm{CAL}} = T_\mathbf{x}\mathbf{b}_\mathbf{h}$, where

$$\mathbf{b}_\mathbf{h} = \left(\sum_{k\in S} d_k\mathbf{h}'\mathbf{x}_k\right)^{-1}\sum_{k\in S} d_k\mathbf{h}' y_k,$$

when $\sum_S d_k y_k - \sum_S d_k\mathbf{x}_k\left(\sum_S d_k\mathbf{h}_k'\mathbf{x}_k\right)^{-1}\sum_S d_k\mathbf{h}_k' y_k = 0$. This will happen when there is a vector $\boldsymbol{\theta}$ such that $h_k\boldsymbol{\theta} = \boldsymbol{\theta}'\mathbf{h}_k' = 1$ for all $k \in S$ (to see why, rewrite $\sum_S d_k\mathbf{x}_k$ in $\sum_S d_k y_k - \sum_S d_k\mathbf{x}_k\left(\sum_S d_k\mathbf{h}_k'\mathbf{x}_k\right)^{-1}\sum_S d_k\mathbf{h}_k' y_k$ as $\sum_S d_k\boldsymbol{\theta}'\mathbf{h}'\mathbf{x}_k$); that is to say, when

a linear combination of the components of $\mathbf{h}_k$ (or a single component) is unity. For example, in the ratio estimator, $t_y^{\mathrm{CAL}} = T_x \left(\sum_S d_k x_k\right)^{-1} \sum_S d_k y_k = T_x \left(t_y^E / t_x^E\right)$, $\mathbf{x}_k$ is the scalar $x_k \geq 0$ with at least one positive value in the sample, and $\mathbf{h}_k = 1$. The calibration weight for element k can be rendered $w_k = d_k \left(T_x / \sum_S d_i x_i\right)$. The ratio estimator can also be expressed as a GREG with $c_k = 1/x_k$, but only when all the $\mathbf{x}_k$ in the sample are positive.

A popular extension of the ratio estimator is the group-ratio-model estimator in which $\mathbf{x}_k = x_k(u_{1k}, \ldots, u_{Pk})$, where the u_{pk} are group-membership indicators, the P groups are exhaustive and mutually exclusive, and x_k is a nonnegative scalar. If $T_{\mathbf{x}}$ is known and there is a positive sampled x-value in each group, then setting $\mathbf{h}_k = (u_{1k}, \ldots, u_{Pk})$ yields an estimator expressible in projection form with $w_k = d_k \left(\sum_{U \cap p} d_i x_i / \sum_{S \cap p} d_i x_i\right)$ for each k in group p. When the groups are design strata, the group-ratio-model estimator is also called a "separate ratio estimator."

Another common example of a linear calibration estimator expressible in projection form is the GREG estimator with $c_k = 1$ and $u_k = 1$ as a component of $\mathbf{x}_k$.

A linear calibration estimator can be put in prediction form, $t_y^{\mathrm{CAL}} = \sum_S y_k + \sum_{U-S} x_k \mathbf{b}_{\mathbf{h}}$, when $\sum_S (d_k - 1)\, y_k - \sum_S (d_k - 1)\, \mathbf{x}_k \left(\sum_S d_k \mathbf{h}_k' \mathbf{x}_k\right)^{-1} \sum_S d_k \mathbf{h}_k' y_k = 0$. This will happen when there is a vector $\boldsymbol{\theta}$ such that $\mathbf{h}_k \boldsymbol{\theta} = \boldsymbol{\theta}' \mathbf{h}_k' = (d_k - 1)/d_k = 1 - \pi_k$ for all $k \in S$ (to see why, replace $\sum_S (d_k - 1)\, \mathbf{x}_k$ in $\sum_S (d_k - 1)\, y_k - \sum_S (d_k - 1)\, \mathbf{x}_k \left(\sum_S d_k \mathbf{h}_k' \mathbf{x}_k\right)^{-1} \sum_S d_k \mathbf{h}_k' y_k$ with $\sum_S (d_k - 1)\, \boldsymbol{\theta}' \mathbf{h}' \mathbf{x}_k \left[d_k / (d_k - 1)\right]$). This implies that a linear combination of the components of $\mathbf{h}_k$ (or a single component) is $1 - \pi_k$. "Prediction form" gets its name because effectively the y-value for each population element k not in the sample is predicted by $\mathbf{x}_k \mathbf{b}_{\mathbf{h}}$.

When $\mathbf{x}_k$ is the scalar $x_k \geq 0$ and $\mathbf{h}_k$ the scalar $1 - \pi_k$, the linear calibration weight for element k can be rendered $w_k = 1 + (d_k - 1)\left[\sum_{U-S} x_i \Big/ \sum_S (d_i - 1)\, x_i\right]$, which is never less than unity if at least one sampled k has a positive $(d_k - 1)\, x_k$ value.

In a multivariate setting, if u_k is a component of $\mathbf{x}_k$, then setting $\mathbf{h}_k$ equal to $(1 - \pi_k)\, \mathbf{x}_k$ results in a linear calibration estimator expressible in prediction form. This is the same as the GREG estimator with $c_k = 1 - \pi_k$.

4.2. *Randomization-optimal calibration*

Consider the following possibilities for $\mathbf{h}_k$ in Eq. (11):

$$\mathbf{h}_{(1)k} = \sum_{j \in S} \frac{\left(\pi_{kj} - \pi_k \pi_j\right)}{\pi_{kj} \pi_j} \mathbf{x}_j, \quad \text{and}$$

$$\mathbf{h}_{(2)k} = \sum_{j \in U} \frac{\left(\pi_{kj} - \pi_k \pi_j\right)}{\pi_k \pi_j} \mathbf{x}_j \tag{12}$$

Under many designs, $\sum_S d_k \mathbf{h}_{(m)k}' \mathbf{x}_k$ is a randomization consistent estimator for $\mathbf{Var}_I \left(t_{\mathbf{x}}^E\right)$ when $m = 1$ or 2. Moreover, using either variable, t_y^{CAL} is asymptotically identical to the optimal difference estimator:

$$t_y^{\mathrm{ODIF}} = t_y^E + \left(T_{\mathbf{x}} - t_{\mathbf{x}}^E\right) \left[\mathbf{Var}_I \left(t_{\mathbf{x}}^E\right)\right]^{-1} \mathbf{Cov}_I \left(t_{\mathbf{x}}^E, t_y^E\right),$$

that is, the estimator that minimizes the randomization variance of $t_y^E + \left(T_{\mathbf{x}} - t_{\mathbf{x}}^E\right)\mathbf{b}$ for some fixed $\mathbf{b}$. We will call the version of t_y^{CAL} using either $\mathbf{h}_{(1)k}$ or $\mathbf{h}_{(2)k}$ in Eq. (12) a "randomization-optimal" calibration estimator, with the former denoted ROCE1 and the latter ROCE2.

Observe that when $\pi_{kj} = \pi_k \pi_j$, both $\mathbf{h}_{(1)k}$ and $\mathbf{h}_{(2)k}$ collapse to $[(1-\pi_k)/\pi_k]\,\mathbf{x}_k$. Thus, setting c_k in Eqs. (6) or (7) to $(1-\pi_k)/\pi_k$ produces a randomization-optimal calibration estimator under Poisson sampling.

There are two problems with the randomization-optimal calibration estimator under a more general sampling design. First, $\mathbf{Var}_I\left(t_{\mathbf{x}}^E\right)$ will be singular when the sampling design is such that a component x_{pk} of $\mathbf{x}_k$ is design balanced; that is, $\sum_S x_{pk}/\pi_k = \sum_U x_{pk}$. Any such component has to be removed from $\mathbf{x}_k$. Similarly, if a linear combination of components of $\mathbf{x}_k$ are design balanced, then (at least) one of the components must be removed from $\mathbf{x}_k$. Second, each $\mathbf{h}_{(m)k}$ may change as the sample and population grow arbitrarily large. Consequently, the regularity conditions in Eq. (4) cannot be made directly relevant for such a variable.

We can flesh out these issues with two examples. Consider first a probability proportional to size sampling scheme with $\pi_k = n\,x_k/\sum_U x_j \leq 1$ for all k. Suppose $\mathbf{x}_k$ is the lone component of $\mathbf{x}_k$. If the sampling design has a fixed sample size n, then x_k is a design-balanced variable. After it is removed from $\mathbf{x}_k$, the randomization-optimal calibration estimator collapses into the expansion estimator: $t_y^E = \sum_S y_k/\pi_k = n^{-1}\sum_S y_k/x_k$, which is also called "the mean of ratios."

Note that under Poisson sampling, which has a random sample size, the randomization-optimal calibration estimator is

$$t_y^{\text{RO}} = \sum_{k\in S}\frac{y_k}{\pi_k} + \left(T_{\mathbf{x}} - \sum_{k\in S}\frac{x_k}{\pi_k}\right)\left[\sum_{k\in S}\frac{x_k^2}{\pi_k^2}(1-\pi_k)\right]^{-1}\sum_{k\in S}\frac{y_k x_k}{\pi_k^2}(1-\pi_k)$$

$$= \sum_{k\in S}\frac{y_k}{\pi_k} + T_{\mathbf{x}}\left(1-\frac{n_S}{n}\right)\frac{\sum_{k\in S}\frac{y_k}{x_k}(1-\pi_k)}{\sum_{k\in S}(1-\pi_k)}.$$

This is *not* identical to the more widely-used ratio/mean-of-ratios estimator,

$$t_y^{\text{RATIO}} = \sum_{k\in S}\frac{y_k}{\pi_k} + \left(T_{\mathbf{x}} - \sum_{k\in S}\frac{x_k}{\pi_k}\right)\frac{\sum_{k\in S}\frac{y_k}{\pi_k}}{\sum_{k\in S}\frac{x_k}{\pi_k}} = T_{\mathbf{x}}\frac{\sum_{k\in S}\frac{y_k}{\pi_k}}{\sum_{k\in S}\frac{x_k}{\pi_k}} = \frac{1}{n_S}\sum_{k\in S}\frac{y_k}{x_k}.$$

Next consider a stratified simple random sample with $S_\alpha, U_\alpha, n_\alpha$, and N_α, denoting respectively the sample, population, sample size, and population size in stratum α, and A the number of strata. Equation (12) becomes

$$\mathbf{h}_{(1)k} = \sum_{\alpha=1}^{A}\frac{N_\alpha(\mathbf{x}_k - \overline{\mathbf{x}}_{\alpha S})}{n_\alpha - 1}, \quad \text{where } \overline{\mathbf{x}}_{\alpha S} = \frac{1}{n_\alpha}\sum_{j\in S_\alpha}\mathbf{x}_j, \quad \text{and}$$

$$\mathbf{h}_{(2)k} = \sum_{\alpha=1}^{A}\frac{N_\alpha^2(\mathbf{x}_k - \overline{\mathbf{x}}_{\alpha U})}{N_\alpha - 1}, \quad \text{where } \overline{\mathbf{x}}_{\alpha U} = \frac{1}{N_\alpha}\sum_{j\in U_\alpha}\mathbf{x}_j.$$

Note that one can compute $\mathbf{h}_{(2)k}$, but not $\mathbf{h}_{(1)k}$, when $N_\alpha > n_\alpha = 1$ and $k \in S_\alpha$. Nevertheless, $\mathbf{h}_{(1)k}$ (and ROCE1) has been used more often in practice.

It is easy to see that each stratum-membership indicator, $u_{\alpha k}$, is design balanced. There are thus up to A such linearly independent components that will need to be removed from the $\mathbf{x}$-vector, one for each stratum.

For a randomization-optimal calibration estimator to be randomization consistent whenever t_y^E is, we can assume the regularity conditions in Eq. (4) as before and add that $\sum_S d_k \mathbf{h}'_{(m)k} \mathbf{x}_k / N$ ($m = 1$ or 2) is invertible both for the realized N and in the probability limit. In addition, when A is fixed as the sample and population sizes grow arbitrarily large, we assume that the stratum population means of the $\mathbf{x}$-vector implicit in the computation of $\mathbf{h}_{(2)k}$ converges to a vector of positive constants that are components of $\mathbf{z}_k$.

Let $\mathbf{x}_k$ denote the original P-vector benchmark variables, $\tilde{\mathbf{x}}_k$ the $\tilde{P}$-vector with all design-balanced benchmark-variables removed, and $\tilde{\tilde{\mathbf{x}}}_k$ the $\tilde{\tilde{P}}$-vector including all components in $\tilde{\mathbf{x}}_k$ as well as the A stratum indicators. When every $n_\alpha \geq 2$, we can rewrite the calibration weights for ROCE1,

$$w_k = d_k + \left(T_{\mathbf{x}} - \sum_{i \in S} d_i \tilde{\mathbf{x}}_i \right) \left(\sum_{i \in S} d_i \mathbf{h}'_{(1)i} \tilde{\mathbf{x}}_i \right)^{-1} d_k \mathbf{h}'_{(1)k} \tilde{\mathbf{x}}'_k,$$

as

$$w_k = d_k + \left(T_{\mathbf{x}} - \sum_{i \in S} d_i \tilde{\tilde{\mathbf{x}}}_i \right) \left(\sum_{i \in S} c_i d_i \tilde{\tilde{\mathbf{x}}}'_i \tilde{\tilde{\mathbf{x}}}_i \right)^{-1} c_k d_k \tilde{\tilde{\mathbf{x}}}'_k,$$

where $c_k = [n_\alpha / (n_\alpha - 1)] (1 - \pi_k) / \pi_k = [N_\alpha / (n_\alpha - 1)] [1 - (n_\alpha / N_\alpha)]$ for $k \in S_\alpha$. With this in mind (and assuming the n_α were large), Bankier (2002) calls using $c_k = (1 - \pi_k) / \pi_k$ "pseudo [randomization] optimal" calibration weighting.

Despite its name, Montanari and Ranalli (2002) shows that the randomization-optimal calibration estimator employing these weights does not always have the least empirical mean squared error among calibration estimators based on the benchmark variables in $\tilde{\mathbf{x}}_k$ and some combinations of the stratum-indicator variables. The problem is that the optimality of a randomization-optimal calibration estimator is asymptotic. In a finite world, effectively adding a dummy variable to the model for every stratum can be wasteful.

From a purely randomization point of view, $\tilde{\mathbf{b}}_{\mathbf{h}(1)} = \left(\sum_S d_k \mathbf{h}'_{(1)k} \tilde{\mathbf{x}}_k \right)^{-1} \sum_S d_k \mathbf{h}'_{(1)k} y_k$ may be a consistent estimator for $\tilde{\mathbf{B}}_{\mathbf{h}} = \left(\sum_U \mathbf{h}'_{(2)k} \tilde{\mathbf{x}}_k \right)^{-1} \sum_U \mathbf{h}'_{(2)k} y_k = \left[\mathbf{Var}_I \left(t_{\tilde{\mathbf{x}}}^E \right) \right]^{-1} \mathbf{Cov}_I \left(t_{\tilde{\mathbf{x}}}^E, t_y^E \right)$, but it is not $\tilde{\mathbf{B}}_{\mathbf{h}}$ itself. Furthermore, $\tilde{\tilde{\mathbf{b}}} = \left(\sum_S d_k \tilde{\tilde{\mathbf{x}}}'_k \tilde{\tilde{\mathbf{x}}}_k \right)^{-1} \sum_S d_k \tilde{\tilde{\mathbf{x}}}'_k y_k$ is not a randomization consistent estimator for $\tilde{\tilde{\mathbf{B}}} = \left(\sum_U \tilde{\tilde{\mathbf{x}}}'_k \tilde{\tilde{\mathbf{x}}}_k \right)^{-1} \sum_U \tilde{\tilde{\mathbf{x}}}'_k y_k$ when the population mean for one or more stratum-indicator variable approaches 0 as the population grows arbitrarily large, violating a regularity condition in Eq. (5).

Why should anyone be concerned about the population mean for a stratum-indicator variable tending toward 0 as the population grows large? Because that is the sensible way

to set up the asymptotics when there is deep stratification: many strata and few sampled units per stratum. The sample sizes within strata stay fixed as the overall sample size and the number of strata grow. Consequently, $\bar{y}_{\alpha S}$, a component of $\sum_S d_k \tilde{\tilde{\mathbf{x}}}' y_k$, does *not* converge to $\bar{y}_{\alpha U}$ as the sample grows arbitrarily large.

When n_α is small, $u_{\alpha k}$ should not be treated as a benchmark variable in calibration, although there may still be some potential gains from pseudo randomization optimal calibration weighting. What "small" means depends on the population and variable of interest. Popular rules of thumb range from under 6 to under 20. Even when "nuisance" strata are stripped of any effect on calibration weighting, they still play a part in variance estimation, as we shall see.

5. Variance estimation

We saw in the last section that under mild conditions the properties of the GREG extend to more general calibration estimators such as those having weights expressible in the form of Eq. (11). These properties include the anticipated variance of an estimation strategy (an estimator coupled with a sampling design) employing linear-calibration weighting. The value of anticipated variance as a measure of a strategy's accuracy diminishes after the sample has been drawn, however. If the model is correct, then the model variance should be estimated given the sample actually drawn rather than averaged over all possible samples. If the model fails, only the randomization mean squared error is relevant.

Suppose the model in Eq. (3) holds, and the element errors are uncorrelated with $E\left(\varepsilon_k^2\right) = \sigma_k^2$, then Eq. (11) tells us that under certain conditions (when either $N \geq O(n^{3/2})$ or σ_k^2 has the form $\mathbf{x}_k\zeta$ for some ζ), the model variance of an estimator in calibration form is (approximately) $V_M = \sum_S \left(w_k^2 - w_k\right)\sigma_k^2$. This suggests the following estimator for the model variance of the linear-calibration estimator:

$$v_M = v_M\left(t_y^{\mathrm{CAL}}\right) = \sum_{k\in S}\left(w_k^2 - w_k\right) r_k^2, \tag{13}$$

where $r_k = y_k - \mathbf{x}_k\mathbf{b}$ is a sample residual, and $\mathbf{b}$ is any model-unbiased estimator for the model parameter, $\boldsymbol{\beta}$.

Under mild assumptions similar to the regularity conditions in Eqs. (5) and (6), $E_\varepsilon\left(r_k^2\right) = \sigma_k^2 + O_P(1/n)$, and $E_\varepsilon\left(v_M\right) = V_M\left[1 + O_P\left(1/n\right)\right]$; that is to say, v_M is nearly unbiased under the model (more formally, v_M is an asymptotically model unbiased estimator of the model variance of t_y^{CAL}) when mild regularity conditions hold and either $N = O(n^{3/2})$ or σ_k^2 has the form $\mathbf{x}_k\zeta$ for some ζ.

5.1. Poisson sampling

From Eq. (10), we can conclude that randomization mean squared error of the linear-calibration estimator under Poisson sampling is approximately $V = \sum_U \left(d_k - 1\right) e_k^2$. If $w_k = d_k\left[1 + O_P\left(1/n^{1/2}\right)\right]$, then v_M is also a reasonable randomization mean-squared-error estimator when $r_k^2 \approx e_k^2$.

Let $r_k = y_k - \mathbf{x}_k \mathbf{b}_\mathbf{h}$ and $e_k = y_k - \mathbf{x}_k \mathbf{B}_\mathbf{h}$, where $\mathbf{b}_\mathbf{h} = \left(\sum_S d_k \mathbf{h}'_k \mathbf{x}_k\right)^{-1} \sum_S d_k \mathbf{h}'_k y_k$, and $\mathbf{B}_\mathbf{h} = \left(\sum_U \mathbf{h}'_{k_k} \mathbf{x}_k\right)^{-1} \sum_U \mathbf{h}'_k y_k$. Because $\mathbf{b} = \mathbf{B}\left[1 + \mathbf{O}_P\left(1/n^{1/2}\right)\right]$ under mild conditions we assume to hold, $r_k^2 = e_k^2 + O_P\left(1/n^{1/2}\right)$. Thus, v_M is simultaneously a nearly unbiased estimator for the model variance of a linear calibration estimator t_y^{CAL} and a nearly unbiased estimator for its randomization mean squared error. Replacing the d_k within $\mathbf{b}_\mathbf{h}$ by w_k has no effect on the aforementioned arguments.

Observe that the relative model bias of v_M is $O_P(1/n)$, whereas its relative randomization bias is $O_P(1/n^{1/2})$. Nevertheless, Kott and Brewer (2001) discuss ways for even this small model bias to be removed. We will introduce two of them later in the section.

5.2. *Some other sampling designs*

The desirable model-based properties of v_M are unchanged when we move from Poisson sampling to an alternative element-sampling design. Unfortunately, the same cannot be said about its randomization-based properties.

Consider instead

$$v_1 = v_M + \sum_{\substack{k,j \in S \\ k \neq j}} \left(1 - \frac{\pi_k \pi_j}{\pi_{kj}}\right) \frac{r_k}{\pi_k} \frac{r_j}{\pi_j}$$

$$= \sum_{k \in S} \left(w_k^2 - w_k\right) r_k^2 + \sum_{\substack{k,j \in S \\ k \neq j}} \left(1 - \frac{\pi_k \pi_j}{\pi_{kj}}\right) \frac{r_k}{\pi_k} \frac{r_j}{\pi_j}, \tag{14}$$

which mimics

$$v_{\mathrm{RAN}} = \sum_{k \in S} (1 - \pi_k) \left(\frac{e_k}{\pi_k}\right)^2 + \sum_{\substack{k,j \in S \\ k \neq j}} \left(1 - \frac{\pi_k \pi_j}{\pi_{kj}}\right) \frac{e_k}{\pi_k} \frac{e_j}{\pi_j},$$

an unbiased estimator for the randomization variance of the idealized general difference estimator, $t_y^{\mathrm{DGIF}} = t_y^E + \left(T_\mathbf{x} - t_\mathbf{x}^E\right)\mathbf{B}_\mathbf{h} = t_y^{\mathrm{CAL}} + O_P\left(N/n\right)$. The right-hand side of Eq. (14) equals v_M under Poisson sampling. If the sampling design is such that $\sum_{k,j \in S} \left|1 - \left(\pi_k \pi_j / \pi_{kj}\right)\right|$ is $O(n)$, then the regulatory conditions in Eqs. (4) and (5) assure that the model expectation of this summation is ignorably small (at most $O(N^2/n^2)$, while v_1 itself is $O(N^2/n)$). The randomization expectation of the difference between v_1 and v_{RAN} is likewise asymptotically ignorable.

An unfortunate property of the model variance/randomization mean-squared-error estimator v_1 is that it can be negative for some samples under certain designs. The same problem, to a lesser extent, plagues the weighted residual variance estimator in Särndal et al. (1989),

$$v_{\mathrm{SSW}} = \sum_{k,j \in S} \left(1 - \frac{\pi_k \pi_j}{\pi_{kj}}\right) w_k r_k w_j r_j,$$

which does not collapse into v_M under Poisson sampling.

The weighted residual variance estimator is guaranteed to be nonnegative under stratified simple random sampling. By contrast,

$$v_1 = v_M + \sum_{\substack{k,j\in S\\ k\neq j}}\left(1-\frac{\pi_k\pi_j}{\pi_{kj}}\right)\frac{r_k}{\pi_k}\frac{r_j}{\pi_j}$$

$$= v_M - \sum_{\alpha=1}^{A}\frac{\left[\left(\frac{N_\alpha}{n_\alpha}\right)^2-\frac{N_\alpha}{n_\alpha}\right]\left[\left(\sum\limits_{k\in S_\alpha} r_k\right)^2-\sum\limits_{k\in S_\alpha} r_k^2\right]}{n_\alpha-1}, \tag{15}$$

can be negative.

Observe that v_1 will certainly be nonnegative when all $\sum_{S_\alpha} r_k = 0$. The equality obtains when all the stratum-indicator variables are components of $\mathbf{h}_k$ (because $\sum_S d_k\mathbf{h}_k' r_k = \sum_S d_k\mathbf{h}_k'\left[y_k - \mathbf{x}_k\left(\sum_S d_j\mathbf{h}_j'\mathbf{x}_j\right)^{-1}\sum_S d_j\mathbf{h}_j' y_j\right] = 0$). For the regularity conditions to be sensible under this scenario, effectively each of the n_α/n should converge to a positive constant as the sample size grows arbitrarily large, that means, the number of strata should be relatively small compared to n and the sample size within each relatively large.

More generally, by replacing the N_α/n_α in Eq. (15) by the appropriate asymptotically-identical w_k, we have the alternative variance/mean-squared-error estimator

$$\begin{aligned} v_2 &= \sum_{\alpha=1}^{A}\left[\sum_{k\in S_\alpha}\left(w_k^2-w_k\right)r_k^2\right] - \sum_{\alpha=1}^{A}\frac{\left(\sum\limits_{k\in S_\alpha}\left(w_k^2-w_k\right)^{1/2}r_k\right)^2-\sum\limits_{k\in S_\alpha}\left(w_k^2-w_k\right)r_k^2}{n_\alpha-1} \\ &= \sum_{\alpha=1}^{A}\left[\frac{n_\alpha}{n_\alpha-1}\sum_{k\in S_\alpha}\left(w_k^2-w_k\right)r_k^2 - \frac{1}{n_\alpha}\left(\sum_{k\in S_\alpha}\left(w_k^2-w_k\right)^{1/2}r_k\right)^2\right], \end{aligned} \tag{16}$$

which will be nonnegative as long as no w_k falls between 0 and 1. This is another reason for wanting the calibration weights to be bounded below by unity.

When finite population correction can be ignored (i.e., when all $N_\alpha \gg n_\alpha$ and almost all $w_k \gg 1$), Eq. (16) can be approximated by

$$v_3 = \sum_{\alpha=1}^{A}\frac{n_\alpha}{n_\alpha-1}\left[\sum_{k\in S_\alpha} w_k^2 r_k^2 - \frac{\left(\sum\limits_{k\in S_\alpha} w_k r_k\right)^2}{n_\alpha}\right].$$

This same variance equation we can use in practice for many stratified designs with unequal selection probabilities within each strata if all the selection probabilities are small or the sampling is with replacement.

Turning to a multistage sample, let $n_{1\alpha}$ be the size of the PSU sample from stratum α, now denoted $S_{1\alpha}$, and S_i be the element subsample from PSU i. The multistage analogue

of the last equation is

$$v_3 = \sum_{\alpha=1}^{A} \frac{n_{1\alpha}}{n_{1\alpha}-1} \left[\sum_{i\in S_{1\alpha}} \left(\sum_{k\in S_i} w_k r_k \right)^2 - \frac{\left(\sum_{i\in S_{1\alpha}} \sum_{k\in S_{\alpha i}} w_k r_k \right)^2}{n_{1\alpha}} \right]. \tag{17}$$

The estimator is nearly unbiased for the randomization mean squared error of the linear calibration estimator when $\left[\pi_{1ig}/\left(\pi_{1i}\pi_{1g}\right)\right] - \left[\left(n_{1\alpha}-1\right)/n_{1\alpha}\right]$ is ignorably small for any distinct pair of PSUs i and g in stratum α. It is easy to see that v_3 is also nearly unbiased for the model variance of the linear calibration estimator as an estimator for $\sum_U \mathbf{x}_k\boldsymbol{\beta}$.

We can generalize the error structure of the model. Instead of requiring $E\left(\varepsilon_k\varepsilon_j\right)$ to be zero when k and j are different elements, we now require only that this covariance be bounded when the two are from the same PSU. When k and j are from different PSUs, $E\left(\varepsilon_k\varepsilon_j\right)$ is again assumed to be zero.

The new error structure allows elements within the same PSU to be correlated in complex patterns, which need not be specified. Correlations can differ across PSUs and even within PSUs when there are additional levels of clustering (e.g., individuals within households, households within blocks, and blocks with PSUs). Observe that under this more general error structure both $E_\varepsilon\left(v_3\right)$ and $E_\varepsilon\left[\left(t_y^{\mathrm{CAL}} - \sum_U \mathbf{x}_k\boldsymbol{\beta}\right)^2\right]$ are (asymptotically in the case of the former) equal to $\sum_{\alpha=1}^{A}\sum_{i\in S_{1\alpha}} E_\varepsilon\left[\left(\sum_{k\in S_i} w_k\varepsilon_k\right)^2\right]$. Thus, v_3 retains its model-based properties under the more general error structure.

5.3. *Reducing the model bias even further*

We have restricted our attention to estimators with good randomization-based properties because the linear model in Eq. (3) can fail. When the survey is designed to collect information on a set of variables, as is usually the case in practice, the linear model in Eq. (3) may be very reasonable for some survey variables, but not so for others.

The good randomization-based properties of a calibration estimator are asymptotic, but the world we live in is finite. That is why it is helpful to employ models. They effectively "speed up" the asymptotics. In Eqs. (15)–(17), we have good estimators for both the model variance and randomization mean squared error of the linear calibration estimator. Under certain conditions, the model bias of these as model-variance estimators is of a smaller asymptotic order than its randomization counterpart. Variance/mean-squared-error estimators possessing this smaller model bias have been shown empirically to produce confidence intervals with closer to nominal coverage properties (see, for example, Wu and Deng, 1983).

We can potentially reduce the model bias of the three variance estimators even further under by replacing each with

$$v_{ca} = v_c \frac{E_\varepsilon\left[\left(t_y^{\mathrm{CAL}} - T_y\right)^2\right]}{E_\varepsilon\left(v_c\right)}, \tag{18}$$

where $c = 1$, 2, or 3. Even assuming the ε_k are uncorrelated, we would need to further assume the σ_k^2 are known to take the two model expectations in (18) (more generally, we could assume that $E[(\varepsilon_1, \ldots, \varepsilon_N)'\,(\varepsilon_1, \ldots, \varepsilon_N)] = \Omega$ is a known block diagonal matrix, but only the simpler diagonal-Ω variant will be discussed here). Both the numerator and denominator are linear in the σ_k^2. As a result, we only need to assume these values up to a constant to compute v_{ca}.

A method that does not depend on an assumption about the relative sizes of the σ_k^2 follows from the observation that

$$
\begin{aligned}
E\left(r_k^2\right) &= \sigma_k^2\left[1 - 2\mathbf{x}_k\left(\sum_S d_j\mathbf{h}_j'\mathbf{x}_j\right)^{-1} d_k\mathbf{h}_k'\right] \\
&\quad + \mathbf{x}_k\left(\sum_S d_j\mathbf{h}_j'\mathbf{x}_j\right)^{-1}\sum_S d_j^2\sigma_j^2\mathbf{h}_j'\mathbf{h}_j\left(\sum_S d_j\mathbf{x}_j'\mathbf{h}_j\right)^{-1}\mathbf{x}_k' \\
&= \sigma_k^2\left[1 - \mathbf{x}_k\left(\sum_S d_j h_j'\mathbf{x}_j\right)^{-1} d_k\mathbf{h}_k'\right] \\
&\quad + \mathbf{x}_k\left(\sum_S d_j\mathbf{h}_j'\mathbf{x}_j\right)^{-1}\sum_S d_j^2\sigma_j^2\mathbf{h}_j'\mathbf{h}_j\left(\sum_S d_j\mathbf{x}_j'\mathbf{h}_j\right)^{-1}\mathbf{x}_k' \\
&\quad - \mathbf{x}_k\left(\sum_S d_j\mathbf{h}_j'\mathbf{x}_j\right)^{-1} d_k\mathbf{h}_k'\sigma_k^2,
\end{aligned}
$$

where the last two terms on the right-hand side are $\mathbf{O}_P(1/n)$ and tend to have opposite signs. This suggests an ad hoc alternative to Eq. (18): replacing the r_k in the three variance estimators by

$$
\tilde{r}_k = r_k\frac{1}{\sqrt{1 - \mathbf{x}_k\left(\sum_S d_j\mathbf{h}_j'\mathbf{x}_j\right)^{-1} d_k\mathbf{h}_k'}}. \tag{19}
$$

It is not hard to see that using either method of model-bias reduction or their combination (first changing the r_k using Eq. (19) and then applying Eq. (18) assuming values for the σ_k^2 up to a constant) has no affect on the asymptotic randomization-based properties of the resulting estimators. How effective these methods are in practice awaits broad empirical evaluation, although using Eq. (19) worked well in a small study described in Kott (2006).

6. Nonlinear calibration

6.1. The not-necessarily-linear calibration estimator

We can generalize the linear form for the calibration weights in Eq. (11) to

$$
w_k^{\mathrm{GEN}} = d_k f(\mathbf{h}_k\mathbf{g}), \tag{20}
$$

where f is a monotonic, twice-differentiable function with bounded second derivatives everywhere, $f(0) = 1$, $f'(0) = 1$ ($f'(0)$ is the first derivative of $f(.)$ evaluated at 0), and $\mathbf{g}$ is chosen so that the calibration equation holds.

Strictly speaking, there should be an additional symbol on w_k^{GEN} (and later on w_k^{LIN}) to denote the particular choice of $\mathbf{h}_k$. It has been dropped for convenience. An extension of Eq. (8) allowing $f_k(.)$ to vary across the sampled elements, as we will do in subsequent subsections, is straightforward but adds nothing to the discussion immediately following.

A solution, $\mathbf{g}$, to Eq. (20) can be approached iteratively. One can start with $\mathbf{g}^{(0)} = \mathbf{0}$; that is, $\sum_S w_k^{(1)} y_k = \sum_S w_k^{\text{LIN}} y_k + O_P(N/n) = T_y\,[1 + O_P(1/n)]$, where $w_k^{(0)} = d_k f(0) = d_k$. For $r = 1, 2, \ldots$, one then sets $\mathbf{g}^{(r)} = \mathbf{g}^{(r-1)} + \left[\sum_S f'\left(\mathbf{h}_k \mathbf{g}^{(r-1)}\right) d_k \mathbf{h}_k' \mathbf{x}_k\right]^{-1} \left(T_{\mathbf{x}} - \sum_S w_k^{(r-1)} x_k\right)'$, and $w_k^{(r)} = d_k f\left(\mathbf{h}_k \mathbf{g}^{(r)}\right)$. Iteration stops at r^* when $T_{\mathbf{x}} = \sum_S w_k^{(r*)} x_k$ for all practical purposes.

Note that $\mathbf{g}^{(1)}$ equals the $\mathbf{g}$ in $w_k^{\text{LIN}} = d_k\,(1 + \mathbf{h}_k \mathbf{g})$. A Taylor expansion around zero reveals $f(\mathbf{h}_k \mathbf{g}^{(1)}) = 1 + \mathbf{h}_k \mathbf{g}^{(1)} + O_P(1/n)$ under our usual regularity conditions, so $\sum_S w_k^{(1)} y_k = \sum_S w_k^{\text{LIN}} y_k + O_P(N/n) = T_y\,[1 + O_P(1/n)]$. Furthermore, it is not difficult to see that $w_k^{\text{GEN}} = w_k^{\text{LIN}}\,[1 + O_P(1/n)]$, an equality that proves helpful in variance estimation. One should be aware, however, that there may not be a set of weights that can be expressed in the form of Eq. (20) while satisfying the calibration equation.

The most common example in practice of a nonlinear f is $f(\mathbf{h}_k \mathbf{g}) = \exp(\mathbf{h}_k \mathbf{g})$, where $\mathbf{h}_k = \mathbf{x}_k = \mathbf{u}_k$ is a vector made up entirely of group-membership indicator variables, some linear combination of which is unity. The groups themselves need not be mutually exclusive. The standard way of computing calibration weights with this form uses Deming and Stephan's iterative proportional fitting. One striking advantage of the alternative routine described here is that, unlike iterative proportional fitting, it can be used even when some of the components of $\mathbf{x}_k$ are continuous. Note that the resulting generalized-raking calibration weights, if they can be found, will always be nonnegative.

Another nonlinear f that always yields positive calibration weights, at least when $\mathbf{h}_k$ is equal to $\mathbf{x}_k$, is $f(\mathbf{x}_k \mathbf{g}) = (1 + \mathbf{x}_k \mathbf{g})^{-1}$. This method, which grows out of pseudo empirical likelihood theory, is discussed in depth in Chapter 30.

Returning to the general case, because $w_k^{\text{GEN}} = w_k^{\text{LIN}}\,[1 + O_P(1/n)]$ under conditions we assume to hold, it is not hard to show that the variance estimators in Section 5 apply equally well to the calibration estimator based on the w_k^{GEN} with $r_k = y_k - \mathbf{x}_k \mathbf{b}_{\mathbf{h}}$, and $\mathbf{b}_{\mathbf{h}} = \left(\sum_S d_k \mathbf{h}_k' \mathbf{x}_k\right)^{-1} \sum_S d_k \mathbf{h}_k' y_k$. This is asymptotically unchanged if $\mathbf{b}_{\mathbf{h}}$ is replaced by $\mathbf{b}_{\mathbf{h}f} = \left(\sum_S d_k f(\mathbf{h}_k \mathbf{g})\, \mathbf{h}_k' \mathbf{x}_k\right)^{-1} \sum_S d_k f(\mathbf{h}_k g)\, \mathbf{h}_k' y_k$ or $\mathbf{b}_{\mathbf{h}f'} = \left(\sum_S d_k f'(\mathbf{h}_k \mathbf{g})\, \mathbf{h}_k' \mathbf{x}_k\right)^{-1} \sum_S d_k f'(\mathbf{h}_k \mathbf{g})\, \mathbf{h}_k' y_k$ since $f(0) = f'(0) = 1$.

6.2. *Truncated linear calibration*

A common version of nonlinear calibration is truncated linear calibration, which puts restrictions on the range of the calibration weights. Allowing a potentially different f for each element, truncated linear calibration takes the form:

$$f_k(\delta) = \begin{cases} L_k & \text{for} \quad \delta < L_k - 1 \\ 1 + \delta & \text{for} \quad L_k - 1 \le \delta \le U_k - 1, \\ U_k & \text{for} \quad \delta > U_k - 1 \end{cases} \tag{21}$$

where $U_k = \infty$ when $f(\delta)$ has no lower bound and $L_k = -\infty$ when $f(\delta)$ has no lower bound.

Setting $L_k = L$ and $U_k = U$ for all k in Eq. (21) puts bounds on the relative weight adjustment ($L \leq w_k/d_k \leq U$, as in Jayasuriya and Valliant, 1996). Similarly, setting $L_k = L/d_k$ or $U_k = U/d_k$, puts bounds on the weights themselves (e.g., $w_k \geq L$, as in Brewer, 1999b). Finally, setting $U_k = U/(d_k\tau_k)$, where τ_k is a measure for the size of element k, assures that the weighted size of the element, $w_k\tau_k$, is no greater than U. This can be helpful in establishment surveys where an element with a high $w_k\tau_k$ may have an unreasonably large influence on the estimate for T_y.

Although $f_k(\delta)$ in Eq. (21) is no longer twice differentiable everywhere, our weight-computing algorithm still applies by treating $f_k'(\delta)$ at the breakdown points (U_k and L_k) as 0. The literature contains other equally-good methods for finding truncated linear calibration weights. Singh and Mohl (1996) describes many of them.

6.3. The example

Using the USDA data from Section 3.4, Table 3 contrasts truncated linear calibration with $c_k = 1$ and weights bounded from below by zero with generalized-raking calibration. Both versions of calibration assured that no calibration weight was negative. The first was very similar to linear calibration with $c_k = 1$, also displayed, which only produced one (of 1436) negative weight.

The minimum and maximum calibration weights were larger with generalized-raking calibration than the other two calibration methods as was the maximum calibration adjustment.

Nonetheless, the three methods produced very similar results. This is not surprising because for values of $\mathbf{x}_k\mathbf{g}$ close to zero, $f(\mathbf{x}_k\mathbf{g}) \approx 1 + \mathbf{x}_k\mathbf{g}$ for all three. Formally, if $\mathbf{x}_k\mathbf{g}$ is $O_P(1/n^{1/2})$, then $f(\mathbf{x}_k\mathbf{g}) = 1 + \mathbf{x}_k\mathbf{g} + O_P(1/n)$ under all three weighting schemes.

Table 3
Comparing the calibration weights under different forms of general calibration

	GREG with $c_k = 1$	GREG with $c_k = 1$ and Weights Truncated at 0	$w_k = d_k\exp(\mathbf{x}_k\mathbf{g})$
$\dfrac{\sum_S \lvert w_k - d_k\rvert}{\sum_S d_k}$	0.0408	0.0408	0.0421
$\dfrac{\text{Max}\,\{w_k\}}{\text{Max}\,\{d_k\}}$	1.0500	1.0502	1.0533
$\text{Max}\left\{\dfrac{w_k}{d_k}\right\}$	1.5290	1.5288	1.6546
$\text{Min}\,\{w_k\}$	−1.0846	0.0000	0.5596
$\text{Min}\left\{\dfrac{w_k}{d_k}\right\}$	−0.0587	0.0000	0.3028
Percent of $w_k < 1$ (number)	0.6964(10)	0.6964(10)	0.4875(7)
Percent of $w_k \leq 0$ (number)	0.0696(1)	0.0000	0.0000
$\dfrac{\sum_S w_k^2 - \sum_S w_k}{\sum_S d_k^2 - \sum_S w_k}$	1.0748	1.0748	1.0776
$\dfrac{\sum_S w_k^2 x_k - \sum_S w_k x_k}{\sum_S d_k^2 x_k - \sum_S w_k x_k}$	1.0357	1.0356	1.0340

Note: x_k, calculated cropland.

7. Calibration and quasi-randomization

7.1. Unit nonresponse

One popular way of handling unit (whole-element) nonresponse is to treat response as an additional phase of Poisson sampling. This type of response modeling is called "quasi-randomization," because it treats the model as if it were part of the sample selection mechanism.

Under quasi-randomization, each element k in the original sample, now denoted S^0, is assumed to have a probability of response, p_k. The probability of jointly "choosing" elements k and j is $p_k p_j$, and the magnitude of p_k is independent of whether k is chosen for the original sample. It is often possible to construct a set of weights so that the calibration estimator is randomization consistent under the quasi-randomization model.

We are interested here in a particular way of constructing those weights. To this end, we assume that the quasi-random response model is correct. Each element has attached to it a row vector of benchmark variables, $\mathbf{x}_k$, for which $T_{\mathbf{x}} = \sum_U \mathbf{x}_j$ is known. In addition, each response propensity p_k is assumed to have the form:

$$p_k = \rho(\mathbf{h}_k\boldsymbol{\gamma}) = 1/f(\mathbf{h}_k\boldsymbol{\gamma}), \tag{22}$$

where $\boldsymbol{\gamma}$ is unknown, $\mathbf{h}_k$ is a row vector with the same dimension as $\mathbf{x}_k$, and the matrix $\sum_S d_k f(\mathbf{h}_k\boldsymbol{\gamma})\mathbf{h}_k'\mathbf{x}_k/N$, where S now denotes the "subsample" of respondents, is invertible both for the realized N and in the probability limit. The function f is assumed to be monotonic and twice differentiable with bounded second derivatives everywhere. Its functional form is known, but the value of the governing parameter, $\boldsymbol{\gamma}$, is not. Unlike in the previous section, $f(0)$ and $f'(0)$ need not be unity.

Using the iterative method described in Section 6.1 to find $\mathbf{g}$, we will often be able to uncover a row vector, $\mathbf{g}$, such that $T_{\mathbf{x}} = \sum_S d_k f(\mathbf{h}_k\mathbf{g})\mathbf{x}_k$. As a result, estimating T_y with $t_y^{\mathrm{CAL}} = \sum_S w_k y_k$, where the adjusted calibration weights have the form, $w_k = d_k f(\mathbf{h}_k\mathbf{g})$, may have good properties under the linear prediction model, $y_k = \mathbf{x}_k\boldsymbol{\beta} + \epsilon_k$, when each h_{pk} is a function of the $\mathbf{x}_g (g \in U)$, and $E\{\varepsilon_k|\{\mathbf{x}_g, I_g; g \in U\}\} = 0$ for all $k \in U$. Note that I_g is now an indicator that g is in both the original sample and the respondent subsample. In this context, prediction-model unbiasedness is simply a result of the weights satisfying the calibration equation (the prefix "prediction" is needed to distinguish this model from the quasi-random one).

Whether or not t_y^{CAL} can reasonably be called prediction-model unbiased has no effect on its quasi-randomization-based properties. Because $T_{\mathbf{x}} = \sum_S d_k f(\mathbf{h}_k\mathbf{g})\,\mathbf{x}_k$, our assumptions and the mean value theorem reveal

$$T_{\mathbf{x}} = \sum_{k\in S} d_k f(\mathbf{h}_k\boldsymbol{\gamma})\,\mathbf{x}_k + \sum_{k\in S} d_k f'(\theta_k)\,[\mathbf{h}_k(\mathbf{g}-\boldsymbol{\gamma})]\mathbf{x}_k = \mathbf{O}_P\left(\frac{N}{n^{1/2}}\right)$$

for some θ_k between $\mathbf{h}_k\mathbf{g}$ and $\mathbf{h}_k\boldsymbol{\gamma}$ (recall $f(.)$ is monotonic). From this we see that if $\sum_S d_k f'(\mathbf{h}_k\boldsymbol{\gamma})\mathbf{h}_k'\mathbf{x}_k/N$ is invertible both for the realized N and at the probability limit, then

$$(\mathbf{g}-\boldsymbol{\gamma})' = \left[\sum_{k\in S} d_k f'(\mathbf{h}_k\boldsymbol{\gamma})\,\mathbf{h}_k'\mathbf{x}_k\right]^{-1}\left[T_{\mathbf{x}} - \sum_{k\in S} d_k f(\mathbf{h}_k\boldsymbol{\gamma})\,\mathbf{x}_k\right] + \mathbf{O}_P\left(\frac{1}{n}\right).$$

The estimator t_y^{CAL} has an error of

$$\begin{aligned} t_y^{\mathrm{CAL}} - T_y &= \sum_{k\in S} d_k f(\mathbf{h}_k\mathbf{g}) y_k - \sum_{k\in U} y_k \\ &= \sum_{k\in S} d_k f(\mathbf{h}_k\mathbf{g}) e_k - \sum_{k\in U} e_k, \end{aligned}$$

where $e_k = y_k - \mathbf{x}_k\left(\sum_U f'\left(\mathbf{h}_j\boldsymbol{\gamma}\right)p_j\mathbf{h}_j'\mathbf{x}_j\right)^{-1}\sum_U f'\left(\mathbf{h}_j\boldsymbol{\gamma}\right)\,p_j\mathbf{h}_j'y_j$, and $p_j = 1/f(\mathbf{h}_j\boldsymbol{\gamma})$ so $\sum_S f'(\mathbf{h}_k\boldsymbol{\gamma})\,\mathbf{h}_k'e_k = \mathbf{O}_P\left(N/n^{1/2}\right)$. Continuing:

$$\begin{aligned} t_y^{\mathrm{CAL}} - T_y &= \sum_{k\in S} d_k f(\mathbf{h}_k\boldsymbol{\gamma})e_k - \sum_{k\in U} e_k + \sum_{k\in S} d_k[f(\mathbf{h}_k\mathbf{g}) - f(\mathbf{h}_k\boldsymbol{\gamma})]e_k \\ &= \sum_{k\in S} d_k f(\mathbf{h}_k\boldsymbol{\gamma})e_k - \sum_{k\in U} e_k + \sum_{k\in S} d_k f'(\mathbf{h}_k\boldsymbol{\gamma})\,\mathbf{h}_k\,(\mathbf{g}-\boldsymbol{\gamma})e_k + O_P\left(\frac{N}{n}\right) \\ &= \sum_{k\in S} d_k f(\mathbf{h}_k\boldsymbol{\gamma})e_k - \sum_{k\in U} e_k + (\mathbf{g}-\boldsymbol{\gamma})'\sum_{k\in S} d_k f'(\mathbf{h}_k\boldsymbol{\gamma})\,\mathbf{h}_k'e_k + O_P\left(\frac{N}{n}\right) \\ &= \sum_{k\in S} d_k f(\mathbf{h}_k\boldsymbol{\gamma})e_k - \sum_{k\in U} e_k + O_P\left(\frac{N}{n}\right). \end{aligned} \tag{23}$$

Thus, t_y^{CAL} is quasi-randomization consistent under mild conditions whenever $t = \sum_S d_k f(\mathbf{h}_k\boldsymbol{\gamma})\,y_k$ is.

To estimate the quasi-randomization mean squared error of t_y^{CAL} (i.e., the estimator's randomization mean squared error under the quasi-random response model), we first note that the probability that distinct elements k and j are both in the respondent subsample is $\ddot{\pi}_{kj} = \pi_{kj}p_kp_j$. Let $\ddot{\pi}_k = \pi_k p_k$, and recall that $d_k = 1/\pi_k$ and $1/p_k = d_k f(\mathbf{h}_k\boldsymbol{\gamma})$. From Eq. (23), we see that the randomization mean squared error of t_y^{CAL} is approximately

$$\begin{aligned} E_I\left[\left(t_y^{\mathrm{CAL}} - T_y\right)^2\right] &\approx \sum_{k\in U}\sum_{j\in U}\left(\ddot{\pi}_{kj} - \ddot{\pi}_k\ddot{\pi}_j\right)\frac{e_k}{\ddot{\pi}_k}\frac{e_j}{\ddot{\pi}_j} \\ &= \sum_{k\in U}(1-\ddot{\pi}_k)\frac{e_k^2}{\ddot{\pi}_k} + \sum_{\substack{k,j\in U\\ j\neq k}}\left(\pi_{kj} - \pi_k\pi_j\right)\frac{e_k}{\pi_k}\frac{e_j}{\pi_j}, \end{aligned}$$

which can be estimated by v_1 in Eq. (15), where now

$$\begin{aligned} r_k &= y_k - \mathbf{x}_k b_{\mathbf{h}f'} \\ &= y_k - \mathbf{x}_k\left(\sum_{j\in S} d_j f'\left(\mathbf{h}_j\mathbf{g}\right)\mathbf{h}_j'\mathbf{x}_j\right)^{-1}\sum_{j\in S} d_j f'\left(\mathbf{h}_j\mathbf{g}\right)\mathbf{h}_j'y_j. \end{aligned} \tag{24}$$

This serves as both a reasonable estimator for the prediction-model variance and quasi-randomization mean squared error under mild conditions, since $w_k \approx 1/\ddot{\pi}_k$ and $r_k \approx e_k$.

When the actual sample is multistage, and the first stage selection probabilities are ignorably small, v_3 in Eq. (17) can be used as the variance/mean-squared-error estimator with r_k defined once more by Eq. (24).

Observe that when there is no nonresponse, $\boldsymbol{\gamma} = 0$, so that $f'(\mathbf{h}_j\mathbf{g}) = f'(0) + f''(0)\,\mathbf{h}_j\mathbf{g} + O_P(1/n) = f'(0) + O_P(1/n^{1/2})$. As a result, the f'-terms in Eq. (24) are all asymptotically identical and can be removed from the definition of r_k without altering the asymptotics of the variance/mean-squared-error estimators.

When f is linear, $f'(\delta) = f'(0) = 1$ for all δ, and the r_k in Eq. (24) are computed as if there were no nonresponse. The same holds true for the variance/mean-squared-error estimators v_1, v_2, and v_3. For generalized-raking calibration, $f'(.) = f(.) = \exp(.)$, and the r_k are computed accordingly.

7.2. *Coverage adjustment*

We can also use calibration weighting to adjust for undercoverage. In this context, the sampling frame itself is assumed to be a quasi-random sample from a hypothetical complete population. The actual sample is treated as the second phase of a two-phase design. The frame becomes S^0, whereas the hypothetical complete population is U. Attached to U is the known vector $T_{\mathbf{x}}$. The probability element $k \in U$ is in S^0 is assumed to be modeled correctly by Eq. (22). If the first (from U to S^0) and second (from S^0 to S) phases of sampling are independent, then all the theory developed for using calibration weighting to handle nonresponse carries over to handling undercoverage.

Overcoverage (duplication) or a combination of under and overcoverage can be handled in the same way. The definition of p_k in Eq. (22) becomes the expected number of times k is in the frame, which can now exceed 1 due to potential duplication.

We have seen that the calibration weights described in this section can produce estimators with good prediction-model-based properties [under Eq. (3)] when the prediction model is correct (in particular, $E\left(\varepsilon_k|\{\mathbf{x}_g, I_g; g \in U\}\right) = 0$ and each h_{pk} is a function of the $\mathbf{x}_g$), and good quasi-randomization properties when the response or coverage model [in Eq. (21)] is correct. In some sense, one model provides protection against the failure of the other. See Kott (1994).

7.3. *The general exponential model*

In the general exponential model or GEM, the $f_k(\delta)$ are again defined for each individual element and have the flexible form:

$$f_k(\delta) = \frac{U_k\,(C_k - L_k)\exp(\delta) + L_k\,(U_k - C_k)}{(U_k - C_k) + (C_k - L_k)\exp(\delta)}, \tag{25}$$

where $L_k \geq 0$, $1 < U_k \leq \infty$, and $L_k < C_k \leq U_k$ are predetermined constants. Observe that if $C_k = 1$, $U_k = \infty$, and $L_k = 0$, then $f_k(\delta) = \exp(\delta)$, while $p_k = 1/f_k(\delta) = \exp(-\delta)$. Similarly, if $C_k = 2$, $U_k = \infty$, and $L_k = 1$, then $p_k = [1 + \exp(\delta)]^{-1}$; that is to say, the probability of element response (or coverage) is logistic. The values L_k and U_k serve as bounds on the calibration adjustment, $f_k(\delta)$, while $C_k = f_k(0)$ is effectively its center.

Although it is tempting to set all the $f_k(\delta)$ to $[1 + \exp(\delta)]^{-1}$ so that the response model can be logistic, it is not practical to do so in the calibration context. The lack of an upper bound may allow some elements to have an unreasonable impact on the estimated total, which can result in an unacceptably large mean squared error. Perhaps even more problematic, when some element or elements has a probability of response very close to 1, the lower limit on $f(\delta)$ can make finding a calibration weight, w_k, impossible.

The (U.S.) National Survey on Drug Use and Health (formerly the National Household Survey on Drug Abuse) is based on a sample of dwelling units and a subsample of individuals. GEM modeling is employed independently within nine mutually exclusive domains. Weights are calibrated at various phases of the estimation process. When adjusting the weights of a respondent subsample to full-sample benchmark totals, the L_k in Eq. (25) are set at 1, the C_k at the inverse of the overall domain response rate, and the U_k by trial and error depending on the original size of the sample weights. Further details can be found in Chen et al. (2000).

7.4. Prediction and response model variables

Consider the popular group-ratio-model estimator discussed in Section 4. Recall $\mathbf{x}_k = (u_{1k}x_k, \ldots, u_{Pk}x_k)$, where x_k is a nonnegative scalar, and the P groups for which the u_{pk} are indicator functions are exhaustive and mutually exclusive. Setting $\mathbf{h}_k = \mathbf{u}_k = (u_{1k}, \ldots, u_{Pk})$ and $f(\delta) = 1 + \delta$ yields the group-ratio-model estimator, which is expressible in calibration form with the weight $w_k = d_k\left(\sum_{U\cap p} d_j x_j \big/ \sum_{S\cap p} d_j x_j\right)$ for each k in the sample and group p. Because $f(\mathbf{u}_k\mathbf{g}) = f(g_p)$ when element k is in group p, the choice of f does not matter with this $\mathbf{h}_k$ except, perhaps, when the range $f(\delta)$ is limited; for example, $f(\delta) = [1 + \exp(\delta)]^{-1}$ cannot be less than 1.

In the presence of nonresponse (or coverage error), this linear calibration is unbiased under the prediction model where for every element in the group p, $y_k = \beta_p x_k + \varepsilon_k$, and $E\big(\varepsilon_k|\{\mathbf{x}_g, I_g; g \in U\}\big) = 0$. It is also quasi-randomization consistent under the response model in which the response propensity, p_k, is constant within each group. Observe that the predictor model variable is $\mathbf{x}_k$, whereas the response model variable is $\mathbf{u}_k$, where each $u_{pk} = (x_{pk})^0$. If *either* model is correct, the group-ratio-model estimator is in some sense (nearly) unbiased.

It is possible to extend this to a generalized-raking calibration estimator where the u_{pk} are indicators of not-mutually-exclusive groups, and some linear combination of the groups is unity. Using business survey data cross-classified by industry and expected size, Hidiroglou and Patak (2006) compute calibration weights satisfying $\sum_S w_k\mathbf{x}_k = \sum_U \mathbf{x}_k$ of the form $w_k = d_k \exp(\mathbf{u}_k\mathbf{g})$ (now the choice of f matters) using a variant of iterative proportional fitting. The routine in Section 6.1 would have worked as well. Again, in the presence of nonresponse, the $\mathbf{x}_k$ serves as the prediction-model variable while $\mathbf{u}_k$ is the response-model variable. That is to say, the expected value of $y_k/\mathbf{x}_k$ under the prediction model is the sum of the parameters associated with the two groups containing k (a size group and an industry group) whereas the probability of the element responding under the response model is the product of the probabilities (the $\exp(\gamma_p)$) associated with those groups (since $\exp(\mathbf{u}_k\boldsymbol{\gamma}) = \prod_{p=1}^{P} \exp\big(u_{pk}\gamma_p\big)\big)$.

Table 4
Comparing different versions of calibration weighting for nonresponse

	GREG with $c_k = 1$ and Weights Truncated at 0	$w_k = d_k \exp(\mathbf{x}_k \mathbf{g})$	$w_k = d_k \exp(\mathbf{h}_k \mathbf{g})$ and $h_{pk} = (x_{pk})^{1/2}$
Weight characteristics			
$\frac{\text{Max}\{w_k\}}{\text{Max}\{d_k\}}$	1.4178	1.4190	1.4722
$\text{Max}\left\{\frac{w_k}{d_k}\right\}$	2.5363	2.8703	2.5130
$\text{Min}\{w_k\}$	0.0000	0.7972	1.0000
$\text{Min}\left\{\frac{w_k}{d_k}\right\}$	0.0000	0.3350	0.8685
Some estimated total planted acres (and their estimated CVs)			
Corn	1,199,257 (2.38)	1,198,278 (2.39)	1,193,302 (2.45)
Alfalfa	126,850 (7.43)	127,046 (7.43)	127,141 (7.49)
Winter wheat	140,455 (6.04)	140,748 (6.10)	141,213 (6.32)

7.5. The example

Table 4 displays three different ways of using calibration to adjust for the unit nonresponse in the June 2005 Quarterly Agricultural Survey in Pennsylvania. Of the 1436 sampled potential agricultural places selected for the sample, 1015 fully responded to the crops section of the survey.

Two of the calibration methods featured in the new table extends what was done for the whole sample in Table 3 to the respondent subsample. Not surprisingly, the average calibration adjustment increased under both methods, but the maximum still fell below 3. Again, generalized-raking calibration produced larger minimum and maximum calibration weights and a larger maximum weight adjustment than truncated-from-below-at-zero linear calibration with $c_k = 1$. The actual estimated totals for three typical survey crops were very similar, as were their estimated coefficients of variation computed using v_M in Eq. (13) with r_k from Eq. (24) (applying the adjustment in Eq. (18) usually had no visible effect).

Table 4 also displays a variant of generalized raking where the components of the response-model variable, $\mathbf{h}_k$, were equal to the square-roots of the components of the benchmark (and prediction-model) variable, $\mathbf{x}_k$. Because all the elements in the frame had reported cropland, setting all $h_{pk} = \left(x_{pk}\right)^0$, as in the last subsection, would have lead to a singularity in the $\sum_S d_k f'(\mathbf{h}_k \mathbf{g})\, \mathbf{h}_k' \mathbf{x}_k / N$ matrix.

This variant produced calibration adjustments in a narrower range than the other two and estimators with slightly higher estimated CVs. The estimated totals themselves were very similar, however.

8. Other approaches, other issues

This chapter showed how calibration weighting combines linear prediction modeling and random sampling. Although several important contributions from the literature

were noted, better reviews contrasting and combining the model and randomization-based approaches to survey sampling can be found elsewhere. See, for example, Brewer (1994) and Fuller (2002). Chapter 23 discusses the model-based approach in depth, with Chapter 29 doing so from a Bayesian viewpoint.

Some maintain that it is misleading to connect, as was done here, calibration weighting with the linear prediction model. Calibration, in this view (apparently now held by Särndal himself; see the interview in Kott et al., 2005), is a purely randomization-based technique. Auxiliary information on benchmark variables can be used to reduce the randomization mean squared error of many survey variables while, at worst, marginally increasing the randomization mean squared errors of a few others.

There is another approach worth considering. Suppose the model is correct in the population, but the sampling design is not ignorable; in particular, the linear model in Eq. (3) holds under the assumption $E\left(\varepsilon_k|\{\mathbf{x}_g;\, g \in U\}\right) = 0$ but not $E\left(\varepsilon_k|\{\mathbf{x}_g,\, I_g;\, g \in U\}\right) = 0$. As a result, taking expectations first under the sampling mechanism and then the prediction model makes sense. This framework, more suitable for estimating model parameters than finite population totals, is explored in Chapter 39. Estimating model parameters is the subject of Chapter 24.

In the discussion of variance estimation in Section 5, we moved quickly from Poisson sampling to stratified simple random sampling and then to situations where finite population correction can be ignored entirely. Chapter 2 deals with a variety of unequal probability sampling schemes. Some of these have been developed to produce samples under which the calibration equation (nearly) holds.

From the randomization point of view, the calibration estimator is nonlinear; that is to say, t_y^{CAL} is a nonlinear function of the I_k. Effectively, the randomization mean-squared-error estimators discussed here are based on a linearization of this nonlinear estimator. Chapter 28 discusses how to use resampling techniques instead. Kott (2006) describes a jackknife analogue of v_3 in Eq. (17).

Although Section 6 dealt with nonlinear calibration, the prediction model itself was always linear. Chapter 27 discusses nonlinear prediction modeling.

In Section 6, we saw how calibration can be used to adjust for unit nonresponse by treating the respondent subsample as if it resulted from a two-phase sampling process. A more thorough treatment of when and how to adjust for unit nonresponse can be found in Chapter 9. Adjustments for item nonresponse are discussed in Chapter 10. Chapter 3 gives a full treatment of two-phase sampling and estimation.

An excellent text by Särndal and Lundström (2005) discusses many of the practical issues in using linear calibration to adjust for nonresponse and coverage errors. The awkward form of the response propensities in linear calibration (i.e., $p_k = [1 + \mathbf{h}_k\boldsymbol{\gamma}]^{-1}$) is excused as a useful approximation. This leads to quasi-randomization mean-squared-error estimation that is, at best, ad hoc.

It has been noted in this chapter, but not addressed, that the iterative procedure in Section 6.1 may fail to converge to a solution. Sometimes when convergence fails after a large number of iterations, replacing some $\mathbf{g}^{(r)} = \mathbf{g}^{(r-1)} + \left[\sum_S f'\left(\mathbf{h}_k\mathbf{g}^{(r-1)}\right) d_k\mathbf{h}_k'\mathbf{x}_k\right]^{-1}\left(T_{\mathbf{x}} - \sum_S w_k^{(r-1)}x_k\right)'$ with the half step: $\mathbf{g}^{(r)} = \mathbf{g}^{(r-1)} + 1/2\left[\sum_S f'\left(\mathbf{h}_k\mathbf{g}^{(r-1)}\right) d_k\mathbf{h}_k'\mathbf{x}_k\right]^{-1}\left(T_{\mathbf{x}} - \sum_S w_k^{(r-1)}x_k\right)'$ will prove effective. For a thorough discussion of computational methods, see Gentle (1998).

A simple alternative when convergence fails is either to alter the calibration-weight range restrictions or to drop some benchmark variables (and an equal number of components of $\mathbf{h}_k$). More generally, Rao and Singh (1997) propose an algorithm for balancing the requirements of the calibration equations and the range restrictions.

Chang and Kott (2007) discusses the calibration for nonresponse when there are more benchmark variables than response-model variables. They also treat the possibility that one of the survey variables is a response-model variable. The theory underlying the quasi-random response model is unaffected, but prediction-modeling as described in this chapter fails because $E\left(\varepsilon_k|\{\mathbf{h}_g; g \in U\}\right) \neq 0$.

Acknowledgements

I would like to thank Matt Fetter of the National Agricultural Statistics Service of USDA who did the programming for the tables in this chapter. Also, Ken Brewer, Michail Sverchkov, Ari Veijanen, and an anonymous reader for their thoughtful reviews of an earlier version of this chapter.

Sample Surveys: Inference and Analysis, Vol. 29B
ISSN: 0169-7161
© 2009 Elsevier B.V. All rights reserved
DOI: 10.1016/S0169-7161(09)00226-0

Estimating Functions and Survey Sampling

V. P. Godambe and Mary E. Thompson

1. Introduction

In statistical inference, the estimation of parameters plays an important role. Often, a parameter estimator is chosen to maximize or minimize some "objective function" of the data and the parameter: maximum likelihood estimation and least squares estimation are two examples. The point estimate is then the solution of an *estimating equation*, of which the left-hand side (the *estimating function*) is the derivative of the objective function with respect to the parameter, and the right-hand side is 0. The concept of an estimating function unifies the discussion of estimation in parametric and semiparametric contexts.

Both the maximum likelihood estimating equation and the least squares (normal) estimating equations are *unbiased*, in the sense that the corresponding estimating functions have expectation 0 under the motivating models. This unbiasedness is the fundamental property to be retained in all extensions and generalizations because it generally leads to consistency of the estimators.

The first results in estimating function theory (Durbin, 1960; Godambe, 1960) were optimality theorems. Godambe (1960) showed that among all unbiased estimating functions for a scalar parameter θ indexing a parametric family, the best in a certain sense was the score function, the derivative of the log likelihood function. This result and its successors guide the choice of estimating function in a wide variety of contexts. Bera et al. (2006) have provided a very thoughtful survey paper.

An estimating function can be viewed both as a vehicle for estimation and as a way of defining a parameter, the object of estimation.

For example, if ξ is a distribution for a random variable Y and ξ has parameter θ, let $\mathcal{E}$ denote expectation with respect to ξ, and suppose

$$\mathcal{E}\phi(Y, \theta) = 0. \tag{1}$$

Then, if $Y_1, \ldots, Y_N$ each have distribution ξ,

$$\mathcal{E}\left(\sum_{j=1}^{N} \phi(Y_j, \theta)\right) = 0;$$

the equation $\sum_{j=1}^{N} \phi(Y_j, \theta) = 0$ is an *unbiased estimating equation* for θ, and the function

$$\sum_{j=1}^{N} \phi(y_j, \theta)$$

is an *unbiased estimating function* for θ. We solve the estimating equation to obtain a point estimate of θ.

In other contexts, ϕ can *define* the parameter $\theta = \theta(\xi)$ as the value of θ which solves the Eq. (1) for any given ξ. For example, if $\phi(Y, \theta) = Y - \theta$, Eq. (1) defines the *mean* of the distribution ξ.

This dual role of estimating functions is seen in another aspect in survey sampling theory. To continue the same example, in survey sampling theory, a finite population mean is the solution of the *census estimating equation*

$$\sum_{j=1}^{N} (Y_j - \theta) = 0 \tag{2}$$

or its realization

$$\sum_{j=1}^{N} (y_j - \theta) = 0.$$

The Eq. (2) is unbiased also for estimation of the mean $\theta(\xi)$ of a hypothetical superpopulation distribution ξ for population values $Y_1, \ldots, Y_N$.

As will be seen in Section 2, many finite population quantities of interest are meaningfully thought of as roots of estimating functions for superpopulation parameters. A notable exception is the finite population total

$$T_y = \sum_{j=1}^{N} y_j.$$

But whenever we try to improve the estimation of T_y by thinking of it as $N\mu_y$, where μ_y is the finite population mean, or as $T_{\mathbf{x}}B$, where $T_{\mathbf{x}}$ is a vector of totals and B is a vector of finite population regression coefficients, it is useful to return to estimating functions or systems for the auxiliary parameters μ_y or B.

In the following, sections we will elaborate on these definitions, discuss optimality of estimating functions in survey sampling, set forth some of their asymptotic properties, and outline their role in interval estimation. We will conclude with some remarks on bootstrapping, on multivariate and nuisance parameters, and on imputation of estimating functions. A more detailed treatment of some of the parts is provided in Thompson (1997).

2. Defining finite population and superpopulation parameters through estimating functions

We begin by considering finite population parameters that are defined implicitly by a population equation of the form

$$\sum_{j=1}^{N} \phi_j(y_j, x_j, \theta_N) = 0. \tag{3}$$

Here y_j and x_j are values of observable variates, the ϕ_j are known p-dimensional functions, and θ_N is a p-dimensional quantity defined by (3).

Example 2.1. Several important finite population quantities are naturally defined as in (3):

(i) the population mean μ_y of y is defined by

$$\sum_{j=1}^{N}(y_j - \theta_N) = 0; \tag{4}$$

(ii) if y and x are real, the population ratio R of y to x is defined by

$$\sum_{j=1}^{N}(y_j - \theta_N x_j) = 0; \tag{5}$$

(iii) if y is real, its population cumulative distribution function (c.d.f.) evaluated at a real number y is defined by

$$\sum_{j=1}^{N}(I(y_j \le y) - \theta_N) = 0, \tag{6}$$

where

$$\begin{aligned} I(y_j) &= 1 \quad \text{if} \quad y_j \le y, \\ &= 0 \quad \text{if} \quad y_j > y; \end{aligned}$$

(iv) if y is real, its population median can be defined as the least value of θ_N such that

$$\sum_{j=1}^{N}(I(y_j \le \theta_N) - 1/2) \ge 0, \tag{7}$$

and this is approximately of the form (3).

In general, a population γth quantile is obtained from the definition (7) with 1/2 replaced by γ.

3. Design-unbiased estimating functions

A design-based methodology for estimating θ_N was first set out in general form by Binder (1983). It starts from the observation that the estimating function (3) can be estimated in an unbiased manner from the sample, using a Horvitz–Thompson form (Horvitz and Thompson, 1952). Thus, when a population function or parameter is defined by (3), an estimator for it can be defined as a solution of the sample estimating equation

$$\sum_{j \in s} \phi_j(y_j, x_j, \theta)/\pi_j = 0, \tag{8}$$

where π_j denotes the probability of inclusion of unit j. For any sampling design, even a so-called informative design where π_j depends on the responses y even after conditioning on x, the left-hand side of (8) is design-unbiased for the left-hand side of (3). Thus, we may expect a solution $\hat{\theta}_s$ of (8) to be close to θ_N for large samples in typical applications.

Example 3.2. In example (i) of Section 2, Eq. (8) is $\sum_{j\in s}(y_j - \theta)/\pi_j = 0$ and the resulting estimator of μ_y has the form

$$\hat{\theta}_s = \left(\sum_{j\in s} y_j/\pi_j\right) \Big/ \left(\sum_{j\in s} 1/\pi_j\right). \tag{9}$$

For any design with equal inclusion probabilities, even one which is not of fixed size, this estimator is the sample mean $\bar{y}_s$. In general, the estimator (9) need not be design-unbiased; however, for any choice of inclusion probabilities π_j, the estimator is error-free when all components of $\mathbf{y} = (y_1, \ldots, y_N)$ are the same—a property not shared by the unbiased estimator $\left(\sum_{j\in s} y_j/\pi_j\right)/N$. It can therefore be expected to be relatively efficient when the responses y are homogeneous, as they would be if generated from an independent and identically distributed (i.i.d) model or mixture of i.i.d. models. The estimator (9) is sometimes called the Hájek estimator since it was proposed in Hájek (1971).

In examples (ii)–(iv), we obtain similarly the estimator

$$\hat{R}_s = \left(\sum_{j\in s} y_j/\pi_j\right) \Big/ \left(\sum_{j\in s} x_j/\pi_j\right) \tag{10}$$

for the population ratio, the estimator

$$\hat{F}_s(y) = \left[\sum_{j\in s} I(y_j \le y)/\pi_j\right] \Big/ \left(\sum_{j\in s} 1/\pi_j\right) \tag{11}$$

for the value of the population c.d.f. at y, and the estimator

$$\hat{F}_s^{-1}\left(\frac{1}{2}\right) \tag{12}$$

for the population median, if we interpret (12) as the least value of y for which $\hat{F}_s(y) \ge 1/2$.

Note that $\hat{F}_s(y)$ of (11) is a true distribution function, in that it increases from 0 to 1 as y increases; this would not always be the case if $\sum_{j\in s} 1/\pi_j$ in the denominator were replaced by N.

An example of an informative design is *length-biased sampling*, where π_j is proportional to y_j. This design is seldom implemented precisely but can be a useful approximation. For example, the approximation might apply when y_j is the waiting time for person j for a certain medical procedure, and the sampling design is to take all population members waiting for the procedure at a certain date. Alternatively, the y_j might be the area of underground deposits detected by testing in random locations. In the length-biased sampling case, we obtain as estimator of the population c.d.f.

$$\hat{F}_s(y) = \left[\sum_{j\in s} I(y_j \le y)/y_j\right] \Big/ \left[\sum_{j\in s} 1/y_j\right]. \tag{13}$$

In estimation from large-scale surveys, the inverse inclusion probability weights $1/\pi_j$ are typically replaced by survey weights w_j that are sample dependent, incorporating adjustments for nonresponse and auxiliary information. The sample estimating equations in (8) are then replaced by the approximately unbiased

$$\sum_{j \in s} w_j \phi_j(y_j, x_j, \theta) = 0,$$

and these are also intended to yield approximately unbiased point estimates for θ_N.

4. Optimality

The sample estimating functions of (8) in Section 3 are natural but they are by no means the only design-unbiased estimators of the census estimating functions. Thus, it is of interest to see whether there is a sense in which they are optimal. The following describes a possible framework.

Suppose we have a superpopulation model describable in terms of a class $C = \{\xi\}$ of distributions ξ for the population array $\mathbf{Y} = (Y_1, \ldots, Y_N)$. Let $\theta = \theta(\xi)$ be a superpopulation parameter, namely a real- or vector-valued function defined on C. If $Y_1, \ldots, Y_N$ are independent under distributions in C, then in many practically important cases, an optimal (in terms of the model) estimating function (system) for θ exists in the form

$$\Phi^*(\mathbf{y}, \theta) = \sum_{j=1}^{N} \phi_j(y_j, x_j, \theta), \tag{14}$$

where each ϕ_j has the dimension of θ, and

$$\mathcal{E}\{\phi_j(Y_j, x_j, \theta(\xi))\} = 0 \quad \text{for all } \xi \in C. \tag{15}$$

We then seek an optimal sample estimating function (system) to correspond. Godambe and Thompson (1986a) have discussed the relevant optimality criteria in detail. For simplicity, let us take θ to be real, and suppress possible dependence on x, in the following outline.

When Φ^* of (14) is optimal for estimating θ, we regard θ_N, defined by

$$\sum_{j=1}^{N} \phi_j(y_j, \theta_N) = 0, \tag{16}$$

as the finite population parameter associated with θ. We then consider estimating θ_N from the sample by solving equation

$$g(\chi_s, \theta) = 0, \tag{17}$$

where

$$\chi_s = \{(j, y_j) : j \in s\} \tag{18}$$

represents the sample data. When the data are being obtained via a randomized sampling design p, it is natural to require design unbiasedness for the estimating function,

namely, that

$$E_p\{g(\chi_s, \theta)\} = \sum_{j=1}^{N} \phi_j(y_j, \theta) \tag{19}$$

for each population array $\mathbf{y}$ and parameter value θ. Here, E_p denotes expectation under the sampling design. In particular, if the inclusion probabilities π_j are all positive, $j = 1, \ldots, N$, then the function

$$g^*(\chi_s, \theta) = \phi_s(\theta) = \sum_{j \in s} \frac{\phi_j(y_j, \theta)}{\pi_j} \tag{20}$$

satisfies (19).

As anticipated, g^* of (20) is optimal in certain senses, as seen in the following theorem.

THEOREM 1. *(Godambe and Thompson, 1986a): If $Y_1, \ldots, Y_N$ are independent and the estimating function terms are unbiased as in (15), and if the sampling design is independent of* $\mathbf{Y}$*, then among all g satisfying (19), g^* can be shown to minimize each of*

$$\mathcal{E}E_p g^2 / \left(\mathcal{E}E_p \frac{\partial g}{\partial \theta}\right)^2, \mathcal{E}E_p g^2, \text{ and } \mathcal{E}E_p \left(g - \sum_{j=1}^{N} \phi_j(Y_j, \theta)\right)^2 \tag{21}$$

for all $\xi \in C$.

The independence of $\mathbf{Y}$ and the sampling design is important, in that the proof requires that

$$\mathcal{E}\phi(Y_j/\pi_j) = 0$$

for each j, a condition that might be violated if π_j depends on Y_j. For example, the estimating function leading to (13) does not satisfy the conditions of the theorem.

The independence of the $Y_1, \ldots, Y_N$ is required only because it implies the orthogonality (under the model) of Y_j/π_j and candidate estimating functions evaluated for samples not containing j. It can be relaxed in specific cases.

Note that the estimators resulting from optimal estimating functions need not themselves be unbiased. For example, the usual estimator under simple random sampling of a population ratio R is not design unbiased but nevertheless comes from an optimal unbiased estimating function under the criteria of the theorem.

The corresponding theorem for a multivariate parameter is a straightforward extension. For p-dimensional g satisfying (19), let $J(\chi_s, \theta)$ be the matrix with abth element $\partial g_a/\partial \theta_b$. Let $J^*(\chi_s, \theta)$ be the corresponding matrix for g^* of (20).

THEOREM 2. *Assume that $Y_1, \ldots, Y_N$ are independent, with distribution depending on a p-dimensional parameter θ. Assume that (15) holds and that the sampling design is independent of* $\mathbf{Y}$*. Then for the matrix version of each criterion in THEOREM 1, the difference between the criterion value for g and its value for g^* is non-negative definite for all θ.*

For example, corresponding to the first criterion,

$$(\mathcal{E}E_p J)^{-1}[\mathcal{E}E_p gg^{\tau}](\mathcal{E}E_p J^{\tau})^{-1} - (\mathcal{E}E_p J^*)^{-1}[\mathcal{E}E_p g^* g^{*\tau}](\mathcal{E}E_p J^{*\tau})^{-1}$$

is non-negative definite, where τ denotes transpose.

Example 4.3. Suppose that under the model ξ, the model for $\mathbf{Y}$ is a regression model, expressible as

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon, \tag{22}$$

where the components of $\mathbf{X}$ are fixed and the components of ϵ are independent and identically distributed with mean 0. The optimal census estimating equations for β are

$$\sum_{j=1}^{N} x_j(Y_j - x_j^{\tau}\beta) = 0, \tag{23}$$

the root of which defines the population or census regression parameter B_N or B.

The optimal sample estimating equation system is easily seen to be

$$\sum_{j\in s} \frac{x_j(Y_j - x_j^{\tau}\beta)}{\pi_j} = 0, \tag{24}$$

and the corresponding estimator for β and B is

$$\hat{B}_s = \left(\sum_{j\in s} \frac{(x_j x_j^{\tau})}{\pi_j}\right)^{-1} \sum_{j\in s} \frac{(x_j y_j)}{\pi_j}.$$

At the same time, it is clear from (23) that if the first component of x is constant, then

$$T_y = T_{\mathbf{x}} B, \tag{25}$$

and hence that the regression estimator $T_{\mathbf{x}}\hat{B}_s$ of T_y may be regarded as optimal under the given model and the sampling design.

Thus, whenever the survey weights are determined so as to make the estimators of population totals regression estimators, the resulting estimates can be thought of as optimal in the sense of the theorems when the model (22) is correct.

5. Asymptotic properties of sample estimating functions and their roots

5.1. Single parameter case

Let us introduce the notation $\phi_s(\theta)$ for the left-hand side of the sample estimating equation

$$\sum_{j\in s} \phi_j(y_j, x_j, \theta)/\pi_j = 0;$$

and let $\hat{\theta}_s$ denote the solution when it exists. The first part of the following discussion is taken from Thompson (1997), Chapter 4.

The key to establishing properties of consistency and asymptotic normality for the estimator $\hat{\theta}_s$ is the Taylor series expansion

$$-\tilde{\phi}_s(\theta) = \tilde{\phi}_s(\hat{\theta}_s) - \tilde{\phi}_s(\theta) = \frac{\partial \tilde{\phi}_s}{\partial \theta}(\hat{\theta}_s - \theta) + \frac{1}{2}\frac{\partial^2 \tilde{\phi}_s}{\partial \theta^2} \mid_{\bar{\theta}} (\hat{\theta}_s - \theta)^2, \tag{26}$$

where $\tilde{\phi}_s(\theta) = \phi_s(\theta)/N$ and $\bar{\theta}$ is some value between $\hat{\theta}_s$ and θ. This expansion can be formed when ϕ_s and $\partial\phi_s/\partial\theta$ are continuous in θ and where $\partial^2\phi_s/\partial\theta^2$ exists in the region of interest. The region of interest might be a bounded region of the parameter space known to contain all θ_N in its interior. Suppose it can be shown in some asymptotic framework that $\hat{\theta}_s$ eventually exists uniquely and that

(1) $(\hat{\theta}_s - \theta_N)/\theta_N \to 0$ in probability,
(2) for θ in the region of interest, $\mathrm{Var}_p\{\tilde{\phi}_s(\theta)\}$ and $\mathrm{Var}_p(\partial\tilde{\phi}_s/\partial\theta)$ are of order $O(n^{-1})$ for some measure of sample size n,
(3) $\partial\tilde{\phi}_s/\partial\theta$ and its expectation approach a continuous function $A(\theta) \neq 0$ for θ in that region, and
(4) $\partial^2\tilde{\phi}_s/\partial\theta^2$ is uniformly bounded in probability near θ_N.

In applications to complex designs, the measure of sample size n would not necessarily be the actual number of units in the sample but would be assumed to grow proportionally to the amount of information in the sample. Then by solving (26) at $\theta = \theta_N$ for $\hat{\theta}_s - \theta_N$, we can see that

$$\hat{\theta}_s - \theta_N \text{ is of order } O_p(n^{-1/2})$$

(which is a stronger assertion than condition 1 in the sense of giving an order of convergence) and

$$\hat{\theta}_s - \theta_N = \frac{-\tilde{\phi}_s(\theta_N)}{E_p(\partial\tilde{\phi}_s/\partial\theta) \mid_{\theta=\theta_N}} + O_p(n^{-1}). \tag{27}$$

Then, since $\phi_s(\theta_N)$ has expectation 0, the bias in $\hat{\theta}_s$ is of order $O(n^{-1})$ while its root mean squared error is of order $O(n^{-1/2})$. Thus, it is generally the case in this framework that $\hat{\theta}_s$ is *asymptotically design-unbiased*, in the sense that

$$E_p(\hat{\theta}_s - \theta_N)/\sqrt{E_p(\hat{\theta}_s - \theta_N)^2}$$

approaches 0 asymptotically. Moreover, if conditions are right for the asymptotic normality of the sample sum $\tilde{\phi}_s$, we can conclude that $\sqrt{n}(\hat{\theta}_s - \theta_N)$ is asymptotically normal, with mean 0 and approximate variance $n\mathrm{Var}_p(\tilde{\phi}_s(\theta_N))/[E_p\partial\tilde{\phi}_s/\partial\theta]^2 \mid_{\theta=\theta_N}$. This is the basis of the common practice of estimating the mean squared error of $\hat{\theta}_s$ by

$$v(\phi_s)/(\partial\phi_s/\partial\theta)^2 \tag{28}$$

evaluated at $\theta = \hat{\theta}_s$, where $v(.)$ is the form of a design-unbiased or design-consistent estimator of the variance of a Horvitz–Thompson sample sum.

The condition that $(\hat{\theta}_s - \theta_N)/\theta_N \to 0$ in probability is a design-consistency condition. It would follow, for example, if it could be shown that $\tilde{\phi}_s(\theta_N)/\theta_N$ approaches 0 in probability (a consequence of condition 2 if θ_N^{-1} is bounded) and that the convergence to the nonstochastic limit in condition 3 is uniform (so that a first-order Taylor series

expansion suffices). See Yuan and Jennrich (1998) for a discussion of conditions for estimating function asymptotics.

Suppose $\theta = \theta(\xi)$ is a parameter of a superpopulation distribution under which $Y_1, \ldots, Y_N$ are independent, and suppose the terms of the census estimating function $\sum_{j=1}^{N} \phi_j(Y_j, \theta) = 0$ are unbiased with respect to ξ. Let the sampling design be independent of $\mathbf{Y}$. Then, the sample estimating function $\phi_s(\theta)$ is also an unbiased estimating function for θ. Under regularity conditions on the development of the model and the sampling design, the same expansion as in (26) applies. We can conclude that under the model, or the model and design combined, $\sqrt{n}(\hat{\theta}_s - \theta)$ is asymptotically normal, with mean 0 and approximate variance $n\mathcal{E}(\tilde{\phi}_s^2(\theta)/[\mathcal{E}\partial\tilde{\phi}_s/\partial\theta]^2$. The model mean squared error of $\hat{\theta}_s$ as an estimator of θ can be estimated by an expression like (28), where $v(.)$ is the form of a model-based estimator of variance. Rubin-Bleuer and Schiopu Kratina (2005) have given a formal treatment of the asymptotics in a two-phase framework.

But even if the sampling design is not independent of $\mathbf{Y}$, the sample estimating function $\phi_s(\theta)$ is unbiased under the combined expectation $\mathcal{E}E_p$. It is not difficult to see that the expression in (28) is in that case a justifiable estimator for the approximation (from linearization) of the mean squared error

$$\mathcal{E}E_p(\hat{\theta}_s - \theta_N)^2.$$

(An adjustment would be required to estimate the mean squared error $\mathcal{E}E_p(\hat{\theta}_s - \theta(\xi))^2$.) To the extent to which the design-based and model-based estimators of variance are close to each other, the use of estimating functions leads naturally to inferences which are valid in both frameworks.

The theory does not apply directly to the estimation of the population distribution function and quantiles because the estimating function for the distribution function is not continuous in θ. The most natural approach for quantiles is to assume that as $N \to \infty$, the population c.d.f. $F_N(y)$ uniformly (with error $O(N^{-1/2})$) approaches a c.d.f. $F(y)$, which is continuous, and has a continuous positive derivative f in the neighborhood of the quantile of interest. The next step is to establish a Bahadur representation for the sample quantile $\hat{\theta}_s$:

$$\hat{\theta}_s - \theta_N = \frac{1}{f(\theta_N)}[\hat{F}_s(\theta_N) - F_N(\theta_N)] + o_p(n^{-1/2}). \tag{29}$$

Francisco and Fuller (1991) have given sufficient conditions for the representation (29) to hold.

The theory also needs some modification to deal with sample estimating functions where the inverse inclusion probability weights are replaced by sample-dependent weights, since such sample estimating functions are not design-unbiased in general and since assumptions on the construction or evolution of the weights must be incorporated in the asymptotic framework. Rao et al. (2002) have set out linearization variance estimation and some asymptotic theory for estimating functions with poststratification weights.

5.2. Multivariate parameter case

The same theory is applicable in the multivariate case as in the univariate case. Under analogous conditions, the quantities $\sqrt{n}(\hat{\theta}_s - \theta_N)$ and $\sqrt{n}(\hat{\theta}_s - \theta)$ are asymptotically

p-variate normal with 0 mean. If we here define the matrix $J_s(\theta)$ to have abth element $\partial\phi_{as}/\partial\theta_b$, then the *sandwich estimator*

$$J_s(\theta)^{-1} v(\phi_s) J_s^{\tau}(\theta)^{-1} \tag{30}$$

(analog of (28)) is a robust estimator of the variance–covariance matrix of $\hat{\theta}_s$ as an estimator of θ_N or θ, with proper choice of the estimator form $v(.)$.

6. Interval estimation from estimating functions

6.1. Interval estimation from approximate normality

Let θ be a real parameter. In the notation of previous sections, we have seen that under appropriate conditions, in large samples, we can take

$$\frac{\phi_s(\theta) - \sum_{j=1}^{N} \phi_j(y_j, x_j, \theta)}{\sqrt{v(\phi_s)}}$$

to be approximately standard normal, where $v(\phi_s)$ is a design-consistent estimator of the variance of $\phi_s(\theta)$. This fact suggests two possibilities for constructing interval estimates for the finite population parameter θ_N.

For one possibility, let $\hat{v}(\phi_s)$ be $v(\phi_s)$ with θ replaced by $\hat{\theta}_s$ so that it is calculable from the sample. If ϕ_s is a monotone function of θ, we can construct limits for an approximate two-sided $100(1-2\alpha)\%$ confidence interval for θ_N as the values of θ satisfying

$$\phi_s(\theta) = \pm z_{1-\alpha}\sqrt{\hat{v}(\phi_s)}, \tag{31}$$

where $z_{1-\alpha}$ is the $(1-\alpha)$ quantile of the standard normal distribution.

The second possibility is to retain the dependence on θ in $v(\phi_s)$ and to try to find limits that satisfy

$$\frac{\phi_s(\theta)}{\sqrt{v(\phi_s)}} = \pm z_{1-\alpha}. \tag{32}$$

This method will not be applicable so generally because (32) is less likely than (31) to have exactly two solutions in θ. It is suggested here because the left-hand side of (32) may in some cases have a distribution closer to normality than $\phi_s(\theta)/\sqrt{\hat{v}(\phi_s)}$. This is partly because the bias tends to be smaller and partly because the distribution may be narrower. Analogously, as pointed out by Godambe and Thompson (1999), if Y_j, $j \in s$ are i.i.d. $N(\theta, \sigma^2)$, then the pivot

$$\tau = \frac{\sum_{j\in s}(Y_j - \theta)}{\sqrt{\sum_{j\in s}(Y_j - \theta)^2}}$$

is closer to $N(0, 1)$ than is the t-statistic, since $\mathrm{Var}(\tau) = 1$, and the kurtosis of τ is $3 - 6/(n+2)$.

In both cases, the interval for θ_N consists of values θ for which the hypothesis H : $\theta_N = \theta$ would not be rejected by a corresponding two-sided significance test at level 2α, assuming the approximate normality of the corresponding root or *pivot*.

Example 6.4. If θ_N is a population ratio R, there are several approximate confidence intervals to be considered for use with simple random sampling. Confidence intervals corresponding to (31) would come from the pivot

$$\frac{\phi_s(R)}{\sqrt{\hat{v}(\phi_s)}} = \frac{\sum_{j\in s}(y_j - Rx_j)}{\sqrt{\left(1-\frac{n}{N}\right)\frac{n}{n-1}\sum_{j\in s}(y_j - \hat{R}x_j)^2}}. \tag{33}$$

Confidence intervals corresponding to (32) could come from the pivot

$$\frac{\phi_s(R)}{\sqrt{v(\phi_s)}} = \frac{\sum_{j\in s}(y_j - Rx_j)}{\sqrt{\left(1-\frac{n}{N}\right)\frac{n}{n-1}\sum_{j\in s}(z_j - \overline{z}_s)^2}}, \tag{34}$$

or alternatively

$$\frac{\phi_s(R)}{\sqrt{v(\phi_s)}} = \frac{\sum_{j\in s}(y_j - Rx_j)}{\sqrt{\frac{N}{N-1}\left(1-\frac{n}{N}\right)\sum_{j\in s} z_j^2}}, \tag{35}$$

where $z_j = y_j - Rx_j$. It is shown by Thompson (1997, p. 101) that the design expectation of (34), where R is retained in the denominator, is closer to 0 than the design expectation of (33), where R is replaced by an estimate in the denominator.

Thus, for particular populations, the distribution of (34) and (35) may be closer to standard normal than the distribution of (33). The approximate confidence intervals based on the first two quantities are, respectively,

$$\hat{R}_s \pm z_{1-\alpha}\sqrt{v(\overline{y}_s) - 2\hat{R}_s\mathrm{cov}(\overline{y}_s, \overline{x}_s) + \hat{R}_s^2 v(\overline{x}_s)}/\overline{x}_s \tag{36}$$

and

$$\frac{\hat{R}_s - z^2 b_s}{1 - z^2 a_s} \pm \frac{z\sqrt{z^2(b_s^2 - a_s c_s) + c_s - 2b_s\hat{R}_s + a_s\hat{R}_s^2}}{1 - z^2 a_s}, \tag{37}$$

where $z = z_{1-\alpha}$, $a_s = v(\overline{x}_s)/\overline{x}_s^2$, $b_s = \mathrm{cov}(\overline{x}_s, \overline{y}_s)/\overline{x}_s^2$, $c_s = v(\overline{y}_s)/\overline{x}_s^2$.

The limits in (37) were used by Fieller (1932) and have been discussed in a sampling context by Cochran (1977, p. 156).

The next example deals with estimation of the corresponding superpopulation parameter.

Example 6.5. Suppose under the superpopulation model ξ, we have

$$Y_j = \beta x_j + \epsilon_j$$

for all $j = 1, \ldots, N$, where x_j is real, and ϵ_j has mean 0 and variance $\sigma^2 x_j$. The optimal census estimating function for β is well known to be

$$\sum_{j=1}^{N}(y_j - \beta x_j),$$

and thus the corresponding census parameter is the population ratio R. Under the model, the pivot

$$\frac{\phi_s(\beta)}{\sqrt{v(\phi_s)}} = \frac{\sum_{j\in s}(y_j - \beta x_j)}{\sqrt{\sum_{j\in s}(y_j - \beta x_j)^2}} \tag{38}$$

is approximately $N(0, 1)$, in a manner that is robust to misspecification of the variance function. At the same time, if the design is simple random sampling, the same pivot is approximately $N(0, 1)$ under the combined model of ξ and the design, and it is thus suitable for providing confidence intervals for β. The similarity between (35) and (38) can reinforce the validity of both.

6.2. *Interval estimation from inverse testing*

Certain improvements to interval estimation are available, depending on more refined approximations than normality. They can be described in an inverse testing formulation. For simplicity, in this section, we suppose that the design is simple random sampling and that ϕ_j is not dependent on j.

The inverse testing method works by excluding from the confidence set for θ_N all values of θ, which would be rejected by a level-α significance test of the hypothesis $\theta_N = \theta$. For a given value of θ, if the hypothesis were true, the population arrays $\mathbf{x}$, $\mathbf{y}$ would satisfy

$$\sum_{j=1}^{N} \phi(y_j, x_j, \theta) = 0. \tag{39}$$

As already mentioned, the intervals from (31) and (32) have inverse testing interpretations.

Thus, the first step is to imagine artificial population arrays $\mathbf{x}(\theta)$, $\mathbf{y}(\theta)$ which are consistent with the sample values, which satisfy (39), and which have an associated distribution close to the one found in the sample. This enables the conceptual imputation of $\phi(y_j(\theta), x_j(\theta), \theta)$ for the unseen units of the population. The second step is to approximate the simple random sampling distribution of $\phi_s(\theta)$ in this artificial population by an appropriate approximation or by simulation. In particular, we approximate the probability that $\phi_s(\theta)$ differs from zero by more than its observed sample value ϕ_s^0, given the artificial population. If this probability is less than or equal to α, the value of θ is excluded. Thus, if $\phi_s(\theta)$ were decreasing in θ, an upper confidence limit $\hat{\theta}_U$ would satisfy

$$P(\phi_s(\hat{\theta}_U) \le \phi_s^0(\hat{\theta}_U) \mid \mathbf{x}(\hat{\theta}_U), \mathbf{y}(\hat{\theta}_U)) = \alpha,$$

whereas a lower confidence limit $\hat{\theta}_L$ would satisfy

$$P(\phi_s(\hat{\theta}_L) \ge \phi_s^0(\hat{\theta}_L) \mid \mathbf{x}(\hat{\theta}_L), \mathbf{y}(\hat{\theta}_L)) = \alpha,$$

approximately. The two limits together would define a $100(1 - 2\alpha)\%$ two-sided confidence interval.

One way of constructing the population arrays $\mathbf{x}(\theta)$, $\mathbf{y}(\theta)$ would be as follows. Suppose for simplicity that y is real and x is a scalar constant. Let $y^1, \ldots, y^r$ be the distinct

sampled values of y, occurring, respectively, $a_1, \ldots, a_r$ times. Define population weights $w_1, \ldots, w_r$ with the intention that in the array $\mathbf{y}(\theta)$, there should be w_i occurrences of y^i for $1 \le i \le r$. Thus, we would aim to have the weights $w_i = w_i(\theta)$ satisfy

$$\sum_{i=1}^{r} w_i = N, \tag{40}$$

$$\sum_{i=1}^{r} w_i \phi(y^i, \theta) = 0. \tag{41}$$

We would then require that the population distribution of y be close to the sample distribution in some sense. For example, suppose we require that the Kullback–Leibler distance $\sum_{i=1}^{r} w_i \log(w_i/a_i)$ be minimized. Then,

$$w_i = Nq_i / \sum_{i=1}^{r} q_i, \tag{42}$$

where

$$q_i = a_i \exp\{t\phi(y^i, \theta)\} \tag{43}$$

and $t = t(\theta)$ is a solution of

$$\sum_{i=1}^{r} a_i \phi(y^i, \theta) e^{t\phi(y^i, \theta)} = 0. \tag{44}$$

For another example, we could require that the empirical likelihood be maximized and that would correspond to minimizing $\sum_{i=1}^{r} a_i \log(a_i/w_i)$, yielding

$$w_i = Nq_i / \sum_{i=1}^{r} q_i, \tag{45}$$

where

$$q_i = a_i/(1 + t\phi(y^i, \theta)) \tag{46}$$

and $t = t(\theta)$ is a solution of

$$\sum_{i=1}^{r} a_i \phi(y^i, \theta)/(1 + t\phi(y^i, \theta)) = 0. \tag{47}$$

The weights satisfying (42)–(47) exactly will not in general correspond to an array $\mathbf{y}(\theta)$; the weights actually used would be approximations of these which fulfilled the further condition of being positive integers greater than the corresponding sample frequencies.

Related ideas are seen in some formulations of the finite population bootstrap (see Section 7), but more closely in the 'scale load' approach of Hartley and Rao (1968). An empirical likelihood approach for single stage complex designs has been developed in several papers including Chen and Sitter (1999) and Wu and Rao (2006). In the simple context above, an empirical likelihood ratio would be defined from the probabilities

w_i/N and a chi-squared approximation to its distribution would be used for testing and estimation.

7. Bootstrapping estimating functions

The bootstrap t method (see, e.g., DiCiccio and Romano, 1988) would aim to find an approximate distribution for

$$\phi_s(\theta)/\sqrt{v(\phi_s)} \tag{48}$$

or for

$$\phi_s(\theta)/\sqrt{\hat{v}(\phi_s)} \tag{49}$$

by resampling. The resampling could take place from an imagined artificial population, as in Section 6.2, or from the sample in a manner adjusted for the sizes of the sample components. The method would then solve for θ equations of the form (31) or (32) with $\pm z_{1-\alpha}$ replaced with $\hat{H}^{-1}(1-\alpha)$ and $\hat{H}(\alpha)$, the $1-\alpha$ and α quantiles of the resampling distribution. If the equations had unique solutions, these would serve as endpoints for an approximate confidence interval for θ_N.

The estimating function bootstrap of Hu and Kalbfleisch (2000), originally proposed for independent sampling, finds the approximate distribution of the pivot by resampling from the terms of ϕ_s at the point estimate $\hat{\theta}_s$, then proceeds in a similar manner with the pivot of (48).

For application to complex surveys, it is important to note that for the bootstrap to work, the denominator $v(\phi_s)$ need not be a consistent estimator of the variance of ϕ_s. However, if it differs from a consistent estimator by a close-to-constant factor, the resulting confidence intervals will be better.

The most commonly used bootstrap method for complex survey designs is the Rao and Wu (1988) method. A review of its application to estimating functions is given by Rao (2006). An earlier paper by Rao and Tausi (2004) investigated a jackknife version of the Hu and Kalbfleisch estimating function resampling. The Rao–Wu bootstrap is valid for stratified multistage designs and is implemented through the provision of sets of bootstrap weights. Each set of weights is to yield a point estimate of θ so that the user can estimate (usually a little conservatively) the variance of $\hat{\theta}_s$ as the variance in the ensemble of point estimates. Letting b index the bootstrap samples, it has been found that numerically stable point estimates can be obtained from the first Newton–Raphson iteration of the solution as

$$\tilde{\theta}(b) = \hat{\theta}_s - [J_s(\hat{\theta}_s)]^{-1}\phi_{sb}(\hat{\theta}_s),$$

and the variance-covariance of θ_s can be estimated as

$$B^{-1}\sum_{b=1}^{B}[\tilde{\theta}(b) - \hat{\theta}_s][\tilde{\theta}(b) - \hat{\theta}_s]^{\tau}$$

– or effectively as a transformed bootstrap variance–covariance of the components of $\phi_s(\theta)$ evaluated at θ_s.

See also Roberts et al. (2006) for an application to survey weighted Generalized Estimating Equation estimation in the (longitudinal) National Population Health Survey of Statistics Canada.

8. Multivariate and nuisance parameters

For simultaneous estimation of the components of a multivariate parameter without resampling point estimates, the considerations in Sections 6 and 7 can be extended in a straightforward manner to produce confidence regions based on chi-squared or other approximations to the distribution of the pivot

$$\phi_s^\tau(\theta)v^{-1}(\phi_s)\phi_s(\theta), \tag{50}$$

where v is an estimated variance–covariance matrix for $\phi_s(\theta)$.

The estimating function approach can also be extended to the estimation of quantities, which need other parameters, or "nuisance parameters," and a system of estimating functions for their definition.

Example 8.6. The *variance* $\sigma_N^2 = \sum_{j=1}^N (y_j - \mu_{\mathbf{y}})^2/N$ satisfies

$$\sum_{j=1}^N [(y_j - \lambda_N)^2 - \sigma_N^2] = 0 \tag{51}$$

$$\sum_{j=1}^N (y_j - \lambda_N) = 0; \tag{52}$$

here, $\lambda_N = \mu_{\mathbf{y}}$, another parameter which must be estimated and which may be regarded as a nuisance parameter.

Example 8.7. The *mean* of a stratified population can be written as

$$\theta_N = \sum_{h=1}^H W_h\theta_h, \tag{53}$$

with stratum mean parameters θ_h satisfying the system

$$\sum_{j=1}^N \delta_{jh}(y_j - \theta_h) = 0, h = 1, \ldots, H. \tag{54}$$

The indicator $\delta_{jh} = 1$ if $j \in \mathcal{S}_h$, or $\delta_{jh} = 0$ otherwise. Viewing θ_N as the quantity of interest and θ_h, $h = 1, \ldots, H-1$, as a vector-valued nuisance parameter can lead naturally to a justification of poststratification (Binder and Patak, 1994).

Example 8.8. The *regression coefficients* B_N and A_N satisfy

$$\sum_{j=1}^N x_j(y_j - B_N x_j - A_N) = 0 \tag{55}$$

$$\sum_{j=1}^N (y_j - B_N x_j - A_N) = 0; \tag{56}$$

here, it might be the case that B_N is of interest, while A_N is a nuisance parameter.

In general, let us think of a system of population estimating functions

$$\sum_{j=1}^{N} \phi_{1j}(y_j, x_j; \theta_N, \lambda_N) = 0 \tag{57}$$

$$\sum_{j=1}^{N} \phi_{2j}(y_j, x_j; \theta_N, \lambda_N) = 0 \tag{58}$$

with (52) and (53) having the dimensions of θ_N and λ_N, respectively. Typically, these equations would have the form of population maximum likelihood equations for θ_N and λ_N. Suppose that θ_N is the parameter of interest, while λ_N is a nuisance parameter.

The sample version of this estimating function system at a general parameter value (θ, λ) is

$$\phi_{1s}(\theta, \lambda) = \sum_{j \in s} \frac{\phi_{1j}(y_j, x_j; \theta, \lambda)}{\pi_j} \tag{59}$$

$$\phi_{2s}(\theta, \lambda) = \sum_{j \in s} \frac{\phi_{2j}(y_j, x_j; \theta, \lambda)}{\pi_j}. \tag{60}$$

If $\hat{\lambda}_\theta$ satisfies $\phi_{2s}(\theta, \hat{\lambda}_\theta) = 0$, then the estimating equation system to be solved for the estimate $\hat{\theta}_s$ of θ_N becomes the *profile estimating function*

$$\phi_{1s}(\theta, \hat{\lambda}_\theta) = 0. \tag{61}$$

Binder and Patak (1994) have shown that to a first-order approximation (for real θ), the MSE of $\phi_{1s}(\theta, \hat{\lambda}_\theta)$ can be estimated by

$$v\left(\sum_{j \in s} \frac{z_{\theta j}}{\pi_j}\right),$$

where v is a variance estimator form and

$$z_{\theta j} = \phi_{1j}(y_j, x_j; \theta, \hat{\lambda}_\theta) - \hat{J}_{1\lambda} \hat{J}_{2\lambda}^{-1} \phi_{2j}(y_j, x_j; \theta, \hat{\lambda}_\theta), \tag{62}$$

with

$$\hat{J}_{i\lambda} = \sum_{j \in s} \frac{1}{\pi_j} \frac{\partial}{\partial \lambda} \phi_{ij}(y_j, x_j; \theta, \lambda) \mid_{\hat{\lambda}_\theta}, \quad i = 1, 2.$$

Note that $\sum_{j \in s}(z_{\theta j}/\pi_j)$ is the combination of the estimating functions in (59) and (60) that changes least as the nuisance parameter λ changes, near $\hat{\lambda}_{\hat{\theta}}$. Interval estimates for θ_N are then obtainable from an $N(0, 1)$ approximation to the distribution of

$$\frac{\phi_{1s}(\theta, \hat{\lambda}_\theta)}{\sqrt{v(\sum_{j \in s} z_{\theta j}/\pi_j)}}. \tag{63}$$

This approximation is likely to be particularly effective if $\phi_{2s}(\theta, \lambda)$ is linear in λ. In some situations where the numerator of (63) is significantly biased as an estimating function for θ, improvements may be expected from modifications that reduce the bias.

A further alternative would be to use a $N(0, 1)$ approximation to the distribution of

$$\frac{\phi_{1s}(\theta, \hat{\lambda}_\theta)}{\sqrt{v(\sum_{j\in s} \tilde{z}_j/\pi_j)}}, \tag{64}$$

where $\tilde{z}_j$ is $z_{\theta j}$ evaluated at $\hat{\theta}$. See Hu and Kalbfleisch (2000) for a discussion of the relative merits of pivots like (63) and (64), and prescriptions for associated bootstrapping.

Example 8.9. Consider the estimation of population variance σ_N^2.
Since $\hat{\lambda}_\sigma = \hat{\lambda} = \hat{T}_{\mathbf{y}}/\hat{N} = \hat{\mu}_{\mathbf{y}}$, then

$$\begin{aligned}
\phi_{1s}(\sigma, \hat{\lambda}_\sigma) &= \sum_{j\in s}((y_j - \hat{\mu}_{\mathbf{y}})^2 - \sigma^2)/\pi_j;\\
\hat{\sigma}^2 &= \left[\sum_{j\in s}(y_j - \hat{\mu}_{\mathbf{y}})^2/\pi_j\right]/\hat{N};\\
\hat{J}_{1\lambda} &= 0, \quad \hat{J}_{2\lambda} = -\hat{N};\\
z_{\theta j} &= (y_j - \hat{\mu}_{\mathbf{y}})^2 - \sigma^2; \tilde{z}_j = (y_j - \hat{\mu}_{\mathbf{y}})^2 - \hat{\sigma}^2.
\end{aligned}$$

According to the prescription above, interval estimates of σ^2 are obtained by setting (63) or (64) equal to $N(0, 1)$ quantiles and solving.

At the same time, correcting the profile likelihood equation for bias is likely to produce improved accuracy. For example, if the design is simple random sampling, the corrected estimating function

$$\phi_{1cs}(\sigma, \hat{\lambda}_\sigma) = \sum_{j\in s} \frac{N}{n}\left((y_j - \hat{\mu}_{\mathbf{y}})^2 - \sigma^2 + \frac{N-n}{N-1}\sigma^2\right)$$

is unbiased and can be used as a starting point for inference.

9. Estimating functions and imputation

We conclude with a note on estimating functions and missing data.

It is well known that in cases of unit nonresponse, an unbiased estimator of the complete data estimating function can be obtained by inverse response probability weighting of the observed data estimating function. Godambe and Thompson (1986b) have provided an optimality theorem for this case. Although it tends to be inefficient for parametric models (Lawless et al., 1999), this form is widely used in biostatistics for its simplicity and robustness (see, e.g., Robins et al., 1995).

However, item nonresponse invites imputation, practically or conceptually. Beaumont (2005) and Haziza and Rao (2006) have taken an estimating function approach to imputation for regression estimation of a population total. In particular, they have made use of systems of estimating equations for joint estimation of the response probability parameters, the regression parameters, and the imputed estimator.

Imputation becomes relatively simple when estimation of a finite population parameter involves solution of an estimating equation with terms that are superpopulation unbiased. Suppose that some terms in the equation are missing because the corresponding observations are incomplete. Suppose that the absence of these terms does not change the fact that the terms that are present have superpopulation expectation 0. Then, if we impute 0 for each of the missing terms, the resulting estimating equation is still superpopulation unbiased.

Suppose also that the terms of the complete data estimating function are independent and that the absence of certain terms does not change the lack of correlation of the terms remaining, under the superpopulation model ξ. (The conditions of unchanged mean 0 and unchanged lack of correlation constitute an estimating function version of missingness at random.) Then, a quantity such as

$$\frac{\sum_{j\in s}\phi_j(Y_j,\theta)}{\sqrt{\sum_{j\in s}\phi_j^2(Y_j,\theta)}} \tag{65}$$

is still an approximately standard normal pivot under the superpopulation model if we impute 0 for each of the missing terms. Moreover, the imputed estimating function is optimal in a certain sense: it is closest to the complete data estimating function in terms of superpopulation variance of the difference. (Godambe and Thompson, 2006).

By an extension, classical mean imputation and regression imputation can be regarded as examples of the imputation of 0 for missing estimating function terms. For the estimation of a population total, consider the system

$$T_Y - \sum_{i=1}^{N} x_i\beta - \sum_{i\in s}\frac{(Y_i - x_i\beta)}{\pi_i} = 0,$$
$$\sum_{i\in s}\frac{x_i(y_i - x_i\beta)}{\pi_i} = 0,$$

where β is the regression coefficient of y on x. The first equation defines a regression estimator for T_Y with unknown parameter β; the second is an estimating equation system for β itself. The terms in the estimating functions are independent. If some of the sample y_is are missing and we impute the corresponding elementary estimating functions $y_i - \beta x_i$ and $x_i(y_i - \beta x_i)$ with 0s, we are effectively imputing for missing y_i the value $x_i\beta$ in the estimating equations and the value $x_i\hat{B}_{s'}$ in the estimates, where s' is the part of the sample with complete observations. The justification for this as presented depends on the correctness of the model. Indeed, under a nonresponse model, the sampling expectation of the left-hand side of the first equation is not 0 but could be regarded as

$$\sum_{i=1}^{N}(1-\alpha_i)(Y_i - x_i\beta),$$

where α_i is the probability that unit i, if sampled, is observed.

If we impute 0 for missing terms in an approximate pivot from a finite population sample, such as (33)–(35), we are effectively assuming that the imputed census estimating function is very close to 0 at the true census parameter. This would be true, for example, if those in the population who would have responded constitute a fairly large

Bernoulli sample of the whole. Other conditions more suited to complex designs can also be formulated.

More generally, in terms of the superpopulation model, often the most natural imputation of a missing estimating function term is its conditional expectation under ξ, given the data which are present (McLeish, 1984). In fact, if the complete data estimating function is a score function, its conditional expectation is also a score function and is optimal. Wang and Chen (2006) have considered the case, sometimes applicable in sampling theory, where independent observations are of form (X, Y), and Y is sometimes missing. They propose a method of multiple estimation of each missing estimating function term from a kernel smoothed estimate of the conditional distribution of Y given X. The imputed terms are then used to produce an empirical likelihood ratio statistic. The distribution of the empirical likelihood ratio is estimated through resampling with a bootstrap that incorporates the imputation method, and confidence intervals for θ are obtained by the inversion of an empirical likelihood ratio test.

At the same time, as is well known, imputation via projection does not fulfill all purposes. Consider a superpopulation model in which $Y_1, \ldots, Y_N$ form the initial segment of an AR(1) time series with mean 0 and variance 1, and the census estimating equation for the parameter θ is

$$\sum_{j=2}^{N} Y_{j-1}(Y_j - \theta Y_{j-1}) = 0.$$

The left-hand side is a martingale, and the estimating function for θ_N from a sample is most naturally a martingale also:

$$\sum_{i=2}^{n} a_i(\theta) Y_{j_{i-1}}(Y_{j_i} - \theta^{(j_i - j_{i-1})} Y_{j_{i-1}}),$$

where $j_1, \ldots, j_n$ is the sample in order, and the a_i are determined for maximum efficiency. In a sense, we can think of the terms in the sample estimating function as imputing the sums of unseen terms in the census estimating function, but it is not obtained by conditional expectation of the unseen terms given the terms which are present.

Acknowledgment

This work was supported by research grants from the Natural Sciences and Engineering Research Council of Canada. We would like to thank a reviewer for very helpful comments.

Sample Surveys: Inference and Analysis, Vol. 29B
ISSN: 0169-7161
© 2009 Elsevier B.V. All rights reserved
DOI: 10.1016/S0169-7161(09)00227-2

Nonparametric and Semiparametric Estimation in Complex Surveys

F. Jay Breidt and Jean D. Opsomer

1. Introduction

1.1. Nonparametric and semiparametric methods

Nonparametric and semiparametric methods are rich classes of statistical tools that have gained acceptance in most areas of statistics. They make it possible to analyze data, estimate trends and conduct inference without having to fully specify a parametric model for the data. In the survey context, their use is much less widespread. In this chapter, we will focus on nonparametric and semiparametric methods in two important statistical areas: estimation of densities and estimation of regression functions. Both of these areas have applications in survey estimation, for both descriptive and analytical uses.

We begin with a necessarily brief overview of the main nonparametric and semiparametric methods relevant to survey estimation. In this section, we describe them for the case of independent and identically distributed (*iid*) data to introduce the methods. Subsequent sections will deal with the situation in which the observations are obtained from a complex survey.

We would like to note that the terms "nonparametric" and "semiparametric" have not been used consistently in the statistical literature, so there is no agreement on which methods exactly fall into each of these two categories. Generally speaking, nonparametric methods are those that do not assume a parametric form for the main features of interest in the data (though there might be parametric assumptions on some of the "nuisance features," e.g., the variance in the case of regression). In contrast, semiparametric methods use a combination of parametric and nonparametric specification for the main features of interest. Clearly, these descriptions are somewhat subjective and open to interpretation, so one person's nonparametric method is another person's semiparametric approach.

1.2. Kernel methods

Kernel methods are used for both density estimation and regression. We begin by describing the kernel density estimator and restrict ourselves to the univariate case. Suppose we observe $X_1, \ldots, X_n$ and we assume these x_i are *iid* from an unknown density $f_x(\cdot)$. The

density is assumed to be a smooth function of x but otherwise unspecified. *Kernel density estimation* methods aim to estimate the density $f_x(\cdot)$ nonparametrically. Wand and Jones (1995) give a good introduction to these methods, and we only describe the main idea here. For a given value x, a simple kernel density estimator $\widehat{f}_x(x; h)$ is defined as

$$\widehat{f}_x(x; h) = \frac{1}{n\,h} \sum_{i=1}^{n} K\left(\frac{x_i - x}{h}\right), \tag{1}$$

where $K(\cdot)$ denotes the *kernel function* and the constant h is referred to as the *bandwidth*. To estimate the density function $f_x(\cdot)$ using (1), $\widehat{f}_x(\cdot; h)$ is computed at each value x for which an estimator is needed, for instance on a dense grid of x-values, which can then be plotted or interpolated.

The kernel $K(\cdot)$ is usually a symmetric probability density, with the standard normal density being a common choice, but other functions can be used as well. The crucial feature of the kernel function is that it determines distance-based weights for the sample observations, to be used in the construction of (1). The bandwidth h determines the smoothness of the estimator $\widehat{f}_x(x; h)$, with small values of h leading to more "wiggly" estimates and large values resulting in smoother estimates. More precisely, the bandwidth determines the bias-variance trade-off for the kernel density estimator $\widehat{f}_x(x; h)$, with large h having potentially larger bias but smaller variance than small h. A large literature is devoted to the determination of the best value for the bandwidth, and we will briefly return to this issue in later sections.

We now turn to the kernel-based regression estimation problem. Suppose that we have a data set with observations $(X_1, Y_1), \ldots, (X_n, Y_n)$, and we are interested in estimating the function $m(\cdot)$ in the model

$$Y_i = m(x_i) + \varepsilon_i, \tag{2}$$

where $m(\cdot)$ is smooth but not further specified, and for simplicity, we assume that the ε_i are *iid* with mean 0 and variance σ^2. The most commonly used kernel method of nonparametric estimation of $m(\cdot)$ is *local polynomial regression*, with local linear regression a popular choice.

Let q represent the degree of the local polynomial regression. For a given value x, the estimator $\widehat{m}(x)$ is defined as $\widehat{\beta}_0$, where $\widehat{\beta}_0, \ldots, \widehat{\beta}_q$ are found by solving the following weighted least squares problem:

$$\min_{\beta} \sum_{i=1}^{n} K\left(\frac{x_i - x}{h}\right) \left(Y_i - \beta_0 - \beta_1 (x_i - x) - \cdots - \beta_q (x_i - x)^q\right)^2.$$

This estimator can be written in matrix notation as

$$\widehat{m}(x) = \boldsymbol{e}_1^T \left(\boldsymbol{X}_x^T \boldsymbol{W}_x(h) \boldsymbol{X}_x\right)^{-1} \boldsymbol{X}_x^T \boldsymbol{W}_x(h) \boldsymbol{Y}, \tag{3}$$

with $\boldsymbol{e}_1 = (1, 0, \ldots, 0)^T$, $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$, $\boldsymbol{W}_x = \text{diag}\{K((x_1 - x)/h), \ldots, K((x_n - x)/h)\}$, and

$$\boldsymbol{X}_x = \begin{bmatrix} 1 & x_1 - x & \cdots & (x_1 - x)^q \\ \vdots & \vdots & & \vdots \\ 1 & x_n - x & \cdots & (x_n - x)^q \end{bmatrix}. \tag{4}$$

As was the case with kernel density estimation, the function $m(\cdot)$ is estimated by computing $\widehat{m}(x)$ for any value x where an estimator of the function is needed. The bias-variance tradeoff for $\widehat{m}(x)$ also again depends on the bandwidth h. It is clear from (3) that the local polynomial estimator can be written as a linear combination of the Y_i, $\widehat{m}(x) = \sum w_i(x) Y_i$, which will be useful when applying this nonparametric method in the survey context. We refer to Wand and Jones (1995) for further information on local polynomial regression, including its theoretical properties.

1.3. Spline methods and other methods

While kernel methods span both density and regression function estimation, spline methods are typically only used for the latter problem. We therefore again consider a data set with observations $(X_1, Y_1), \ldots, (X_n, Y_n)$, which are assumed to follow the model (2) with *iid* errors. While the function $m(\cdot)$ in (2) is still assumed to be smooth but otherwise unspecified, we now make the additional assumption that it is well approximated by a *spline function*. Polynomial spline functions are defined as

$$m(x; \boldsymbol{\beta}) = \beta_0 + \beta_1 x + \cdots + \beta_p x^p + \sum_{j=1}^{J} \beta_{p+j} (x - \kappa_j)_+^p, \tag{5}$$

where $p \geq 1$ is the order of the spline, $\kappa_1, \ldots, \kappa_J$ are a set of pre-specified breakpoints called *knots* and the function $(\cdot)_+^p$ denotes

$$(x - \kappa)_+^p = \begin{cases} (x - \kappa)^p & \text{if } x > \kappa \\ 0 & \text{otherwise.} \end{cases}$$

The linear ($p = 1$) and cubic ($p = 3$) spline models are common choices in practice. The linear splines are simple and continuous, and the cubic splines match up with a common type of smoothing splines, the natural cubic splines, which arise from a penalized optimization with penalty on the squared second derivative of the function. Other formulations of the spline function $m(x; \boldsymbol{\beta})$ are possible, in which the set of *basis functions* $\{1, x, \ldots, x^p, (x - \kappa_1)_+^p, \ldots, (x - \kappa_J)_+^p\}$ are replaced by a different set. For instance, *B-splines* (de Boor, 2001) are a widely used set of basis functions with better numerical properties than polynomial splines. Most of these formulations, including *B*-splines, can be equivalently rewritten into the above polynomial spline, so that we will restrict our attention to (5).

A number of different spline regression methods exist, but we will focus here on *penalized spline regression* because of its ease of use and relevance to the applications in survey estimation. An excellent overview of this method and its applications in a wide range of regression contexts is provided in Ruppert et al. (2003). It is clear from (5) that $m(x; \boldsymbol{\beta})$ is essentially a parametric function (albeit a complicated one), and deviations from a global pth order polynomial can only occur at the knots so that the flexibility of the spline as a representation of an unknown function is determined by the number and location of the knots. To ensure that $m(x; \boldsymbol{\beta})$ is sufficiently flexible, the penalized spline approach sets the number of knots J to be large, say as high as $J = n/4$, and places them at the appropriate quantiles of the x_i.

Fitting of the spline model to the observations is done by least squares minimization but with a penalty added to ensure the existence of a solution and to reduce the potential

increase in variance due to the large number of parameters needing estimation. Specifically, the estimator of $m(\cdot)$ is $m(\cdot; \widehat{\boldsymbol{\beta}})$ using expression (5), where $\widehat{\boldsymbol{\beta}}$ is the minimizer of

$$\sum_{i=1}^{n}\left(Y_i - \beta_0 - \beta_1 x_i - \cdots - \beta_p x_i^p - \sum_{j=1}^{J}\beta_{p+j}(x_i - \kappa_j)_+^p\right)^2 + \lambda\sum_{j=1}^{J}\beta_{p+j}^2 \tag{6}$$

and λ is a fixed penalty. The penalty λ plays an analogous role to the bandwidth h in the kernel regression methods, in that it determines the bias-variance trade-off for the estimator $m(\cdot; \widehat{\boldsymbol{\beta}})$, with large values of λ resulting in potentially larger bias and smaller variance than small values. Since only the nonpolynomial part of the spline coefficients is penalized in (6), λ determines the amount of deviation from a pth degree polynomial function.

Because of the very flexible nature of the spline function (5) and the presence of the penalty λ that serves as a tuning constant, penalized spline regression is typically considered a nonparametric method. Nevertheless, it also shares many characteristics of parametric regression because the number of parameters is fixed (at $J + p + 1$) and the estimator is found as a solution to a global least squares problem.

Other spline regression methods are (unpenalized) spline regression and smoothing spline regression. In the former, a spline function with a small number of knots is specified and the function is fitted without penalization so that careful attention needs to be paid to knot placement to avoid bias. In smoothing spline regression, the formulation of the approach is different from the above, but the estimator is essentially equivalent to a polynomial spline as in (5) but with a knot at every observation point x_i and a penalty term on the derivative of the function. We do not pursue these methods further here and instead refer to Ruppert et al. (2003, Chapter 3).

Other important classes of nonparametric methods are available, many of which could be adapted for use in survey estimation. Orthogonal decompositions, in particular *wavelet* decomposition (Vidakovič, 1999), is a nonparametric regression method with good statistical properties that is applicable in situations where the mean function is not necessarily smooth. Neural networks (Ripley, 1996) are a class of methods conceptually related to penalized spline regression, in which the parameters are found by nonlinear regression. Finally, methods based on classification such as classification and regression trees (Breiman et al., 1984) and multivariate adaptive regression splines (Friedman, 1991) can be used as nonparametric regression methods.

1.4. Fitting more complex models

So far, we have discussed the situation in which the x_i are univariate observations. In surveys, the number of variables is typically large, so we would like to be able to apply nonparametric and semiparametric methods for multivariate data. In principle, it is indeed possible to directly extend all the methods from the previous sections to the multivariate case, but a number of constraints make this impractical for more than two or three dimensions. One issue is the so-called "curse of dimensionality," which implies that model flexibility has to decrease as the dimension of the covariate space increases to obtain satisfactory fits. This could be done by increasing the amount of smoothing (by using a larger bandwidth or penalty) or using a reduced number of knots (in the case

of splines), but more useful approaches are to replace the fully nonparametric model itself by more restricted model specifications. We discuss two important special cases of such models here: additive models and semiparametric models.

Let $\boldsymbol{X}_i = (X_{1i}, \ldots, X_{iD})^T$ represent a vector of D covariates. In additive models, model (2) is replaced by

$$Y_i = m_1(X_{1i}) + \cdots + m_D(X_{Di}) + \varepsilon_i, \tag{7}$$

where the functions $m_d(\cdot)$ are (typically) univariate and smooth but otherwise not restricted to belong to a specific parametric family. This model was made popular by Hastie and Tibshirani (1990), who proposed estimation methods based on *backfitting*. This approach, which is implemented in S-Plus and R, relies on iteratively applying one-dimensional methods such as local polynomial regression or spline regression to the residuals from the fits with respect to the other covariates. While other methods for fitting model (7) have since been proposed, backfitting remains popular today. When penalized spline regression is used as the fitting method, it is possible to fit model (7) without iterating by writing it as a penalized multiple regression problem, from which the spline parameters can be estimated directly (see Ruppert et al., 2003, for details). The package SemiPar (Wand et al., 2005) implements this approach in R.

In a semiparametric model, a nonparametric term is combined with parametrically specified components. Let $\boldsymbol{Z}_i = (Z_{1i}, \ldots, Z_{Pi})^T$ represent the additional covariates to be modeled parametrically. A typical example of a semiparametric model is

$$Y_i = m(X_i) + \boldsymbol{Z}_i^T \boldsymbol{\beta} + \varepsilon_i, \tag{8}$$

where the nonparametric term $m(X_i)$ could itself be multivariate and modeled as an additive model. Backfitting can be applied to fit model (8) as well, but other methods specially designed for semiparametric models are available. The semiparametric model is particularly useful when some of the covariates in a data set are categorical, which by definition cannot be smoothed.

In addition to nonparametric regression for multivariate data, another important extension is for models with more complex mean structures, including nonparametric equivalents of generalized linear models. The generalized additive model (GAM) described in Hastie and Tibshirani (1990) has mean structure

$$E(Y_i|\boldsymbol{X}_i, \boldsymbol{Z}_i) = g\left(m_1(X_{1i}) + \cdots + m_D(X_{Di}) + \boldsymbol{Z}_i^T \boldsymbol{\beta}\right), \tag{9}$$

which combines a known link function $g(\cdot)$ with a mean additive model or a semiparametric model. This model makes it possible to perform common types of regression such as Poisson or logistic regression nonparametrically. The most common fitting method for this type of model is an iterative algorithm called *local scoring*, a combination of Fisher scoring and backfitting. Just like for additive models, this method uses univariate regression methods such as local polynomial and spline regression for the component functions.

2. Nonparametric methods in descriptive inference from surveys

We now consider the use of nonparametric methods in making inference about a finite, labelled population $U = \{1, \ldots, i, \ldots, N\}$. Associated with each label i are study

variables y_i, z_i, etc (possibly vector-valued), which can in principle be observed without error if label i were sampled. Assume that for each $i \in U$, an auxiliary vector $\mathbf{x}_i$ is observed. Let $t_x = \sum_{i \in U} \mathbf{x}_i$. A probability sample $s \subset U$ is drawn according to a fixed-size sampling design $p(\cdot)$, where $p(s) = \Pr[\text{sample } s \text{ is selected}]$. Let $\pi_i = \Pr[i \in s] = \sum_{s: i \in s} p(s) > 0$ and $\pi_{ij} = \Pr[i, j \in s]$ for all $i, j \in U$.

We first consider *descriptive inferences* for this finite population, often done in terms of a point estimate and an associated confidence interval for a finite population parameter such as a total $t_y = \sum_{i \in U} y_i$ or mean $\bar{y}_U = N^{-1} t_y$. A proportion is a special case of the mean, with y_i equal to an indicator on some event. In particular, the finite population distribution function, denoted $F_y(z) = N^{-1} \sum_{i \in U} I_{\{y_i \le z\}}$ with $I_{\{A\}} = 1$ if the event A is true, and 0 otherwise, is a proportion for each fixed z. Other interesting finite population parameters include ratios $\sum_{i \in U} y_i \big/ \sum_{i \in U} z_i$ and vectors of regression coefficients,

$$\boldsymbol{B} = \left(\sum_{i \in U} \mathbf{x}_i^T \mathbf{x}_i \right)^{-1} \sum_{i \in U} \mathbf{x}_i y_i.$$

Each of these examples is built up from finite population totals, and so a canonical problem of interest is estimation of the population total for a generic study variable y.

The Horvitz–Thompson estimator of t_y,

$$\widehat{t}_y = \sum_{i \in s} \frac{y_i}{\pi_i}, \tag{10}$$

(Horvitz and Thompson, 1952) provides an unbiased estimator for the population total t_y, with variance under the sampling design

$$\mathrm{Var}_p\left(\widehat{t}_y\right) = \sum_{i,j \in U} \left(\pi_{ij} - \pi_i \pi_j\right) \frac{y_i}{\pi_i} \frac{y_j}{\pi_j} \tag{11}$$

(see Chapter 1). If auxiliary variables are available for a survey, it might be possible to obtain estimators that are more efficient than $\widehat{t}_y$.

It is of interest to improve upon the efficiency of the Horvitz–Thompson estimator by using the auxiliary information $\mathbf{x}_i$. Motivation for such estimators is often provided by modeling the finite population of y_i as a realization from an infinite superpopulation, ξ, relating $\mathbf{x}_i$ to y_i via

$$y_i = \mu(\mathbf{x}_i) + \varepsilon_i, \tag{12}$$

where ε_i is an independent sequence of random variables with mean zero and variance $v(\mathbf{x}_i)$. Standard superpopulation models are parametric, and typically linear, that is $\mu(\mathbf{x}_i) = \mathbf{x}_i^T \boldsymbol{\beta}$. The potential disadvantage of estimators motivated by a superpopulation model is inefficiency under model misspecification. If the regression model does not fit the data well, there is no improvement over a simple Horvitz–Thompson estimator and potentially even a loss of efficiency. To avoid the consequences of model misspecification, it is natural to replace the parametric specification by a nonparametric specification, in which $\mu(\cdot)$ is a smooth function of $\mathbf{x}$ and $v(\cdot)$ is smooth and strictly positive.

Once the model (whether parametrically or nonparametrically specified) is fitted to the sample data, there are at least two ways to incorporate its predictions into estimation

of the finite population total. The first is a *model-based* approach, in which model-fitted values $\tilde{\mu}(\mathbf{x}_i)$ are used to predict only the nonsampled values of y:

$$\widehat{t}_{\mathrm{MB}} = \sum_{i \in U \setminus s} \tilde{\mu}(\mathbf{x}_i) + \sum_{i \in s} y_i. \tag{13}$$

Typically, model-based estimators of this type are asymptotically model unbiased and highly efficient when $\mu(\mathbf{x}_i)$ and $v(\mathbf{x}_i)$ are correctly specified but biased and even inconsistent if the model is wrong. Inspired by the general applicability of nonparametric models, Kuo (1988), Dorfman (1992), and Chambers et al. (1993) have developed model-based estimators using nonparametric regression.

The second way to incorporate model predictions is *model-assisted* and avoids the potential problems of model misspecification through a design bias adjustment. Model-assisted estimation relies on a model-fitted prediction $\widehat{\mu}_i$ for all the population elements but then corrects the possible design bias in that prediction. The resulting model-assisted regression estimator is of the form

$$\widehat{t}_{\mathrm{MA}} = \sum_{i \in U} \widehat{\mu}_i + \sum_{i \in s} \frac{y_i - \widehat{\mu}_i}{\pi_i}. \tag{14}$$

An intuitive explanation of the design properties of the model-assisted regression estimator proceeds as follows. Let μ_i represent the regression fit for $\mu(\mathbf{x}_i)$ if the entire population were observed. If these $\mu_i s$ were known, then an exactly design-unbiased estimator of t_y would be the generalized difference estimator

$$t_y^* = \sum_{i \in U} \mu_i + \sum_{i \in s} \frac{y_i - \mu_i}{\pi_i} \tag{15}$$

(see Särndal et al. (1992), p. 221, for the parametric case). The design variance of the estimator would be

$$\mathrm{Var}_p\left(t_y^*\right) = \sum_{i,j \in U} \left(\pi_{ij} - \pi_i \pi_j\right) \frac{y_i - \mu_i}{\pi_i} \frac{y_j - \mu_j}{\pi_j}, \tag{16}$$

which we would expect to be smaller than (11) because the $y_i s$ should be "close to" the μ_i's for any reasonable smoothing procedure under the superpopulation model.

In practice, the μ_i are not known, but they are well-defined "parameters" of the finite population that can be estimated by the $\widehat{\mu}_i s$ even if the superpopulation model (12) does not hold. As will be discussed further below, the resulting nonparametric model-assisted estimator can share many properties of linear model-assisted estimators familiar to survey statisticians, including design consistency.

As noted at the beginning of this section, the finite population distribution function $F_N(z) = N^{-1} \sum_{i \in U} I_{\{y_i \le z\}}$ for each z is a special case of a population mean. The nonparametric model-based and model-assisted methods discussed below can thus be used without further modification to improve the precision of estimators of the finite population distribution function. The advantage of doing so is that the same survey weights can be used for estimating $F_N(z)$ for any z as well as the population mean $\overline{y}_U$, for all the survey variables. This approach is discussed in Johnson et al. (2008). However, a number of special estimation methods have also been developed that take advantage of the special structure of $F_N(z)$. We refer to Dorfman (Chapter 36) for further information on this topic.

2.1. *Nonparametric survey regression estimation using kernels*

We now describe a number of ways in which nonparametric estimation can be implemented for descriptive inference. We begin by considering local polynomial regression (LPR) for scalar x_i, as in Section 1.2. Let $\mathbf{y}_s = [y_i]_{i\in s}$ be the vector of y_i's in the sample and define the local design matrix

$$\mathbf{X}_{si} = \left[1\, x_j - x_i \cdots (x_j - x_i)^q\right]_{j\in s}, \tag{17}$$

corresponding to the design matrix in (4) evaluated at $x = x_i$, and the diagonal weighting matrix

$$\mathbf{W}_{si} = \operatorname{diag}\left\{\frac{1}{h}K\left(\frac{x_j - x_i}{h}\right)\right\}_{j\in s}.$$

The unweighted LPR estimator of $\mu(x_i)$ is then given by the intercept in the local, weighted least squares fit of the polynomial:

$$\tilde{\mu}(x_i) = (1, 0, \ldots, 0)\left(\mathbf{X}_{si}^T\mathbf{W}_{si}\mathbf{X}_{si}\right)^{-1}\mathbf{X}_{si}^T\mathbf{W}_{si}\mathbf{y}_s. \tag{18}$$

Plugging these model fits into (13) then yields the model-based kernel regression estimator of Dorfman (1992).

One approach to producing a model-assisted estimator begins instead with the finite population local polynomial fit. Let $\mathbf{y}_U = [y_i]_{i\in U}$ be the vector of y_is for the entire finite population. Define the $N \times (q+1)$ matrix

$$\mathbf{X}_{Ui} = \left[1\, x_j - x_i \cdots (x_j - x_i)^q\right]_{j\in U}$$

and define the $N \times N$ matrix

$$\mathbf{W}_{Ui} = \operatorname{diag}\left\{\frac{1}{h}K\left(\frac{x_j - x_i}{h}\right)\right\}_{j\in U}.$$

The finite population local polynomial fit is then given by

$$\mu_i = \boldsymbol{e}_1^T\left(\mathbf{X}_{Ui}^T\mathbf{W}_{Ui}\mathbf{X}_{Ui}\right)^{-1}\mathbf{X}_{Ui}^T\mathbf{W}_{Ui}\mathbf{y}_U, \tag{19}$$

as long as $\mathbf{X}_{Ui}^T\mathbf{W}_{Ui}\mathbf{X}_{Ui}$ is invertible. The μ_i are the quantities that would be used in the difference estimation (15) if they were available. Since they are generally not available, they are estimated by design-weighted estimators $\widehat{\mu}_i$, constructed by letting

$$\mathbf{W}_{si\pi} = \operatorname{diag}\left\{\frac{1}{\pi_j h}K\left(\frac{x_j - x_i}{h}\right)\right\}_{j\in s}$$

and

$$\widehat{\mu}_i = \boldsymbol{e}_1^T\left(\mathbf{X}_{si}^T\mathbf{W}_{si\pi}\mathbf{X}_{si}\right)^{-1}\mathbf{X}_{si}^T\mathbf{W}_{si\pi}\mathbf{y}_s, \tag{20}$$

provided $\mathbf{X}_{si}^T\mathbf{W}_{si\pi}\mathbf{X}_{si}$ is invertible. Plugging these fits into (14) then yields the model-assisted LPR estimator of Breidt and Opsomer (2000).

Breidt and Opsomer (2000) discuss the theoretical design and model properties of the local polynomial estimator, showing that the LPR estimator is design consistent and asymptotically design unbiased under a mild set of regularity conditions that we

hereafter assume to hold. Asymptotically, the design mean squared error of $\widehat{t}_{\mathrm{MA}}$ under LPR is equivalent to the variance of the generalized difference estimator,

$$\mathrm{MSE}_p\big(\widehat{t}_{\mathrm{MA}}\big) = \mathrm{E}_p\big(\widehat{t}_{\mathrm{MA}} - t_y\big)^2 \approx \sum_{i,j\in U} (y_i - \mu_i)\big(y_j - \mu_j\big)\frac{\pi_{ij} - \pi_i\pi_j}{\pi_i\pi_j}, \quad (21)$$

recalling that the μ_i have been defined as (unknown) finite population parameters. A design consistent and asymptotically design unbiased estimator of $\mathrm{MSE}_p\left(\widehat{t}_{\mathrm{MA}}\right)$ is

$$\widehat{V}\big(\widehat{t}_{\mathrm{MA}}\big) = \sum_{i,j\in s} \big(y_i - \widehat{\mu}_i\big)\big(y_j - \widehat{\mu}_j\big)\frac{\pi_{ij} - \pi_i\pi_j}{\pi_i\pi_j}\frac{1}{\pi_{ij}}. \quad (22)$$

Because each of the smoothed values $\widehat{\mu}_i$ is a linear combination of the y_i in the sample, the LPR model-assisted estimator (14) can also be written in the same form, that is, $\widehat{t}_{\mathrm{MA}} = \sum_s \omega_{is} y_i$ with ω_{is} not involving the y_i. It is readily checked that the weights ω_{is} are calibrated for the population size as well as for the totals of powers of the x_i up to degree q : $\sum_s \omega_{is} x_i^p = \sum_U x_i^p$ for $0 \le p \le q$. The LPR model-assisted estimator shares this property with the generalized regression estimators.

In simulation experiments reported in Breidt and Opsomer (2000), the LPR estimator was competitive with the classical survey regression estimator when the population regression function was linear but dominated the regression estimator when the regression function was not linear. The estimator also performed well relative to other parametric and nonparametric estimators, both model-assisted and model-based. It generally dominated the Horvitz–Thompson estimator and it dominated cubic regression and poststratification estimators, provided it was not oversmoothed. It was sometimes much better and never much worse than two competing model-based nonparametric estimators. Though the efficiency of the nonparametric estimator depended on the choice of bandwidth parameter, the results were fairly insensitive to this choice, suggesting that large gains in efficiency can be attained for a variety of bandwidths.

The local polynomial method can be applied in virtually all situations where generalized regression estimation is used, as long as the value of auxiliary variable x_i is available for every element in the population and it is a continuous (not categorical) variable. Examples of the type of generalizations that are possible are provided by Deville and Goga (2004), who applied LPR to improve the efficiency of survey estimators when samples are taken on two occasions, and by Aragon et al. (2006), who considered quantile estimation.

2.2. Nonparametric survey regression estimation using splines

We next consider nonparametric survey regression estimation using splines, focusing on the special case of penalized regression splines with scalar x_i. For the superpopulation model (12), we assume now that the mean function $\mu(\cdot)$ can be written as in (5). We define $\mathbf{x}_i^T = \big(1, x_i, \ldots, x_i^q, (x_i - \kappa_1)_+^q, \ldots, (x_i - \kappa_J)_+^q\big)$, $\mathbf{X}_s = \left[\mathbf{x}_i^T\right]_{i\in s}$, $\mathbf{X}_U = \left[\mathbf{x}_i^T\right]_{i\in U}$, and $\Pi_s = \mathrm{diag}\{\pi_i\}_{i\in s}$. Further, define the diagonal matrix $\mathbf{A}_\lambda = \mathrm{diag}\{0, \ldots, 0, \lambda, \ldots, \lambda\}$, with $q+1$ zeros on the diagonal followed by J penalty constants λ, corresponding to the J truncated polynomial terms in (5).

The unweighted sample spline estimator of $\mu(x_i)$, corresponding to the solution to the penalized least squares minimization (6), is then

$$\tilde{\mu}(x_i) = \mathbf{x}_i^T \left(\mathbf{X}_s^T \mathbf{X}_s + \mathbf{A}_\lambda\right)^{-1} \mathbf{X}_s^T \mathbf{y}_s. \tag{23}$$

Using $x_i = \pi_i$ in (23) and plugging it into (13), Zheng and Little (2003) have proposed a model-based survey regression estimator that uses penalized splines to account for the effect of nonignorable design weights. They have further extended the penalized spline model-based survey regression estimator to the case of two-stage sampling in Zheng and Little (2004).

A model-assisted survey regression estimator based on penalized splines begins by first defining the population fit

$$\mu_i = \mathbf{x}_i^T \left(\mathbf{X}_U^T \mathbf{X}_U + \mathbf{A}_\lambda\right)^{-1} \mathbf{X}_U^T \mathbf{y}_U$$

and then estimates this finite population parameter with a sample-weighted version,

$$\widehat{\mu}_i = \mathbf{x}_i^T \left(\mathbf{X}_s^T \Pi_s^{-1} \mathbf{X}_s + \mathbf{A}_\lambda\right)^{-1} \mathbf{X}_s^T \Pi_s^{-1} \mathbf{y}_s.$$

Plugging these design-weighted fits into (14) yields the penalized spline model-assisted survey regression (PSP) estimator proposed by Breidt et al. (2005).

This estimator has theoretical properties similar to those of the LPR estimator discussed above, including calibration for population totals of powers of x_i up to degree q, design consistency, and asymptotic design unbiasedness (under mild conditions), and its design mean squared error can also be written as in (21). In simulation experiments reported in Breidt et al. (2005), it is shown that the PSP estimator is most often very similar to the LPR estimator. However, penalized spline regression offers a number of advantages over kernel-based methods that make it an attractive smoothing method in the model-assisted context. Incorporating multiple covariates as well as combinations of categorical variables, parametric and nonparametric terms, is straightforward, as shown in Aerts et al. (2002). Another important advantage is the relative ease with which PSP estimators can be computed, even for large data sets or data sets with regions of sparse data. Finally, an important practical consideration is that, since they are more closely related to parametric models, estimators based on spline models are easier to implement in existing survey estimation procedures.

Another class of nonparametric model-assisted estimators based on splines has been studied by Goga (2004, 2005). In both of these papers, Goga uses unpenalized regression splines for which the domain of the auxiliary variable is divided by a number of knots, a B-spline basis function is associated with each knot, and the number of knots goes to infinity so that the B-splines become dense on the domain. Goga (2005) shows that the regression spline estimator is asymptotically design-unbiased and consistent, proposes a design-based variance approximation, and shows that the anticipated variance is asymptotically equivalent to the Godambe–Joshi lower bound. Simulations show that the regression spline estimator has good properties. Goga (2004) applies this methodology to construct model-assisted estimators in the case of sampling on two occasions, with complete auxiliary information available on each occasion.

2.3. Other smoothing methods for survey regression estimation

While the estimators of Sections 2.1 and 2.2 can, in principle, be generalized directly to handle multivariate $\boldsymbol{x}_i$, the modeling approaches described in Section 1.4 are likely

to be more useful in practice. Breidt et al. (2007) extend the LPR estimator to the semiparametric model (8) and show that the semiparametric model-assisted estimator is design consistent and asymptotically normal. They also show that it is calibrated for the population totals of the auxiliary variables in both the parametric and nonparametric portions of the model.

Montanari and Ranalli (2005) proposed neural networks as a multivariate smoothing technique for model-assisted estimation. Opsomer et al. (2008) considered the generalized additive model (9) as a multivariate superpopulation model specification and fitted it by local scoring. One issue with both methods is that they do not lead to estimators calibrated to population totals of the auxiliary variables. In addition, the local scoring estimator for GAM is not a linear function of the y_i so that the resulting model-assisted estimator cannot be written as a weighted sum, making it difficult to integrate GAMs into the usual survey estimation context. Both Montanari and Ranalli (2005) and Opsomer et al. (2008) applied *model calibration*, originally proposed by Wu and Sitter (2001) as a way to obtain calibrated weighted forms for their estimators. Letting $\widehat{\mu}_i$ denote the fits obtained by either neural network fitting or local scoring, model calibration uses the same expression as the model-assisted estimator $\widehat{t}_{\mathrm{MA}}$ in (14) but replaces the $\widehat{\mu}_i$ by $\widehat{\mu}_i^* = \widehat{\mu}_i\widehat{\beta}$, with $\widehat{\beta}$ the estimated coefficient from regressing the y_i on the $\widehat{\mu}_i$ using design-weighted least squares regression. The resulting estimator is then calibrated for $\sum_U \widehat{\mu}_i$. In Opsomer et al. (2008), the idea of model calibration is further extended by combining the $\widehat{\mu}_i$ from the GAM with additional covariates into a multivariate linear model, with the resulting estimator calibrated for all the variables included in that linear model. This estimator can again be written as a weighted sum of the observations (ignoring the fact that the $\widehat{\mu}_i$ themselves depend on the y_i).

2.4. Smoothing parameter selection

Nonparametric regression applications require the specification of one or several smoothing parameters, such as the bandwidth in kernel regression or the penalty in spline regression. Selecting the "right" amount of smoothing is a challenging topic in the model-assisted context, further complicated by the fact that in a typical survey application, a single set of survey regression weights is applied to all the survey variables. Because the best smoothing parameter choice depends on the variable being smoothed, no single parameter value (and hence single set of survey weights) will be optimal for all variables in the survey. Nevertheless, it is of interest to have a method to select the amount of smoothing for those cases when precision for a single variable or a small set of them can justify the development of a specifically targeted estimation procedure.

Opsomer and Miller (2005) proposed an automated bandwidth selection method for the LPR estimator that estimates the bandwidth h minimizing the design mean squared error (21). They note that minimizing the traditional estimator of the design mean squared error, $\widehat{V}\left(\widehat{t}_{\mathrm{MA}}\right)$ in (22), tends to pick bandwidths that are much too small and instead propose a cross-validation-based estimator. The estimator is the minimizer of

$$\mathrm{CV}(h) = \sum_{i,j\in s}\left(y_i - \widehat{\mu}_i^{(-)}\right)\left(y_j - \widehat{\mu}_j^{(-)}\right)\frac{\pi_{ij}-\pi_i\pi_j}{\pi_i\pi_j}\frac{1}{\pi_{ij}},$$

where $\widehat{\mu}_i^{(-)}$ is the LPR fit computed as in (20) but with observation i removed from the sample. The criterion $\mathrm{CV}(h)$ is a complicated function of h that needs to be evaluated numerically to find its minimum. Computations can be greatly simplified because $\widehat{\mu}_i^{(-)}$ is easily written as a function of $\widehat{\mu}_i$ so that it is only necessary to fit the LPR once for each value of h. Simulations in Opsomer and Miller (2005) show that the minimizer of $\mathrm{CV}(h)$ is able to successfully adjust the amount of smoothing to the characteristics of the underlying function $\mu(\cdot)$, even for moderate sample sizes. Smoothing parameter selection for the other nonparametric model-assisted estimators described in this section has not been formally studied, but the principle of using a cross-validation-based criterion based on the design mean squared error should apply for the PSP estimator, as well as the more complicated GAM and neural network cases as well.

3. Nonparametric methods in analytic inference from surveys

In contrast to descriptive inferences about the current state of a real, finite population, analytic inferences are about model parameters for a hypothetical, infinite population considered to be a generating mechanism for the current state of the finite population. Analytic inference from surveys using nonparametric methods is relatively rare in the literature and not always cleanly divided from descriptive inference. Both nonparametric density estimation and regression estimation have been used in analytic inferences.

3.1. Nonparametric density estimation

Because no probability density function exists for a finite population, density estimation must be regarded as asymptotic descriptive inference about a limiting sequence of finite populations (i.e., the finite population distribution function $F_N(z)$ defined above is assumed to converge to a differentiable function $F(z)$ as $N \to \infty$) or as analytic inference about an infinite superpopulation. We focus here on the case of analytic inference, in which the goal is to estimate the hypothetical probability density function generating the realized y-values in the finite population.

Bellhouse and Stafford (1999) consider both asymptotic descriptive inference and analytic inference. Using a design-based approach, they compute design-weighted kernel smooths from sample data as estimates of the corresponding finite population smooths. They also consider a binned version of the problem, with the range of y divided into equally-sized bins, leading to a histogram estimate. Finally, they consider a smoothed version of the histogram. They informally develop asymptotic integrated mean square errors of the various estimators under model and design and illustrate with data from the Ontario Health Survey.

Breunig (2001) takes a purely model-based approach to density estimation in the context of survey data from a clustered design. This work accounts for the correlation structure induced by the clustering but ignores all other design features.

Buskirk and Lohr (2005) extend earlier work of Buskirk (1998, 1999) to a thorough exploration of design-weighted kernel density estimation in design-based, model-based, and combined settings. They develop asymptotic theory in these various settings, discuss bandwidth selection and density estimation near boundaries, and apply the methods to data from the U.S. National Crime Victimization Survey and from the U.S. National Health and Nutrition Examination Survey III.

3.2. Nonparametric regression estimation

One particular type of analytic inference for which nonparametric methods are well-suited is exploratory data analysis, in which nonparametric scatterplot smoothers are used to suggest the functional form of the regression relationship between y and $\mathbf{x}$. Korn and Graubard (1998) use LPR to smooth survey microdata, accounting for complex design. They do not describe the theoretical properties of this methodology. Bellhouse and Stafford (2001) use LPR in the same context of exploratory studies. Their goal is to make inference about the infinite population regression function $m(\cdot)$ from (2). They take a design-based approach to this inferential problem by constructing bins on the x-variable and by using the design weights to estimate bin proportions in the finite population and the y-means within bins. If x_i denotes the x-value characterizing the ith bin, with estimated bin proportion $\widehat{p}_i$ and bin mean $\widehat{\overline{y}}_i$, then the function $m(\cdot)$ is estimated using weighted LPR of $\widehat{\overline{y}}_i$ on x_i, with the usual kernel weights modified by multiplying by $\widehat{p}_i$. The authors approximate the design expectation and variance of the resulting estimator and illustrate with data from the Ontario Health Survey.

Smith and Njenga (1992) take a different approach to incorporating nonparametric regression into analytic inference. They begin by discussing robustness under both design and model-based inference and propose new methods for robust model-based analytic inference by using smoothing techniques. Specifically, they suggest kernel regression of multivariate $\mathbf{y}$ study vectors on covariates $\mathbf{x}$ to estimate conditional mean vectors, followed by kernel regression of the resulting multivariate residuals on $\mathbf{x}$ to estimate conditional covariance matrices. These nonparametric estimates are robust to model misspecification and can be used for analytic inferences about regression coefficients or for multivariate analyses about the relationships among components of the $\mathbf{y}$ vector.

Yet another approach to employing nonparametric regression in analytic inference is proposed by Chambers et al. (2003). This approach builds on the corresponding parametric approach developed in Pfeffermann and Sverchkov (1999). The idea of that article is to avoid the potential bias caused by nonignorable sampling designs by using the sample distribution of the sample measurements in maximum likelihood estimation. This sample distribution is related to the conditional distribution of the study variable and the conditional distribution and conditional mean of the sample selection probabilities. These quantities are estimated parametrically in Pfeffermann and Sverchkov (1999) and nonparametrically in Chambers et al. (2003) under various data scenarios. The simplest of these scenarios, for example, leads to an estimator that uses a design-weighted version of the Nadaraya–Watson estimator.

The demand for flexible, robust procedures in all aspects of inference from complex surveys suggests that nonparametric methods have great potential in this area. The three approaches described in this section have tapped some of that potential, but it is clear that much further work remains to be done, and this should be a fruitful area of future research.

4. Nonparametric methods in nonresponse adjustment

In Section 2, nonparametric methods were used to improve the efficiency of survey estimators by taking advantage of the relationship between auxiliary variables available for

the population and the survey variables. In this section, we describe how nonparametric methods can also be used to adjust survey estimators for the presence of nonresponse, when the response mechanism is related to an auxiliary variable available for the original sample. We focus here on the case of *unit nonresponse.*

Nonresponse is pervasive in surveys and can induce bias if it is not properly accounted for. In the context of unit nonresponse, the most commonly used approach is to adjust the weights of the responding observations by incorporating estimates of the probabilities that the units are respondents. This can be done implicitly, as in the *weighting cell* estimator, or explicitly by specifying and fitting a response probability function and obtaining new weights. In both cases, the response process can be viewed here as a second *phase* of sampling, with unknown probability mechanism. This nonresponse phase follows the first phase of sampling, which is determined by the original sampling design. Särndal and Swensson (1987) formally describe the two-phase framework for nonresponse and the types of approaches that can be used to adjust survey estimators for nonresponse.

Suppose that we have a sampling design $p(\cdot)$ with corresponding inclusion probabilities π_i, $i \in U$, and that, in the absence of nonresponse, we were planning to estimate the population total t_y by the Horvitz–Thompon estimator in (10) based on the sample s. Because of nonresponse, we only observe $r \subseteq s$. Since the random process generating r is typically unknown, we need to assume a model for the response mechanism. Let $R_i = 1$ if $i \in r$ and 0 otherwise. As is often done in the nonresponse modeling context, we will assume that the R_i are independent Bernoulli random variables with

$$\Pr\{R_i = 1\} = \phi_i, 0 < \phi_i \leq 1, \forall i \in U \tag{24}$$

(see also Chapter 8 or 25).

For the weighting cell estimator, the population is divided into G cells, $U = \cup_{g=1}^{G} U_g$, and in each cell, the (average) response probability is estimated by the fraction of the sampled respondents in cell g. This fraction is usually computed as $\sum_{r_g} w_i / \sum_{s_g} w_i$, where $s_g = s \cap U_g$; $r_g = r \cap U_g$, with either $w_i = 1/\pi_i$ or $w_i = 1$. In what follows, we will only consider the former case. The weighting cell estimator for t_y is defined as

$$\widehat{t}_{\text{wc}} = \sum_{g=1}^{G} \left(\frac{\sum_{s_g} w_i}{\sum_{r_g} w_i} \right) \sum_{r_g} w_i y_i. \tag{25}$$

From this expression, it is easy to see that in each cell, the estimator of the cell total is ratio-adjusted by the inverse of the weighted proportion of respondents in the cell.

A number of authors have studied the properties of the weighting cell estimator, including Oh and Scheuren (1983), Särndal et al. (1992, p.578) (using the term "response homogeneity group" for the cells), and Kim and Fuller (1999). A common assumption in the study of the design-based properties of $\widehat{t}_{\text{wc}}$ is that the cells are correctly specified, in the sense that they correspond to well-defined and known population groups in which the response indicators R_i are independent and identically distributed with $\Pr\{R_i = 1\} = \phi_g$, $i \in U_g$. Although these authors showed that $\widehat{t}_{\text{wc}}$ is consistent under this assumption, it was not clear what happens when the cells are not correctly specified.

Da Silva and Opsomer (2004) investigate the theoretical behavior of $\widehat{t}_{\text{wc}}$ using nonparametric methodology. Instead of assuming that the cells correspond to known

response categories in the population, they consider the situation in which the response probability $\phi_i = \phi(x_i)$, with x_i an auxiliary variable observed for all elements in the sample and $\phi(\cdot)$ an unknown smooth function. The weighting cells are formed by sorting the sample on the x_i and dividing the range of x_i into G groups. Under this scenario, the grouping can be thought of as a very simple form of smoothing, with the unknown function $\phi(\cdot)$ approximated by a piecewise constant fit, and G a "smoothing parameter" similar to the bandwidth h or the penalty λ in the previous sections. Da Silva and Opsomer (2004) prove that $\widehat{t}_{\text{wc}}$ is a consistent estimator of t_y under quasi-randomization (i.e., the combination of the sampling design and the response mechanism) under mild conditions, provided that G is allowed to increase as the sample size increases. Unlike the previous authors studying the weighting cell estimator, they do not require the cells to be correctly specified.

While the weighting cell estimator can be considered a simple nonparametric estimator, it is possible to construct nonresponse adjusted estimators that incorporate nonparametric regression methods more fully. This type of estimator will be based on explicit estimation of the unknown response probability function $\phi(\cdot)$ and the ideas of two-phase estimation. Starting again from the Horvitz–Thompson estimator in (10), suppose the response probability function $\phi(\cdot)$ were known. The two-phase estimator

$$\widehat{t}_{\phi} = \sum_{r} \frac{y_i}{\pi_i \phi\,(x_i)} \tag{26}$$

is unbiased and consistent under quasi-randomization. This estimator is infeasible so that it is replaced by

$$\widehat{t}_{\widehat{\phi}} = \sum_{r} \frac{y_i}{\pi_i \widehat{\phi}_i}, \tag{27}$$

with $\widehat{\phi}_i$ an estimate of $\phi(x_i)$.

Many authors have considered parametric specifications for $\phi(\cdot)$, including Kim and Kim (2007). We discuss the nonparametric case here. The use of kernel-type smoothing methods in the nonresponse context was first proposed by Giommi (1984, 1987) and further discussed by Niyonsenga (1994, 1997). Neither of these authors provided formal theoretical results on their nonparametric estimators. Recently, Da Silva and Opsomer (2006) studied the properties of the estimator in (27) with the response probability function $\phi(\cdot)$ estimated by a sample-weighted kernel regression estimator of the response indicators. The estimator is a special case of the estimator in (20), with $\boldsymbol{y}_s$ replaced by the vector of response indicators R_i in the sample and the degree of the local polynomial $q = 0$, that is, the local design matrix in (17) replaced by a vector of ones. The resulting estimator can be written as

$$\widehat{\phi}_i = \left(\sum_{j \in s} K\!\left(\frac{x_j - x_i}{h} \right) \frac{1}{\pi_j} \right)^{-1} \sum_{j \in s} K\!\left(\frac{x_j - x_i}{h} \right) \frac{R_j}{\pi_j}. \tag{28}$$

The results of Da Silva and Opsomer (2006) for the estimator (27) with $\phi(x_i)$ estimated by (28) show that the nonparametric nonresponse adjusted estimator is quasi-randomization consistent for t_y under mild conditions. They also found that $\widehat{t}_{\widehat{\phi}}$ does not

have the same asymptotic distribution as $\widehat{t}_{\phi}$, but that the estimation of the response probability function $\phi(\cdot)$ contributes additional terms in the asymptotic approximation. This implies that if the estimated response function is treated as if it were known for the purpose of inference, it is likely that the variance will be incorrectly estimated. Kim and Kim (2007) found a similar result in the parametric case.

5. Nonparametric methods in small area estimation

As a final application of nonparametric methods in the survey context, we discuss applications in small-area estimation. Cowling et al. (1996) present two applications of spatial smoothing in a small-area estimation context. The first use of smoothing is in making small-area estimates less variable. The procedure adjusts the original weights to allow for deviations from benchmark totals, then the modified weights are spatially smoothed via a kernel over geographic neighborhoods to get less spatial variability in the weights. The result is more stable small-area estimates. The second application uses design-weighted kernel smooths to get maps of estimates of the characteristic over a spatial domain. No properties are derived for either of these methodologies.

In two recent developments, nonparametric methods are brought directly into classical methods for small-area estimation. Mukhopadhay and Maiti (2004) propose an extension of the area-level model in which the linear mean function is replaced by a nonparametric function to be estimated by kernel regression, while Opsomer et al. (2008) consider an element-level model and use penalized spline regression.

Suppose the population contains T small areas of interest, indexed by t. The nonparametric area-level model studied by Mukhopadhay and Maiti (2004) is

$$y_t = m(x_t) + u_t + \varepsilon_t, \tag{29}$$

where u_t and ε_t are distributed independently as $\mathcal{N}\left(0, \sigma_u^2\right)$ and $\mathcal{N}(0, D_t)$ with D_t known. If $m(\cdot)$ is a linear function, this model is usually called the Fay–Herriot model (Fay and Herriot, 1979). The purpose of small-area estimation methods for model (29) is to predict $\tilde{y}_t = m(x_t) + u_t$, and in the linear model case, empirical best linear prediction (EBLUP) methods or hierarchical Bayesian methods are typically used. The prediction procedure starts by estimating $m(\cdot)$ by the LPR estimator (3) with $q = 0$ so that the matrix $\boldsymbol{X}_x$ in (4) is replaced by a vector of ones (as was done in Section 4). The small-area variance σ_u^2 is estimated by $\widehat{\sigma}_u^2 = \sum_{t=1}^{T} \left\{(y_t - \widehat{m}(x_t))^2 - D_t\right\}/T$, possibly adjusted to ensure non-negativity, and the predictor for $\tilde{y}_t$ is defined in analogy to the EBLUP as

$$\widehat{y}_t = \widehat{\gamma}_t y_t + (1 - \widehat{\gamma}_t)\, \widehat{m}(x_t) \tag{30}$$

with $\widehat{\gamma}_t = \widehat{\sigma}_u^2 / \left(\widehat{\sigma}_u^2 + D_t\right)$. Mukhopadhay and Maiti (2004) derive an asymptotic approximation to the prediction mean squared error of $\widehat{y}_t$, $\mathrm{E}\left(\widehat{y}_t - \tilde{y}_t\right)^2$, and a plug-in estimator for that quantity.

Because of close connections between EBLUP and penalized spline regression (see Wand, 2003), penalized splines provide a convenient approach for integrating nonparametric models into small-area estimation. Opsomer et al. (2008) extend the linear element-level mixed model small-area estimation approach described in Battese et al. (1988) to the setting in which the mean function can be nonparametrically (or

semiparametrically) specified. The model is

$$y_i = m(x_i) + \boldsymbol{d}_i^T \boldsymbol{u} + \varepsilon_i, \tag{31}$$

where $\boldsymbol{d}_i = (d_{1i}, \ldots, d_{Ti})^T$ is a vector of indicators with $d_{ti} = 1$ if element i is in the small-area t and zero otherwise, $\boldsymbol{u} = (u_1, \ldots, u_T)^T$ is a vector of mutually independent small-area effects with mean 0 and variance σ_u^2, and ε_i is the random error with mean 0 and variance σ_ε^2, independent of $\boldsymbol{u}$. The nonparametric function $m(\cdot)$ is expressed as a spline function as in (5). Following Wand (2003), we rewrite this as $m(x_i) = \boldsymbol{x}_i^T \boldsymbol{\beta} + \boldsymbol{z}_i^T \gamma$, with $\boldsymbol{x}_i = \left(1, x_i, \ldots, x_i^p\right)^T$, $\boldsymbol{z}_i = \left((x_i - \kappa_1)_+^p, \ldots, (x_i - \kappa_J)_+^p\right)^T$, $\boldsymbol{\beta}$ a vector of unknown parameters, and $\boldsymbol{\gamma}$ a vector of independent random variables with mean 0 and variance σ_γ^2. The full model is therefore

$$y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + \boldsymbol{z}_i^T \boldsymbol{\gamma} + \boldsymbol{d}_i^T \boldsymbol{u} + \varepsilon_i. \tag{32}$$

The term $\boldsymbol{z}_i^T \boldsymbol{\gamma}$ is a random deviation from the fixed linear trend in the population, and $\boldsymbol{d}_i^T \boldsymbol{u}$ is the random effect for small-area i. The goal of the small-area estimation is now the prediction of $\tilde{y}_t = \overline{\boldsymbol{x}}_t^T \boldsymbol{\beta} + \overline{\boldsymbol{z}}_t^T \boldsymbol{\gamma} + u_t$, where we assume that $\overline{\boldsymbol{x}}_t$ and $\overline{\boldsymbol{z}}_t$ are known.

The critical point of the formulation (32) is that we have once again expressed the model as a linear element-level mixed effect model so that the full range of EBLUP methods can be applied. Opsomer et al. (2008) propose restricted maximum likelihood estimation to estimate the parameters $\boldsymbol{\beta}$, σ_γ^2, σ_u^2, σ_ε^2, and predict $\tilde{y}_t$ by

$$\widehat{\tilde{y}}_t = \overline{\boldsymbol{x}}_t^T \widehat{\boldsymbol{\beta}} + \overline{\boldsymbol{z}}_t^T \widehat{\boldsymbol{\gamma}} + \widehat{u}_t, \tag{33}$$

with

$$\widehat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^T \widehat{\boldsymbol{V}}^{-1} \boldsymbol{X}\right)^{-1} \boldsymbol{X}^T \widehat{\boldsymbol{V}}^{-1} \boldsymbol{Y}$$
$$\widehat{\boldsymbol{\gamma}} = \widehat{\sigma}_\gamma^2 \boldsymbol{Z}^T \widehat{\boldsymbol{V}}^{-1} \left(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\right)$$
$$\widehat{\boldsymbol{u}} = \widehat{\sigma}_\gamma^2 \boldsymbol{D}^T \widehat{\boldsymbol{V}}^{-1} \left(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\right),$$

where $\boldsymbol{X} = (\boldsymbol{x}_1 \ldots, \boldsymbol{x}_n)^T$ and $\boldsymbol{Y}$, $\boldsymbol{Z}$, and $\boldsymbol{D}$ are defined analogously. Further, $\widehat{\boldsymbol{V}}$ is the estimated variance–covariance matrix of $\boldsymbol{Y}$ obtained by plugging the REML estimates of the variance parameters in the variance–covariance matrix of $\boldsymbol{Y}$. The asymptotic approximation to the prediction mean squared error of $\widehat{\tilde{y}}_t$ is shown to directly generalize that obtained in the absence of the spline random effect, and a bias-corrected estimator for the prediction mean squared error is provided. Opsomer et al. (2008) also discuss likelihood ratio testing for the variances of the random effects and propose a simple nonparametric bootstrap for inference.

Sample Surveys: Inference and Analysis, Vol. 29B
ISSN: 0169-7161
© 2009 Elsevier B.V. All rights reserved
DOI: 10.1016/S0169-7161(09)00228-4

Resampling Methods in Surveys

Julie Gershunskaya, Jiming Jiang and P. Lahiri

1. Introduction

Much of the sample survey theory has been developed since the 1930s. Neyman (1934) laid the foundation of the randomization or design-based approach for inference in finite population sampling. Neyman's article prompted the publication of a number of influential papers on design-based methods in the late 30s and 40s. See Rao (2005a) for a detailed history on the development of sample survey theory and applications. Under the randomization approach, the probability sampling mechanism that is used to draw a sample from the finite population forms the basis of inference, and the values of the characteristic of interest for the finite population are viewed as nonstochastic.

The model-based and model-assisted approaches for inference from a finite population have more recent origins, although the use of a superpopulation in finite population sampling can be traced back to Cochran (1939). Under the model-based prediction approach, the finite population is assumed to be a realization from a hypothetical superpopulation characterized by a probability distribution, a model serving all inferential purposes. Details of the prediction approach to finite population sampling can be found in Bolfarine and Zacks (1992), Valliant et al. (2000), and Chapter 23 of this volume. The Bayesian approach for inferences from a finite population also assumes a superpopulation model. It differs from the prediction approach in that a prior distribution on the parameters of the superpopulation model is assumed and the posterior distribution – the conditional distribution of the finite population parameter of interest given the sample – is used for all inferences. We refer the readers to Ghosh and Meeden (1997) and Chapter 29 for details on the Bayesian approach to finite population sampling.

The uncertainty associated with a survey estimate is commonly expressed in terms of its standard error estimate or a measure related to the standard error estimates, such as estimated coefficient of variation or a confidence interval. For a linear survey estimator, the estimation of its design-based standard error for a simple probability sample design is straightforward and involves obtaining an exact expression of the true design-based variance and then estimating the true design-based variance by a design-based unbiased estimator. However, the application of such direct approach is quite complicated for a complex sample design, which typically involves multiple stages of sampling and stratification at various stages.

In the context of jute acreage surveys conducted in Bengal, the Indian scientist P. C. Mahalanobis developed the method of interpenetrating subsamples to estimate the total variance of a survey estimator in presence of measurement errors (Mahalanobis, 1946). The method involves a random assignment of half-samples, drawn directly from the population, to two independent groups of interviewers. The interpenetrating subsampling technique is probably the first attempt to simplify variance estimation in complex surveys.

Replicated samples (Deming, 1956) and ultimate cluster technique in multistage sample design (Hansen et al., 1953) exploit ideas similar to that of Mahalanobis' interpenetrating subsamples method. The methods, commonly referred to as random group methods, involve either drawing two or more subsamples from the finite population or splitting the original sample into several random subgroups, constructing separate estimate of the parameter of interest from each subsample and an estimate from the pooled sample, and computing the variance among the several estimates. Thus, the random group methods offer several subsamples from which to draw inferences, a minor hurdle being that the estimates from the replicated subsamples are not independent unless subsamples are being replaced into the population before each draw. The appeal of the simple variance estimation for any estimator using several smaller subsamples is diminished because of the loss of efficiency. If the number of random groups is small, then variance estimation becomes unstable. On the other hand, with an increased number of random groups, the replicated subsamples become small and yield less efficient estimates than the original (or pooled) sample.

One possible remedy for the practical difficulty associated with the random group methods is to consider resampling methods, which are similar to the random group methods in terms of constructing variance estimates from the variation of the estimates for the subsamples, but differ in that they use subsamples that overlap. There are different resampling methods considered in the literature. The major difference among them is the way the subsamples are formed. The challenge of a resampling technique in surveys, for the most part, lies in its ability to resample from the original sample in such a way as to account for the original sample design.

The design-based variance estimation becomes more complex as one considers nonlinear estimators such as ratio, correlation coefficient, etc. for complex sample designs. There are several approaches for estimating the design-based variances of nonlinear estimators. The Taylor series method (also called the delta method) is one of them. Essentially, the method obtains an estimator of the design-based variance of a linear approximation to the nonlinear estimator by one of the methods available for estimating variances of linear estimators, including resampling methods. The delta method is widely used for simple nonlinear smooth estimators and is available in many software packages. However, this approach cannot be used for nonsmooth statistics.

A practical inconvenience of the Taylor linearization approach is that the form of variance estimator changes with the change of the nonlinear estimator. In contrast, resampling methods are available for obtaining a consistent estimator of the design-based variance for any estimator. The resampling methods are versatile in the sense that the same method can be applied for both smooth and nonsmooth statistics and for any complex survey design. In addition, the method remains the same in computation of the prediction variance of any estimator under the prediction approach to the finite population sampling.

One important application of resampling methods is in the small area estimation to estimate the mean squared errors of empirical best predictors (EBP) under nonlinear mixed models, where the Taylor series method is hard to apply, one reason being the complex derivative computations.

In Section 2, we use a simple setting to introduce two of the most common resampling plans used in survey sampling – the jackknife and the bootstrap. In Section 3, we describe modifications of the techniques to adjust for more complex situations of a stratified multistage design and more complex estimators. Section 4 considers variance estimation in presence of imputation for missing values. In Section 5, we review methods for the two-phase sample design. In Section 6, we consider resampling methods to compute the prediction variance of nonlinear predictors under prediction approach to finite population sampling. Section 7 contains a discussion on the application of the resampling techniques in the small area estimation.

2. The basic notions of bootstrap and jackknife

Consider a finite population of N units labeled by $U = \{1, \ldots, N\}$. Let y_i denote the value of a study variable for the ith unit of U $(i = 1, \ldots, N)$. Suppose we are interested in estimating the population total $Y = \sum_{i \in U} y_i$. Let $s = \{i_1, \ldots, i_n\}$ denote a sample of size n drawn from the finite population U. In this section, using a simple random sampling (SRS) setting, we explain the basic concepts of the bootstrap and jackknife and explore how these methods relate to the traditional direct analytic method.

Under the simple random sampling, an unbiased estimator of Y is given by $\hat{Y} = (N/n) \sum_{i \in s} y_i = \sum_{i \in s} w_i y_i$, where $w_i = N/n$, commonly referred to as the sampling weight, is the number of population units represented by the ith sampled unit. Note that $\hat{Y}$ is the well-known Hansen–Hurwitz estimator (Hansen and Hurwitz, 1943) for SRS with replacement (SRSWR), or unrestricted random sample design, and Narain-Horvitz-Thompson estimator (Horvitz and Thompson, 1952; Narain, 1951) for SRS without replacement (SRSWOR) design. We can write $\hat{Y} = \sum_{i \in U} w_i k_i y_i$, where k_i is the number of times the population unit i appears in the sample.

The variance of $\hat{Y}$ under a SRSWR design has an explicit simple form

$$\mathrm{Var}_{\mathrm{WR}}\big(\hat{Y}\big) = \frac{N\,(N-1)}{n} S^2, \tag{1}$$

where $S^2 = (N-1)^{-1} \sum_{i \in U} (y_i - \overline{Y})^2$, the finite population variance, and $\overline{Y} = N^{-1} \sum_{i \in U} y_i$, the finite population mean.

Note the following relationship between the variances under SRSWOR and SRSWR designs:

$$\mathrm{Var}_{\mathrm{WOR}}\big(\hat{Y}\big) = \frac{N-n}{N-1} \mathrm{Var}_{\mathrm{WR}}\big(\hat{Y}\big) \tag{2}$$

(see, e.g., Cochran, 1977).

Let $\overline{y} = n^{-1} \sum_{i \in s} y_i$ and $s^2 = (n-1)^{-1} \sum_{i \in s} (y_i - \overline{y})^2$ be the sample mean and variance, respectively. Note that under the SRSWR sampling, the expectation of s^2 is

$(N-1)N^{-1}S^2$. Hence, an unbiased estimator of $\mathrm{Var}_{\mathrm{WR}}(\hat{Y})$ is given by

$$\mathrm{var}_{\mathrm{wr}}\big(\hat{Y}\big) = \frac{N^2}{n}s^2 \tag{3}$$

The aim of the following subsections is to investigate the extent to which one can reproduce the variance estimator $\mathrm{var}_{\mathrm{wr}}$ using resampling procedures.

2.1. The bootstrap

The bootstrap, as introduced by Efron (1979), is perhaps the most natural and simple resampling method. The algorithm consists of SRSWR from the original sample, the resample size being usually equal to that of the original sample. The sample obtained at one iteration of the algorithm is called the bootstrap sample, and is denoted $s^{(b)}$. Let $k_i^{(b)}$ be the number of times a unit i from the original sample appears in the *bootstrap sample* $s^{(b)}$. Define the *bootstrap weights* as $w_i^{(b)} = w_i k_i^{(b)}$, for $i \in s$. Then the bootstrap estimate of the total is given by

$$\hat{Y}^{(b)} = \sum_{i\in s} w_i^{(b)} y_i. \tag{4}$$

Consider the distribution of $\hat{Y}^{(b)}$ over the n^n possible bootstrap samples. The formula for the *exact* bootstrap variance Var_* with respect to this distribution is analogous to (1):

$$\mathrm{Var}_*\big\{\hat{Y}^{(b)}\big\} = \frac{n(n-1)}{n}\frac{1}{n-1}\sum_{i\in s}\left(\frac{N}{n}y_i - \frac{N}{n}\overline{y}\right)^2 = \frac{(n-1)}{n}\mathrm{var}_{\mathrm{wr}}\Big(\hat{Y}\Big). \tag{5}$$

Hence, the bootstrap underestimates the true variance by the factor $(n-1)/n$.

In practice, resampling procedure stops after B Monte Carlo iterations, and the bootstrap variance estimate is obtained as

$$v_B\big(\hat{Y}\big) = \frac{1}{B}\sum_{b=1}^{B}\left\{\hat{Y}^{(b)} - \hat{Y}\right\}^2. \tag{6}$$

In (6) the replicate average $\overline{\hat{Y}}^{(b)} = B^{-1}\sum_{b=1}^{B}\hat{Y}^{(b)}$ is often used in place of $\hat{Y}$; this alternative variance formula, however, provides a variance estimate that is less conservative than $v_B\big(\hat{Y}\big)$.

Two alternatives to the aforementioned bootstrapping strategies are now presented.

Plan A. Subsamples of size $m \neq n$ are selected *with replacement* from the original sample.

The bootstrap weights are adjusted to $w_{i,m}^{(b)} = \lambda_i^{(b)} w_i k_i^{(b)}$, where $\lambda_i^{(b)} = n/m$, is an adjustment factor needed to maintain the unbiasedness of the replicate estimates. The bootstrap replicate estimate is given by $\hat{Y}_{m,\mathrm{wr}}^{(b)} = \sum_{i\in s} w_{i,m}^{(b)} y_i$ (subscripts in $\hat{Y}_{m,\mathrm{wr}}^{(b)}$ index the size of subsample and the subsampling mechanism). Similar to (5), the exact variance of $\hat{Y}_{m,\mathrm{wr}}^{(b)}$ over all possible such replicated subsamples is given by

$$\mathrm{Var}_*\left\{\hat{Y}_{m,\mathrm{wr}}^{(b)}\right\} = \frac{n}{m}\mathrm{Var}_*\left\{\hat{Y}_{n,\mathrm{wr}}^{(b)}\right\},$$

and

$$\mathrm{var}_{\mathrm{wr}}\left(\hat{Y}\right) = \frac{m}{n-1}\mathrm{Var}_*\left\{\hat{Y}^{(b)}_{m,\mathrm{wr}}\right\}. \tag{7}$$

Thus, with-replacement resampling with $m = n-1$ yields the unbiased estimate of variance.

Plan B. Subsamples of size $m < n$ are drawn *without replacement*.

The bootstrap variances of replicate estimates obtained under the without-replacement bootstrap scheme relate to the similar estimates obtained from the with-replacement sampling as follows:

$$\mathrm{Var}_*\left\{\hat{Y}^{(b)}_{m,\mathrm{wor}}\right\} = \frac{n-m}{n-1}\mathrm{Var}_*\left\{\hat{Y}^{(b)}_{m,\mathrm{wr}}\right\}.$$

This is analogous to (2), after replacing the population and sample counts by n and m, respectively. Therefore,

$$\mathrm{var}_{\mathrm{wr}}\left(\hat{Y}\right) = \frac{m}{n-m}\mathrm{Var}_*\left\{\hat{Y}^{(b)}_{m,\mathrm{wor}}\right\}. \tag{8}$$

It is interesting to note that the delete-one and delete-d jackknife schemes considered later in this chapter correspond to $m = n-1$ and $m = n-d$, $d > 1$ of the bootstrap Plan B, respectively.

Remark. In the bootstrap reweighting scheme described in both Plan A and Plan B, for the bootstrap sample b, the original weight is multiplied by an adjustment factor $\lambda_i^{(b)} = n/m$ if the original sampled unit is selected in the bootstrap sample and zero otherwise. In other words, the omitted original sampled units are removed at the time of estimation from the bootstrap sample. It is, however, possible to develop an alternative bootstrap weighting scheme that uses all the original sampled units. It can be shown that the following adjustment factors meet the desirable properties that (1) the resulting bootstrap estimator $\hat{Y}^{(b)}_m$ is an unbiased estimate of Y and (2) the bootstrap variance matches $\mathrm{var}_{\mathrm{wr}}(\hat{Y})$:

(i) If the bootstrap samples are selected with replacement,

$$\lambda^{(b)}_{i,\mathrm{wr}} = \begin{cases} 1 - \left(\frac{m}{n-1}\right)^{1/2} + \left(\frac{m}{n-1}\right)^{1/2}\frac{n}{m}, & \text{if } i \in s^{(b)}, \\ 1 - \left(\frac{m}{n-1}\right)^{1/2}, & \text{if } i \notin s^{(b)}; \end{cases} \tag{9}$$

(ii) If the bootstrap samples are selected without replacement,

$$\lambda^{(b)}_{i,\mathrm{wor}} = \begin{cases} 1 + \left(\frac{n-m}{m}\right)^{1/2}, & \text{if } i \in s^{(b)}, \\ 1 - \left(\frac{m}{n-m}\right)^{1/2}, & \text{if } i \notin s^{(b)}. \end{cases} \tag{10}$$

2.2. *The jackknife*

The original idea of the method can be attributed to Quenouille (1949) who used this method to reduce the bias of an estimator of the serial coefficient. In a follow-up article, Quenouille (1956) examined the properties of this bias reduction method in the context of infinite population. Tukey (1958) noted that the same approach could be used for

the variance estimation and gave it the name "jackknife." Durbin (1959) was the first to consider the application of jackknife in the design-based approach to the finite population sampling.

The delete-one-unit jackknife samples are constructed from the original sample by omitting one unit at a time. Let $s^{(j)}$ denote a jackknife sample obtained by deleting the jth original sampled unit. Because the jackknife sample size is $n-1$, smaller than the original sample size, the sampling weights need to be adjusted.

The jackknife weights are defined as follows:

$$w_i^{(j)} = \begin{cases} w_i n(n-1)^{-1}, & \text{if } i \in s^{(j)} \\ 0, & \text{otherwise.} \end{cases}$$

Define the jackknife replicate estimates of Y as $\hat{Y}^{(j)} = \sum_{i \in s} w_i^{(j)} y_i$ $(j = 1, \ldots, n)$. Then, the jackknife estimator of the design-based variance of $\hat{Y}$ is given by

$$v_{\mathrm{JK}}(\hat{Y}) = \frac{n-1}{n} \sum_{j=1}^{n} \left\{ \hat{Y}^{(j)} - \hat{Y} \right\}^2. \tag{11}$$

Note that for the simple linear estimator, $\hat{Y}$, we get identical jackknife variance estimator if $\hat{Y}$, in (11), is replaced by $n^{-1} \sum_{j=1}^{n} \hat{Y}^{(j)}$ because $\hat{Y} = n^{-1} \sum_{j=1}^{n} \hat{Y}^{(j)}$. Unlike the bootstrap estimate, the jackknife formula (11) is not a result of Monte Carlo simulations, but is an exact variance of the jackknife replicate estimates over all n possible subsamples. It follows directly from (8) and (11) that $\mathrm{var}_{\mathrm{wr}}(\hat{Y}) = v_{\mathrm{JK}}(\hat{Y})$.

For variance estimation for a general estimator $\hat{\theta}$ in the infinite population context, Tukey (1958) introduced the concept of pseudo values. Here, the pseudo values are defined as:

$$\tilde{\theta}^{(j)} = n\hat{\theta} - (n-1)\hat{\theta}^{(j)}, \quad j = 1, \ldots, n,$$

where $\hat{\theta}^{(j)}$ is similar to the definition of $\hat{Y}^{(j)}$. Treating $\tilde{\theta}^{(j)} (j = 1, \ldots, n)$ as iid, Tukey proposed the following estimator of the variance of $\hat{\theta}$:

$$\begin{aligned} v_{\mathrm{JK}}(\hat{\theta}) &= \frac{1}{n(n-1)} \sum_{j=1}^{n} (\tilde{\theta}^{(j)} - \tilde{\theta}^{(\cdot)})^2 \\ &= \frac{n-1}{n} \sum_{j=1}^{n} (\hat{\theta}^{(j)} - \hat{\theta}^{(\cdot)})^2, \end{aligned}$$

where $\tilde{\theta}^{(\cdot)} = n^{-1} \sum_{j=1}^{n} \tilde{\theta}^{(j)}$ and $\hat{\theta}^{(\cdot)} = n^{-1} \sum_{j=1}^{n} \hat{\theta}^{(j)}$. A replacement of $\hat{\theta}^{(\cdot)}$ in $v_{\mathrm{JK}}(\hat{\theta})$ by $\hat{\theta}$ results in an alternative variance estimator that is more conservative than $v_{\mathrm{JK}}(\hat{\theta})$. Note that for the linear estimator $\hat{Y}$, $v_{\mathrm{JK}}(\hat{Y})$ is obtained as a special case of Tukey's formula.

3. Methods for more complex survey designs and estimators

3.1. Variance estimation in stratified multistage sampling

Consider a finite population in which elements are grouped into primary sampling units (PSU) and the PSUs are grouped into H mutually exclusive and exhaustive strata with N_h PSUs in stratum $h (h = 1, \ldots, H)$. At the first stage, a sample s_h of n_h PSUs is

selected from stratum h with varying selection probabilities and without replacement. At the subsequent stages, elements or clusters of elements are sampled independently from each sampled PSU. Let y_{hik} be the value of the characteristic y associated with the ultimate unit (hik) in the sample s, that is, the sampled element k belonging to the i^{th} selected PSU in stratum h. Let w_{hik} denote the corresponding basic weight, which is simply the inverse of the inclusion probability for the ultimate unit (hik). The basic design-unbiased estimator of the population total Y is then given by:

$$\hat{Y} = \sum_{(hik)\in s} w_{hik} y_{hik}.$$

The direct analytic unbiased estimation of the design-based variance of $\hat{Y}$ is complicated. For the purpose of variance estimation, it is convenient to assume with-replacement sampling at the first stage, even though actual sampling is usually done without replacement. The following simplifying formula for variance estimation is generally used:

$$\operatorname{var}\left(\hat{Y}\right) = \sum_{h=1}^{H} \frac{1}{n_h(n_h-1)} \sum_{i=1}^{n_h} \left(y_{hi} - \overline{y}_h\right)^2,$$

where $y_{hi} = \sum_{k\in(hi)} n_h w_{hik} y_{hik}$ and $\overline{y}_h = n_h^{-1} \sum_{i=1}^{n_h} y_{hi}$. The variance estimator $\operatorname{var}(\hat{Y})$ generally results in overestimation, but the relative bias is negligible provided that the first stage sampling fractions n_h/N_h are small.

We now describe the resampling schemes suitable for the stratified multistage sampling.

3.1.1. The delete-one-cluster jackknife in stratified multistage design

To obtain the delete-one-PSU jackknife estimator, one PSU is omitted at a time from the original sample. Let $s^{(gj)}$ be a jackknife sample obtained after removing PSU j from stratum g. The original basic weights are modified as follows:

$$w_{hik(gj)} = \begin{cases} w_{hik} n_h (n_h-1)^{-1}, & \text{if } h = g, i \neq j, \\ 0, & \text{if } h = g, i = j, \\ w_{hik}, & \text{if } h \neq g. \end{cases}$$

The jackknife variance estimator for the unbiased estimator $\hat{Y}$ is then given by

$$v_{\mathrm{JK}}\left(\hat{Y}\right) = \sum_{h=1}^{H} \frac{n_h - 1}{n_h} \sum_{j\in s_h} \left\{\hat{Y}^{(hj)} - \hat{Y}^h\right\}^2.$$

where $\hat{Y}^{(gj)} = \sum_{(hik)\in s} w_{hik(gj)} y_{hik}$, and $\hat{Y}^g = \sum_{j\in s_g} \hat{Y}^{(gj)}/n_g$.

For a nonlinear function of population totals, $\theta = g(\mathbf{Y})$, Rao and Wu (1985) considered different variations of $v_{\mathrm{JK}}(\hat{\theta})$ and found them to be asymptotically equivalent. For example, $\hat{\theta}^h$ in

$$v_{\mathrm{JK}}(\hat{\theta}) = \sum_{h=1}^{H} \frac{n_h - 1}{n_h} \sum_{j\in s_h} \left\{\hat{\theta}^{(hj)} - \hat{\theta}^h\right\}^2$$

may be replaced by $n^{-1}\sum_{h=1}^{H}\sum_{j\in s_h}\hat{\theta}^{(hj)}$ or $H^{-1}\sum_{h=1}^{H}\hat{\theta}^{h}$.

3.1.2. The balanced half samples

The balanced half samples (BHS), also called the balanced repeated replications (BRR), is another popular method for variance estimation in sample surveys. The basic idea of the balanced half samples method can be attributed to the work of Hurwitz, Gurney, and others in the late 1950s and early 1960s in the context of estimating variances of estimators from the Current Population Survey conducted by the U.S. Census Bureau. The notion of balancing in the method was formalized by McCarthy (1969) who also coined the name "pseudoreplication" for the procedure of resampling from the original sample.

First, consider a stratified design when two elements are selected from each stratum using SRSWR. Under this sampling design, an unbiased estimator of the population total Y is given by $\hat{Y}=\sum_{h=1}^{H}(w_{h1}y_{h1}+w_{h2}y_{h2})$, where $w_{h1}=w_{h2}=N_h/2$.

The variance of $\hat{Y}$ is

$$\operatorname{Var}\bigl(\hat{Y}\bigr)=\frac{1}{2}\sum_{h=1}^{H}N_h(N_h-1)S_h^2,$$

where $S_h^2=(N_h-1)^{-1}\sum_{i\in U_h}(Y_{hi}-\overline{Y}_h)^2$ and $\overline{Y}_h=N_h^{-1}\sum_{i\in U_h}Y_{hi}$.

An unbiased estimator of the variance is

$$\operatorname{var}_{\mathrm{wr}}(\hat{Y})=\frac{1}{2}\sum_{h=1}^{H}N_h^2s_h^2,$$

with $s_h^2=\sum_{i\in s_h}(y_{hi}-\overline{y}_h)^2$ and $\overline{y}_h=(y_{h1}+y_{h2})/2$.

The BHS method consists of drawing replicate subsamples $s^{(\alpha)}$ such that every subsample contains exactly one of the two elements from each stratum. Each subsample $s^{(\alpha)}$ obtained this way has half the original sample size, and hence the name "half-samples." The weight of each unit selected into the sample is, therefore, doubled:

$$w_{hi}^{(\alpha)}=\begin{cases}2w_{hi}, & \text{if } i\in s_h^{(\alpha)},\\ 0, & \text{otherwise,}\end{cases}$$

where $s_h^{(\alpha)}$ denote the set of elements chosen from stratum h in the αth pseudoreplicate. The α^{th} BHS pseudoreplicate estimate of Y is given by $\hat{Y}^{(\alpha)}=\sum_{h=1}^{H}w_{hi}^{(\alpha)}y_{hi}$.

The variance of $\hat{Y}^{(\alpha)}$ over all possible 2^H half-samples is equal to the estimate of variance,

$$\operatorname{Var}_*\Bigl\{\hat{Y}^{(\alpha)}\Bigr\}=\operatorname{var}_{\mathrm{wr}}\Bigl(\hat{Y}\Bigr),$$

which follows immediately from (7).

Note that the total number of replicates increases rapidly with the increase of the number of strata. The word "balanced" in the name of the method refers to an innovative technique used to substantially reduce the number of required replications. The reduction is achieved by using a Hadamard matrix, an orthogonal matrix whose entries are either $+1$ or -1.

A thorough discussion of BHS for complex sampling designs can be found in Wolter (1985). We only give an outline of the procedure.

The PSUs are first randomly divided into two groups so that each random group contains exactly one PSU from each stratum. Let $\delta_{hi} = +1$ if the unit (hi) is in the first group and -1 otherwise. Each stratum is ascribed to a column of the Hadamard matrix of a specific dimension. Rows of the Hadamard matrix define the contents of replicate half-samples. Let $H(\alpha, h)$ denote the element in the αth row and hth column of the Hadamard matrix. The replicate half-sample $s^{(\alpha)}$ consists of the units (hi) for which δ_{hi} and $H(\alpha, h)$ are of the same sign. To achieve the full orthogonal balance, any H columns of the Hadamard matrix can be used except for the column with all $+1$. The full orthogonal balance guarantees a desirable property that the average of the replicate estimates of a linear estimator equals the full sample estimator. The Hadamard matrices are necessarily of order 1, 2, or a multiple of 4. Thus, the dimension of the Hadamard matrix applied to H strata is to be between $H+1$ and $H+4$. This means that the number of required replications, say A, for the BHS method of the variance estimation can be reduced to $H + 1 \leq A \leq H + 4$.

The exact variance $\mathrm{Var}_*\left(\hat{Y}^{(\alpha)}\right)$, based on A replicates, is given by

$$v_{\mathrm{BHS}}\left(\hat{Y}\right) = \frac{1}{A}\sum_{\alpha=1}^{A}\left\{\hat{Y}^{(\alpha)} - \hat{Y}\right\}^2.$$

In case some strata have more than two PSUs, adjustments to the basic BHS plan are necessary. One approach is to form two random groups of PSUs in each stratum and apply the BHS to the grouped PSUs as if they were one PSU. This method is called the grouped BHS (GBHS). The method, however, leads to inefficient variance estimators (Wu, 1991) and does not produce consistent variance estimators, in a sense that $v_{\mathrm{GBHS}}(\hat{Y})/\mathrm{var}_{\mathrm{wr}}(\hat{Y})$ does not converge to 1 in probability as the strata sample sizes n_h go to infinity (Rao and Shao, 1996).

Rao and Shao (1996) proposed a modification to the GBHS. After the GBHS estimate is computed, PSUs are regrouped into new random groups and the procedure is repeated. After R such iterations of the GBHS, an average of the R estimates, $v_{\mathrm{RGBHS}} = R^{-1}\sum_{r=1}^{R} v_{\mathrm{GBHS},r}$, provides an asymptotically consistent estimator of variance when R and strata sample sizes increase. This method is called the repeatedly grouped BHS (RGBHS).

Another modification to the basic BHS method, due to Fay (Judkins, 1990), is particularly useful when deleting half of the observations could lead to inefficient estimates. Instead of doubling or zeroing the original weights, as is done in the basic BHS method, the Fay's approach perturbs the weights as follows:

$$w_{hi}^{(\alpha)} = \begin{cases} (1+\varepsilon)w_{hi}, & \text{if } i \in s_h^{(\alpha)}, \\ (1-\varepsilon)w_{hi}, & \text{otherwise,} \end{cases}$$

where ε is a predefined constant factor. The final formula for the modified or Fay's, as the method is often called, BHS (MBHS or FBHS) needs to be adjusted:

$$v_{\mathrm{MBHS}}\left(\hat{Y}\right) = \frac{1}{A\varepsilon^2}\sum_{\alpha=1}^{A}\left\{\hat{Y}^{(\alpha)} - \hat{Y}\right\}^2. \tag{12}$$

The usual choice for ε is $1/2$.

Rao and Shao (1999) studied the properties of the MBHS for various choices of ε. They considered a more general case where the size m_h of a replicate sample in stratum h is not necessarily half of the original sample. Using result (10),

$$w_{hi}^{(\alpha)} = \begin{cases} 1 + \varepsilon\left(\frac{n_h - m_h}{m_h}\right)^{1/2} w_{hi}, & \text{if } i \in s_h^{(\alpha)}, \\ 1 - \varepsilon\left(\frac{m_h}{n_h - m_h}\right)^{1/2} w_{hi}, & \text{otherwise,} \end{cases}$$

and thus the final formula reduces to (12).

3.1.3. The bootstrap in stratified multistage design

The naïve version of the bootstrap, if applied without adjustments independently in each stratum, may significantly underestimate the variance. As we observe in (5), the bootstrap underestimates by the factor $(n-1)/n$. If the sample size n_h in stratum h is small, the underestimation becomes noticeable, and when there are many strata with small samples, the bias rapidly accumulates. To overcome this problem, several methods have been considered in the literature.

Rao and Wu (1988) proposed to draw m_h PSUs independently from each stratum h using SRSWR and modify the bootstrap estimates of means using a special adjustment. Rao et al. (1992) later enhanced the method by rescaling the bootstrap weights instead of the initially proposed adjustment to values y_{hik}. The bootstrap weights are defined as

$$w_{hik}^{(b)} = \left[\left\{1 - \left(\frac{m_h}{n_h - 1}\right)^{1/2}\right\} + \left\{\left(\frac{m_h}{n_h - 1}\right)^{1/2} \frac{n_h}{m_h}\right\} m_{hi}^{(b)}\right] w_{hik}, \tag{13}$$

where $m_{hi}^{(b)}$ is the number of times PSU (hi) is selected in the bootstrap sample b, $\sum_{(hi)\in s} m_{hi}^{(b)} = m_h$ (see result (9) for motivation of the weight adjustment). This method is called the *rescaling* bootstrap. In case $m_h = n_h - 1$, the formula for weights reduces to $w_{hik}^{(b)} = \left\{\frac{n_h}{n_h - 1}\right\} m_{hi}^{(b)} w_{hik}$. The method gives consistent estimator of variance of the estimator $\widehat{\theta} = g(\widehat{\overline{Y}})$, a smooth function of $\widehat{\overline{Y}}$.

In the linear case, when $n_h > 3$, the choice of $m_h = (n_h - 2)^2/(n_h - 1)$ leads to a bootstrap distribution (i.e., the distribution of the bootstrap estimates) whose third moment equals the unbiased estimate of the third moment of the estimator. The same choice of m_h, in case when strata variances σ_h^2 are known, ensures that the second-order term of the Edgeworth expansion of $Z = \left(\widehat{\overline{Y}} - \overline{Y}\right)/\sigma$ (where σ^2 is the true variance of $\widehat{\overline{Y}}$) matches the second-order term of the bootstrap distribution of this statistic, as the number of strata increases to infinity.

For smooth functions of population means, Rao and Wu (1988) obtained bootstrap-t confidence intervals. The idea is to approximate the distribution of $t = (\hat{\theta} - \theta)/s(\hat{\theta})$ by the bootstrap estimates $t^* = (\hat{\theta}^* - \hat{\theta})/\hat{s}(\hat{\theta}^*)$, where the estimate of $s(\hat{\theta}^*)$ can be obtained using, for example, jackknife variance estimation. The confidence intervals are obtained using the resulting bootstrap histogram of t^*.

3.2. Bootstrap methods for without replacement sampling

In devising various bootstrap methods, special attention was given to accounting for non-negligible sampling fractions $f_h = n_h/N_h$. Gross (1980) proposed a without-replacement bootstrap (BWO) method for variance estimation in case of SRSWOR. The method

consists of generating a pseudo population by replicating each sample measurement k times, where $k = N/n$ is assumed to be an integer. The SRSWOR of size n are repeatedly drawn from the pseudo population a large number (say, B) of times, and the variance is estimated using the usual bootstrap formula. Unfortunately, this method does not yield the unbiased variance estimator even for the linear estimator under SRSWOR design. However, extensions of this method that yield unbiased variance estimates for the linear case under a variety of sampling designs have been proposed in the literature (Bickel and Freedman, 1984; McCarthy and Snowden, 1985; Sitter, 1992b).

Rao and Wu (1988) developed variation of their rescaling bootstrap method that can be applied to sampling with unequal probabilities and without replacement. Sitter (1992a) proposed a method known as mirror-match bootstrap. We describe it for the single-stage stratified sampling, extentions are possible for the two-stage design and the Rao–Hartley–Cochran (1962) method of pps sampling. From each stratum h, SRSWOR of size $n_h^{'} < n_h$ are drawn independently k_h times. After each draw, resamples are replaced into the original sample. The number of times the resamples are drawn is $k_h = n_h\left(1 - f_h^*\right)/n_h'(1 - f_h)$, where $f_h^* = n_h'/n_h$. The resulting bootstrap sample size is $n_h^* = n_h\left(1 - f_h^*\right)/(1 - f_h)$. In case k_h is not an integer, Sitter suggested to use randomization between bracketing integers. The method yields a consistent variance estimator for $\widehat{\theta} = g\big(\widehat{\overline{Y}}\big)$. For linear statistics, when $f_h \geq 1/n_h$, the choice of $n_h^{'} = f_h n_h$ (hence, $f_h^* = f_h$, and the name, "mirror-match", of the procedure) ensures that the bootstrap histogram matches the Edgeworth expansion as the number of strata increases to infinity.

For a multistage design with SRSWOR and with nonnegligible sampling fractions, Funaoka et al. (2006) proposed a bootstrap method that yields consistent variance estimates for smooth and nonsmooth statistics. They called this method the Bernoulli bootstrap. To form a bootstrap sample, at each stage of the design, a unit is either kept in the bootstrap sample with some preassigned probability or replaced by a randomly selected unit from the sample. The procedure is repeated multiple times and the bootstrap estimate of variance is obtained using the standard bootstrap formula. The appeal of the method is its simplicity. The algorithm can be easily applied to sampling designs with three or more stages. In addition, the size of a bootstrap sample equals the original sample size, which is a desirable property in the case of imputation and the necessity to account for the variability associated with the imputation procedure.

3.3. More complex estimators

The estimator $\hat{\theta}$ is a linear estimator if it can be expressed as a linear function of the sample indicators. The sample indicator for a population unit takes the value 1 if the unit is sampled and 0 otherwise. Methods and the exact properties of variance estimators, considered in the previous sections, are readily extendable to the class of all linear estimators, but the main attraction of the replication methods is that the algorithm and the final formula can be applied without changes to more complex estimators.

Consider the class of nonlinear smooth functions of linear estimators. This class will be of interest when, for example, the target population quantity is a smooth function of a vector $\mathbf{Y}$ of population totals $\theta = g(\mathbf{Y})$. An estimator of θ is given by $\hat{\theta} = g(\hat{\mathbf{Y}})$, where $\hat{\mathbf{Y}}$ is a linear design-based unbiased estimator of $\mathbf{Y}$. For this class of estimators, Krewski and Rao (1981) established the consistency of the variance estimators obtained by the

Taylor linearization, jackknife, and BHS methods when the number of strata increases. In the previous subsection, we already mentioned that the rescaling bootstrap method of Rao, Wu, and Yue and the mirror-match approach of Sitter also yield consistent estimators of the variances for a nonlinear smooth functions of linear estimators (Rao and Wu, 1988; Sitter, 1992a) .

3.3.1. Delete-d jackknife for nonsmooth statistics

In many situations, we are interested in nonsmooth statistics such as, for example, the estimator of population quantiles. It is known (see Efron, 1982; Miller R.G., 1974b) that the delete-one-unit jackknife does not produce consistent estimators of variances for nonsmooth estimators. To treat this limitation of the jackknife, Shao and Wu (1989) proposed a modification called the *delete-d* jackknife. Instead of deleting one unit at a time, the delete-d procedure omits d observations. The variance is computed by formula (8) with $m = n - d$. The factor in formula (8) depends on the number d of deleted observations, and d can be chosen according to a measure of smoothness of the estimator to adjust the factor so that the final formula could produce a consistent estimator. In spirit, the delete-d jackknife is very similar to a particular form of bootstrap. The computations intensify with the increase of d. To reduce the number of replications, Shao and Wu proposed a balanced subsampling scheme. Motivated by the results, Rao et al. (1992) suggested that, in case of a multistage design, the delete-one-PSU jackknife might perform somewhat similar to the delete-d and better than the delete-one-unit jackknife for the iid case.

3.3.2. The generalized regression estimator

The generalized regression (GREG) estimator of Y is given by

$$\hat{Y}_{\mathrm{GREG}} = \sum_{(hik)\in s} w^*_{hik} y_{hik},$$

where

$$\begin{aligned} w^*_{hik} &= w_{hik} g_{hik}, \\ g_{hik} &= 1 + x^T_{hik} \hat{A}^{-1} \Big(X - \hat{X} \Big), \\ \hat{X} &= \sum_{(hik)\in s} w_{hik} x_{hik}, \\ \hat{A} &= \sum_{(hik)\in s} w_{hik} x_{hik} x^T_{hik}, \end{aligned}$$

x_{hik} being a vector of auxiliary variables with known population totals X (see also Chapter 25).

The following model-assisted variance estimator of $\hat{Y}_{\mathrm{GREG}}$ has been suggested by Särndal et al. (1989):

$$\mathrm{var}(\hat{Y}_{\mathrm{GREG}}) = \sum_{h=1}^{H} \frac{1}{n_h(n_h - 1)} \sum_{i=1}^{n_h} \big(e^*_{hi} - \overline{e}^*_h \big)^2,$$

where $e^*_{hi} = \sum_{k\in(hi)} n_h w^*_{hik} e_{hik}$, $e_{hik} = y_{hik} - x^T_{hik}\hat{B}$, $\hat{B} = \hat{A}^{-1}\hat{b}$, $\hat{b} = \sum_{(hik)\in s} w_{hik} x_{hik} y_{hik}$, and $\overline{e}^*_h = n_h^{-1} \sum_{i=1}^{n_h} e^*_{hi}$.

The jackknife estimator for $\hat{Y}_{\mathrm{GREG}}$ can be constructed as follows. Define the jackknife variance replicate estimate when PSU (gj) is deleted as

$$\hat{Y}^{(gj)}_{\mathrm{GREG}} = \sum_{(hik)\in s} w^*_{hik(gj)} y_{hik},$$

where $w^*_{hik(gj)} = w_{hik(gj)} g_{hik(gj)}$, and $g_{hik(gj)}$ is obtained from g_{hik} when $\hat{A}$ and $\hat{X}$ are replaced by $\hat{A}^{(gj)}$ and $\hat{X}^{(gj)}$ respectively. The definitions of $\hat{A}^{(gj)}$ and $\hat{X}^{(gj)}$ are exactly the same as $\hat{A}$ and $\hat{X}$ respectively, except that $w_{hik}s$ are replaced by $w_{hik(gj)}$.

The jackknife estimator is

$$v_{\mathrm{JK}}\left(\hat{Y}_{\mathrm{GREG}}\right) = \sum_{h=1}^{H} \frac{n_h - 1}{n_h} \sum_{j\in s_h} \left\{\hat{Y}^{(hj)}_{\mathrm{GREG}} - \hat{Y}_{\mathrm{GREG}}\right\}^2.$$

Yung (1996) established asymptotic equivalence of $v_{\mathrm{JK}}(\hat{Y}_{\mathrm{GREG}})$ and $\mathrm{var}(\hat{Y}_{\mathrm{GREG}})$ of Särndal et al. (1989) to the higher order term for the stratified paired selection design. Yung and Rao (1996) linearized the jackknife variance estimator $v_{\mathrm{JK}}(\hat{Y}_{\mathrm{GREG}})$ and obtained the jackknife linearization variance estimator that is identical to $\mathrm{var}(\hat{Y}_{\mathrm{GREG}})$.

The computation of the jackknife variance estimator $v_{\mathrm{JK}}(\hat{Y}_{\mathrm{GREG}})$ requires the inversion of the matrix $\hat{A}^{(gj)}$ for each replicate. This can be avoided by applying the following approximation to $g_{hik(gj)}$ that uses the inverse of the full sample matrix $\hat{A}$:

$$\tilde{g}_{hik(gj)} = 1 + \left(w_{hik}/w_{hik(gj)}\right) x^T_{hik} \hat{A}^{-1}\left(X - \hat{X}^{(gj)}\right).$$

The resulting variance estimator is identical to the standard linearization variance estimator, which is obtained from $\mathrm{var}(\hat{Y}_{\mathrm{GREG}})$ when w^*_{hik} is replaced by w_{hik} in the definitions of e^*_{hi}, and $\overline{e}^*_h$.

3.3.3. Estimating function approach

Most often the goal of a survey is to estimate descriptive characteristics of a finite population, for example, population means, totals, functions of population totals, etc. In other cases, when more complex interrelationships of population measurements are of interest, a specific model is assumed for the population, and the aim of the survey is to estimate the parameters of a model. Whether there is an assumed model or not, most finite population parameters of interest, θ, can be formulated as a solution to population estimating equations

$$U(\theta) = \sum_{(hik)\in U} u_{hik}(\theta) = 0. \tag{14}$$

See Binder (1983), Godambe and Thompson (1986a), and Chapter 26 of this volume.

The problem reduces to the estimation of the finite population totals $U(\theta)$ by $\hat{U}(\theta) = \sum_{(hik)\in s} w_{hik} u_{hik}(\theta)$ and solving for θ the sample estimating equations $\hat{U}(\theta) = 0$.

In general, the solution to the estimating equations is obtained using iterative algorithms. For example, at the rth step of the Newton–Raphson algorithm

$$\hat{\theta}_{\mathrm{r}} = \hat{\theta}_{r-1} + \left\{\hat{J}\left(\hat{\theta}_{r-1}\right)\right\}^{-1} \hat{U}\left(\hat{\theta}_{r-1}\right),$$

where $\hat{J}(\hat{\theta}_{r-1})$ is $\hat{J}(\theta) = -\partial\hat{U}(\theta)/\partial\theta$ evaluated at $\hat{\theta}_{r-1}$.

The customary jackknife or bootstrap variance estimator uses replicate estimates $\hat{\theta}^{(b)}$ that are solutions to corresponding replicate sample estimating equations

$$\hat{U}^{(b)}(\theta) = \sum_{(hik)\in s} w_{hik}^{(b)} u_{hik}(\theta) = 0,$$

where $w_{hik}^{(b)}$ are the corresponding replication weights. The one-step Newton–Raphson estimator is obtained as

$$\hat{\theta}^{(b)} = \hat{\theta} + \left\{\hat{J}^{(b)}(\hat{\theta})\right\}^{-1} \hat{U}^{(b)}\left(\hat{\theta}\right) \tag{15}$$

by using the full-sample estimate of θ as a starting value. The problem is that matrix $\hat{J}^{(b)}(\hat{\theta})$ in (15) is not always invertable for every replicate. The replicate estimates can be obtained by using $\hat{J}(\hat{\theta})$ that is based on the full sample (Rao and Tausi, 2004):

$$\tilde{\theta}^{(b)} = \hat{\theta} + \left\{\hat{J}(\hat{\theta})\right\}^{-1} \hat{U}^{(b)}\left(\hat{\theta}\right),$$

thus avoiding the need to invert every replicate matrix $\hat{J}^{(b)}(\hat{\theta})$. See also Rao (2006).

4. Variance estimation in the presence of imputation

Nonresponse is a common persistent problem in surveys. Missing responses are often filled in or imputed using different imputation methods. It is important to account for the added uncertainty in the data whenever imputation is used. Ignoring the fact that part of the data is imputed rather than observed and treating the approximate data as true observed values leads to an underestimation of the variance. Therefore, adjustments to the replication procedure need to be considered.

One approach to account for the additional uncertainty due to imputation is the multiple imputation (MI) technique proposed by Rubin (1978). This method entails construction and analyzing multiple sets of independently imputed data. One important requirement for the design-based validity of MI is that the method used for the imputation must be "proper" (as defined in Rubin, 1987, pp. 118–119). However, for various reasons, imputation methods commonly used in practice are "improper." Resampling procedures reviewed in this section can deal with improper imputation methods.

Technical precondition necessary for the ability to implement resampling methods is that the data set with the originally imputed data must contain information on the response status and imputation class. The usual assumption about the response mechanism within each imputation class is that probability of response is the same for all units, and events of nonresponse occur independently (the uniform response mechanism).

In choosing a replication strategy, it is important to distinguish two major types of imputation: deterministic and random. Under the former, an imputed value is nonrandom

given a set of reported values and it equals the expectation with respect to the imputation model. Examples include ratio, regression, and mean imputation. A commonly used method of weight adjustment to account for unit nonresponse can be viewed as a particular case of implicit deterministic imputation. In this case, the mean of the reported values within each weight adjustment class is implicitly used to impute for the missing values.

Under the random imputation method, an imputed data is a random variable and can be viewed as its expectation under the imputation model plus a random noise. Hot-deck imputation is an example of random imputation. Other examples are random ratio or random regression methods, where the imputed value is constructed by adding, to the model expectation, a random term drawn from a set of residuals.

Intuitively, a natural way to account for imputation is to reimpute values independently for each replicate sample in the same way as the original sample data are imputed. Burns (1990) applied it to jackknife replication for a stratified multistage design. Replicate estimates, derived after the reimputation, were used in the standard jackknife formula

$$v_{\mathrm{JK}}\left(\hat{\theta}\right) = \frac{n-1}{n} \sum_{j \in s} \left\{ \hat{\theta}^{(j)} - \hat{\theta} \right\}^2. \tag{16}$$

However, Rao and Shao (1992) showed that, in the case of a random imputation, this method leads to an overestimation of the variance. This is because the original sample estimate $\hat{\theta}$, customarily used in the standard jackknife formula, is itself random when random imputation method is used. Saigo et al. (2001) noted that, with random reimputation, it is necessary to use average of the replicate estimates, $\overline{\hat{\theta}} = n^{-1} \sum_{j \in s} \hat{\theta}^{(j)}$, in the final formula:

$$v_{\mathrm{JK}}\left(\hat{\theta}\right) = \frac{n-1}{n} \sum_{j \in s} \left\{ \hat{\theta}^{(j)} - \overline{\hat{\theta}} \right\}^2.$$

As an alternative to reimputation, Rao and Shao (1992) proposed an adjustment to the imputed value when a responding unit is omitted from the jackknife sample. The imputed values are adjusted using the difference in expectations under the imputation model, $E_*^{(j)} y_i^* - E_* y_i^*$, as follows:

$$y_i^{(j)} = \begin{cases} y_i, & i \in s_{\mathrm{r}}, \\ y_i^* + E_*^{(j)} y_i^* - E_* y_i^*, & i \notin s_{\mathrm{r}}, \end{cases} \tag{17}$$

where $y_i^{(j)}$ is the value ascribed to unit i when unit j is omitted, s_{r} is a set of respondents, y_i^* is the originally imputed value; $E_* y_i^*$ and $E_*^{(j)} y_i^*$ are expectations under the imputation model given the original and the jackknife sets of respondents, respectively. When unit $i \notin s_{\mathrm{r}}$ is omitted, $E_* y_i^* = E_*^{(j)} y_i^*$, and no adjustment is made. The jackknife variance estimator is given by the standard formula and is consistent as the sample size increases.

Next, we describe an important application of the adjusted jackknife to the weighted hot-deck imputation considered in Rao and Shao (1992). For a stratified multistage sampling design, the hot-deck imputation procedure itself needs adjustment. Suppose that the sample is divided into imputation classes and that a uniform response mechanism is assumed within each class. These imputation classes may comprise of a mixture of

strata. The weighted hot-deck method of imputation chooses donors independently within each imputation class, I, from a set of respondents, $I_r \subset I$, with probabilities $w_{hik} \Big/ \sum_{(hik)\in I_r} w_{hik}$ using with replacement sampling. Under this imputation scheme, the expectation of imputed value y^* over the imputation class I is given by

$$E_{I*} y^* = \sum_{(hik)\in I_r} w_{hik} y_{hik} \Big/ \sum_{(hik)\in I_r} w_{hik}.$$

If a unit $j \in I_r$ that belongs to stratum g is omitted, the expectation based on the remaining respondents inside the imputation class I, is given by

$$E^{(gj)}_{I*} y^* = \sum_{(hik)\in I_r} w^{(gj)}_{hik} y_{hik} \Big/ \sum_{(hik)\in I_r} w^{(gj)}_{hik},$$

where $w^{(gj)}_{hik}$ are the jackknife replicate weights.

For the imputed value $y^*_{h'i'k'}$ of a unit $(h'i'k')$ from the imputation class I, the adjusted jackknife value is given by

$$y^{(gj)}_{h'i'k'} = y^*_{h'i'k'} + \frac{\sum\limits_{(hik)\in I_r} w^{(gj)}_{hik} y_{hik}}{\sum\limits_{(hik)\in I_r} w^{(gj)}_{hik}} - \frac{\sum\limits_{(hik)\in I_r} w_{hik} y_{hik}}{\sum\limits_{(hik)\in I_r} w_{hik}},$$

and the resulting jackknife estimator is given by the standard formula. The described methodology also finds its application in the two-phase sampling context considered in the next section.

Rao (1996) reviewed adjusted jackknife replication for various commonly used imputation methods, under stratified simple random sampling and multistage designs. Shao et al. (1998) used similar adjustments for BRR. The BRR has the advantage of producing consistent variance estimators for sample quantiles. This is important for random imputation methods that are often applied because they provide better approximations to the distributional characteristics, such as sample quantiles, than the deterministic methods do.

The operational advantage of the aforementioned value adjustment methods is quite important for the random imputation method. Reimputation for deterministic imputation does not present technical problems. If a program for the original imputation already exists, it is not difficult to adjust it for repeated applications and, with modern computer power, nonrandom type of reimputation is not a significant obstacle. Random reimputation, on the other hand, would entail repeated selection of random samples of donors from the replicated set of respondents, a much more computer intensive procedure than the non-random variant.

For bootstrap, adjustments similar to those of Rao and Shao (1992) are not always appropriate: the method does not provide valid estimators of variances for sample quantiles (Shao and Sitter, 1996). Efron (1994) proposed a bootstrap application with imputed data for simple random sampling, and Shao and Sitter (1996) studied reimputation with the bootstrap under stratified multistage design. The bootstrap procedure is straightforward. Under the assumption of uniform response mechanism within strata, bootstrap samples of size $n_h - 1$ are selected with replacement from the original sample, independently across different strata; reimputation is performed for each bootstrap sample using the original imputation method and resulting bootstrap estimates are used in the

standard formula. Shao and Sitter (1996) showed the consistency of the estimator for large stratum sample sizes.

Saigo et al. (2001) extended the bootstrap approach when the stratum sample sizes are very small. They proposed a method called the repeated half sample bootstrap that can be described as follows. When the stratum sample size n_h is an even number, the sample is randomly divided into halves, and one of them is selected. Instead of doubling the weights, each selected record is repeated twice (hence the name of the method). The trick of the repetition is the key for the method to work properly for random imputation methods. When n_h is odd, the procedure is adjusted using the following method (1) with probability 1/4 and method (2) with probability 3/4 for selecting bootstrap samples from a given stratum h:

(1) select a sample of size $(n_h - 1)/2$, duplicate the selected observations and draw one more observation from the sample that has just been selected in order to obtain n_h observations in total;
(2) select a sample of size $(n_h + 1)/2$, duplicate the selected observations and randomly delete one observation to obtain n_h observations in total.

Saigo et al. (2001) also proposed similar procedures of repeated replications for the BRR method.

5. Resampling methods for sampling designs in two phases

In two-phase sampling, or double sampling, some variables are observed in a large first phase sample. These variables are then used to construct a sampling plan for selecting the second phase units out of the first phase sample. Some additional variables, whose measurement can typically be costly, are observed on the second phase units. The theory available for the two-phase sampling can be adapted to survey nonresponse theory, under certain assumptions about the response mechanism and conditional on the number of respondents, if a set of respondents is viewed as a second-phase sample.

Rao and Sitter (1995) considered the ratio estimator and Sitter (1997) studied linear regression estimator under two-phase sampling, where SRSWOR is used to draw the first-phase sample as well as a subsample from the first-phase sample. Jackknife replicates are created by omitting one unit in turn from the larger first-phase sample, and the resulting estimator is obtained by the standard formula:

$$v_{\mathrm{JK}} = \frac{n_1 - 1}{n_1} \sum_{j \in s_1} \left\{ \hat{\theta}^{(j)} - \hat{\theta} \right\}^2,$$

where n_1 is the first-phase sample size, and the replicate estimate $\hat{\theta}^{(j)}$ is defined as usual analog to the full-sample estimate $\hat{\theta}$. This "full-sample" jackknife variance estimator requires n_1 replicates.

The first-phase sample is often much larger than the second-phase sample. In this case, it is desirable to reduce the number of replications. Fuller (1998) proposed to accomplish this under certain assumptions. Kim and Sitter (2003) extended his ideas. They decomposed the jackknife formula into two parts: in the first part the variance is based on the replicates constructed by omitting units that belong to the smaller sample and in the second part the omitted units belong to the complement of the second-phase

sample. The first part requires n_2 replicates, the number of units in the smaller sample. However, for the second part, in many common situations, it is possible to use much fewer replicates than $n_1 - n_2$, the number of units in the complement to the smaller sample. The resulting variance estimator has similar properties as the full-sample estimator. Kim et al. (2006) considered the situation when the first-phase sample information is used to define strata for the second phase of sampling. They proposed variance estimators for the double-expansion estimator (DEE) and for the reweighted expansion estimator (REE) (also considered by Kott, 1990; Kott and Stukel, 1997). This REE approach is analogous to the weighted hot-deck imputation approach considered in Rao and Shao (1992) and reviewed in Section 4, and the adjusted jackknife variance works in this setting. Kim et al. (2006) noted that DEE can be viewed as particular case of REE and proposed a variance estimator for this case.

6. Resampling methods in the prediction approach

The jackknife variance estimators given in this section follow from Valliant et al. (2000, Chapter 11) when the superpopulation model is a linear model. The consistency of the jackknife variance estimators discussed in this section follows from Jiang et al. (2002). Consider a finite population of N elements. Let y_i and x_i be the values of the dependent and auxiliary variables for unit i of the finite population ($i = 1, \ldots, N$). We assume that the values of the auxiliary variable are known for all units of the finite population. Let s be a sample drawn from the finite population. To illustrate the prediction approach, assume that the values of the dependent variable for the finite population are generated from the following superpopulation model:

$$y_i = \beta x_i + e_i, \tag{18}$$

where e_i are uncorrelated with zero means and variances $\sigma^2 x_i$. Our goal is to estimate the finite population mean, $\overline{Y} = N^{-1} \sum_{i=1}^{N} y_i$. It can be shown (see Chapter 23) that the best linear unbiased predictor (BLUP) of $\overline{Y}$ is given by

$$\overline{Y}^{\mathrm{BLUP}} = N^{-1} \left(\sum_{i \in s} y_i + \hat{\beta} \sum_{i \notin s} x_i \right) = \frac{\overline{y}}{\overline{x}} \overline{X},$$

where $\overline{y} = n^{-1} \sum_{i \in s} y_i$ and $\overline{x} = n^{-1} \sum_{i \in s} x_i$ are the sample means of the dependent and auxiliary variables, respectively; $\overline{X} = N^{-1} \sum_{i=1}^{N} x_i$ is the population mean of the auxiliary variable; $\hat{\beta} = \overline{y}/\overline{x}$ is the weighted least squares estimator of β under model (18). It is interesting to note that here we do not need information on the auxiliary variable on all the unobserved units of the finite population. So, in this case the prediction approach yields the usual ratio estimator, which is approximately unbiased and consistent for large sample size n under the usual design-based framework irrespective of the model.

The prediction variance is defined as $\mathrm{Var}(\overline{Y}^{\mathrm{BLUP}} - \overline{Y})$, where the variance is with respect to the assumed linear model. Note that

$$\begin{aligned} \mathrm{Var}\left(\overline{Y}^{\mathrm{BLUP}} - \overline{Y}\right) &= \mathrm{Var}\left\{(1-f)\overline{X}_{\mathrm{ns}}\hat{\beta} - (1-f)\overline{Y}_{\mathrm{ns}}\right\} \\ &= \mathrm{Var}\left\{(1-f)\overline{X}_{\mathrm{ns}}\hat{\beta}\right\} + \mathrm{Var}\left\{(1-f)\overline{Y}_{\mathrm{ns}}\right\}, \end{aligned}$$

where $f = n/N$, the sampling fraction, $\overline{X}_{\text{ns}} = (N-n)^{-1}\sum_{i\notin s} x_i$ and $\overline{Y}_{\text{ns}} = (N-n)^{-1}\sum_{i\notin s} y_i$ means of the auxiliary and the dependent variables for the finite population units that are not sampled. Because the population size N is usually much larger than the sample size n, the order of the second term ($O[(N-n)^{-1}]$) is much smaller than that of the leading term ($O(n^{-1})$). Thus, essentially we are interested in estimating $\text{Var}\{(1-f)\overline{X}_{\text{ns}}\hat{\beta}\}$, that is the variance of a linear function of the weighted least square estimator of β. Thus, a jackknife estimator of the prediction variance is given by

$$v_{\text{J}} = (1-f)^2\overline{X}_{\text{ns}}^2\frac{n}{n-1}\sum_{k=1}^{n}\left(\hat{\beta}_{-k}-\hat{\beta}\right)^2,$$

where $\hat{\beta}_{-k}$ is the weighted least square estimator of β obtained by deleting the kth observation.

Valliant et al. (2000, Chapter 11) considered the prediction of a smooth nonlinear function of predictors of finite population totals for several variables and proposed a jackknife estimator of the prediction variance of their predictor under a two-stage sample design. Their superpopulation model allows for a general but known within PSU correlation structure. For certain covariance matrix structures such as a block diagonal structure, which is reasonable for a clustered finite population, Lahiri (2008) presents consistent jackknife variance estimators of empirical best predictors (EBP) of finite population means when the covariance matrix is unknown.

Now consider the case when the dependent variable is binary. In this case the linear model for the superpopulation is not appropriate. We assume that the dependent variable for the finite population is generated from the following logistic model:

$$\text{logit}\{P(y_i = 1)\} = \beta x_i$$

for $i = 1, \ldots, N$. In this case, an EBP of the finite population mean is given by:

$$\begin{aligned}\overline{Y}^{\text{EBP}} &= N^{-1}\left\{\sum_{i\in s} y_i + \sum_{i\notin s}\frac{\exp(\hat{\beta}x_i)}{1+\exp(\hat{\beta}x_i)}\right\}\\ &= f\overline{y} + (1-f)h(\hat{\beta}),\end{aligned}$$

where $h(\hat{\beta}) = (N-n)^{-1}\sum_{i\notin s}\exp(\hat{\beta}x_i)/[1+\exp(\hat{\beta}x_i)]$ and $\hat{\beta}$ is a consistent estimator of β. We note that here we need the values of the auxiliary variable for every unit of the finite population. In this case, the prediction variance is given by

$$\begin{aligned}\text{Var}\left(\overline{Y}^{\text{EBP}} - \overline{Y}\right) &= \text{Var}\left\{(1-f)h(\hat{\beta})\right\} + \text{Var}\left\{(1-f)\overline{Y}_{\text{ns}}\right\}\\ &\approx \text{Var}\left\{(1-f)h(\hat{\beta})\right\},\end{aligned}$$

because the order of the second term is much smaller than the leading term. Using the general theory of Jiang et al. (2002), a jackknife estimator of the prediction variance is given by empirical best predictors (EBP):

$$v_{\text{J}} = (1-f)^2\frac{n}{n-1}\sum_{k=1}^{n}\left\{h(\hat{\beta}_{-k}) - h(\hat{\beta})\right\}^2.$$

Following the two examples given in this section, it is possible to obtain a jackknife estimator of the prediction variance under the fairly general set-up of Jiang et al. (2002). See Valliant et al. (2000) for a discussion of the BRR method in the prediction theory.

7. Resampling methods in small area estimation

In a large scale sample survey, sample design is usually developed to provide reliable design-based estimators of parameters of interest for the survey finite population or certain large subgroups of the survey population. Estimation of small subgroups of the survey population is also of interest, but the survey data typically provide small samples or even no sample for small subgroups of the survey population resulting in either highly unstable design-based estimates or no design-based direct estimate. Meza et al. (2003) provide an example where the overall sample size of a statewide telephone survey in Nebraska, USA, is about 4300, which is large enough to produce reliable design-based estimates of the prevalence of alcohol abuse for the entire state and large counties or state level large demographic groups. However, the sample sizes are very small for some small counties or demographic subgroups within counties. For example, in the Boone county the sample size is 14 and there is only one white female in the age-group of 25–44. The estimation problem that arises due to small sample sizes for subgroups of a survey population is referred to as the small area (domain) estimation problem in the sample survey literature. Small area statistics are needed in regional planning and fund allocation in many government programs and thus the importance of producing reliable small area statistics cannot be over-emphasized.

To reduce the sampling errors in the traditional design-based direct estimators for the small areas, one can combine information from the sample survey, various administrative/census records, and even previous surveys using suitable models. A natural question is: how effective are the linear models that are typically used in the prediction approach to finite population sampling? Lahiri (2008) offers a detailed explanation for the problem caused by a linear (fixed effects) model in the small area estimation. To this end, let y_{ij} and x_{ij} be the values of the dependent and auxiliary variables for unit j of the ith area ($i=1,\ldots,m;\ \ j=1,\ldots,N_i$). We are interested in estimating the small area finite population means $\overline{Y}_i = N_i^{-1}\sum_{j=1}^{N_i} y_{ij}$ based on a stratified SRS sample s with areas as the strata and known totals for the auxiliary variables $\overline{X}_i = N_i^{-1}\sum_{j=1}^{N_i} x_{ij}$, $(i=1,\ldots,m)$. We can assume the superpopulation model (18) with either a common slope parameter β for all areas or area specific fixed slope parameters β_i for the areas. Under the prediction approach to finite population sampling, the standard empirical best predictors in both cases are unbiased for $\overline{Y}_i$ under their respective assumed linear models. The predictors are design-unbiased for their respective small area means in the latter case when $x_{ij}=1,\ \ (i=1,\ldots,m;\ j=1,\ldots,N_i)$. However, the predictors are design-biased in general and the bias depends on the sample size and the extent of heterogeneity across the small areas in the former case and the sample size in the latter case. In the former case, the predictors are usually stable in terms of both design-based and prediction variances, the order being $O(n^{-1})$, under certain mild regularity conditions. In contrast, in the latter case the predictors are generally unstable, both in terms of the design-based and prediction variances, the order being $O\left(n_i^{-1}\right)$, under certain regularity conditions.

A compromise between the two versions of the linear models is a linear mixed model:

$$y_{ij} = \beta_i x_{ij} + e_{ij}, \tag{19}$$

where β_i and e_{ij} are uncorrelated with $\beta_i \sim N\left[0, \sigma_\beta^2\right]$ and $e_{ij} \sim N\left[0, x_{ij}\sigma_e^2\right]$. Under this linear mixed model, the best predictor (BP) is given by

$$\overline{Y}_i^{\mathrm{BP}} = N_i^{-1}\left(\sum_{j\in s_i} y_{ij} + \hat{\beta}_i \sum_{j\notin s_i} x_{ij}\right),$$

where s_i is the set of units in area i that belongs to the sample s,

$$\hat{\beta}_i = (1-d_i)\frac{\overline{y}_i}{\overline{x}_i} + d_i\beta, \quad d_i = \frac{\sigma_e^2/\sum_{j\in s} x_{ij}}{\sigma_\beta^2 + \sigma_e^2/\sum_{j\in s} x_{ij}}$$

An EBP of $\overline{Y}_i$, say $\overline{Y}_i^{\mathrm{EBP}}$, is obtained when consistent estimators of the model parameters β, σ_β^2, and σ_e^2 are plugged in the $\overline{Y}_i^{\mathrm{BP}}$ formula. The EBP $\overline{Y}_i^{\mathrm{EBP}}$ generally outperforms the EBP derived from a linear model; see Lahiri (2008) for details.

Model (19) is referred to as a unit level model in the small area estimation literature. However, in many small area estimation problems, it is not possible to use the unit level model simply because of the unavailability of the unit level information. Chen et al. (2007) used the following area level model:

$$y_i = \theta_i + e_i = x_i'\beta + v_i + e_i, \quad i = 1, \ldots, m, \tag{20}$$

where v_i and e_i are all uncorrelated with $v_i \sim [0, A]$ and $e_i \sim [0, D_i]$; $[a, b]$ denotes a distribution with mean a and variance b; the sampling variances D_i are assumed to be known.

In the aforementioned model, e_i is used to account for the sampling variability of the regular survey estimates y_i of true small area means $\theta_i = x_i'\beta$. The area specific random effects v_i link the true small area means θ_i to a vector of p known auxiliary variables x_i, often obtained from various administrative and census records. The parameters β and A of the linking model are generally unknown and are estimated from the available data. To estimate the sampling variability D_i, Fay and Herriot (1979) employed generalize variance function (GVF; see Wolter, 1985) method that uses some external information from the survey. See Hinrichs (2003), and Gershunskaya and Lahiri (2005) for recent developments on variance estimation for the domains.

The well-celebrated Fay–Herriot model (Fay and Herriot, 1979) is obtained as a special case of the model (20) when normality is assumed for both v_i and e_i. Fay and Herriot (1979) used the area specific random effect v_i to relate the true per-capita income (θ_i) to the auxiliary variables (x_i) obtained from the census, housing and Internal Revenue Service records. In other words, Fay and Herriot (1979) used random effects to capture the additional area specific effects not explained by the area specific auxiliary variables. This is achieved at the expense of an additional unknown variance component A to be estimated from the data. In contrast, the corresponding regression model fails to capture this additional area specific variability. Using the U.S. census data, Fay and Herriot (1979) demonstrated that their EB estimator (also an EBLUP) performed better than the direct survey estimator and a synthetic estimator used earlier by the U.S. Census Bureau.

Under the Fay–Herriot model, the BP of θ_i is given by:

$$\hat{\theta}_i(y_i; \phi) = (1 - B_i)y_i + B_i x_i'\beta,$$

where $B_i = D_i/(D_i + A)$, $\phi = (\beta, A)'$. Replacing β by the weighted least square estimator

$$\hat{\beta}(A) = \left(\sum_{i=1}^{m} \frac{x_i x_i'}{D_i + A}\right)^{-1} \sum_{i=1}^{m} \frac{x_i y_i}{D_i + A},$$

we get the following BLUP of θ_i:

$$\hat{\theta}_i(y_i; A) = (1 - B_i)y_i + B_i x_i'\hat{\beta}(A).$$

Replacing A by $\hat{A}$, a consistent estimator of A, we get $\hat{\theta}_i(y_i; \hat{A})$, an EBLUP of θ_i.

The unit level model and the area level model are two examples of the mixed models, which are particularly suitable for small area estimation because of their flexibility in effectively combining different sources of information and explaining different sources of errors. Mixed models typically incorporate area-specific random effects that explain the additional between area variation in the data not explained by the fixed effects part of the model. In contrast, an implicit regression model that motivates a synthetic estimation method assumes no between area variation other than those explained by the area-specific auxiliary variable(s).

The EBP method has been extensively used in small area estimation (e.g., Rao, 2003; Jiang and Lahiri, 2006). Although EBP is fairly easy to obtain, the problem of providing a suitable estimate of its uncertainty measure that accounts for all sources of variation is a highly nontrivial problem and has sparked a huge volume of research over the last couple of decades. The traditional design-based mean squared error, unconditional (over the joint distribution of the observations and random effects) mean squared error, and different types of conditional model-based mean squared errors are of interest. In defining the conditional model-based mean squared error, both conditioning on the area specific random effects and conditioning on the area specific observations have been considered. Using a few illustrative simple examples, Lahiri (2008) points out the difficulty in obtaining stable estimators of design-based mean squared errors of small area estimators, although achieving good design-bias property is usually not a problem.

For the Fay–Herriot model, Prasad and Rao (1990) obtained a second-order unbiased or nearly unbiased estimator of the MSE (i.e., unconditional MSE) defined as $E(\hat{\theta}_i - \theta_i)^2$, where the expectation is taken over the joint distribution of observations y_i and random effects v_i. Rivest and Belmonte (1999) proposed an unbiased estimator of the conditional model-based mean squared error, defined as: $E\{(\hat{\theta}_i - \theta_i)^2|\theta\}$, where the expectation is taken over the observations y_i given θ. Hwang and Rao (1987; unpublished work) obtained a similar unbiased estimator earlier and, using a Monte Carlo simulation study, showed that the estimator is more unstable than the Prasad–Rao estimator of the unconditional mean squared error. Interestingly, their simulation results showed that the Prasad–Rao MSE estimator tracks the conditional mean squared error very well even under a moderate derivation from the model on the random effects. However, in terms of the conditional bias, the Prasad–Rao mean squared error estimator could perform poorly compared to the unbiased estimator of the conditional mean squared error for

an outlying small-area. See Rao (2003, Chapter 4) for more discussion on this type of conditional mean squared error estimation. We can also define conditional model-based MSE as $E[(\hat{\theta}_i - \theta_i)^2 | y_i]$, where the expectation is taken over the random effects v_i given the area specific observartions y_i. Chatterjee and Lahiri (2007) put forward a parametric bootstrap method to estimate such conditional MSE for a fairly general mixed model, which includes the Fay–Herriot model as a special case. Fuller (1990b) use a Taylor series method to estimate this conditional MSE whereas Booth and Hobert (1998) proposed a bootstrap method, which is different from that of Chatterjee and Lahiri (2007).

Extensive research has been conducted to estimate the unconditional mean squared error of EBP and a major portion of the research centered around linear mixed models. In the context of the linear mixed model, a naïve MSE estimator is given by the MSE of the BLUP with the model variance components replaced by suitable consistent estimators. But, it usually underestimates the true MSE of EBLUP mainly for two reasons. First, it fails to incorporate the extra variability incurred because of the estimation of the variance components and the order of this underestimation is $O(m^{-1})$, for large number of small areas m. Second, the naïve MSE estimator underestimates even the true MSE of the BLUP, The order of the underestimation being $O(m^{-1})$. In a pioneering article, Prasad and Rao (1990) demonstrated the importance of accounting for these two sources of underestimation, and using a Taylor linearization method produced a second-order unbiased (or nearly unbiased) MSE estimator of EBLUP when the variance components are estimated by a simple method of moments. The bias of that MSE estimator is of order $o(m^{-1})$. In other words, this is a second-order unbiased or nearly unbiased MSE estimator. Following the work of Prasad and Rao (1990), a huge volume of articles on second-order unbiased MSE estimation has been written. Here we focus our review on resampling methods. Readers interested in the Taylor linearization methods are referred to Rao (2003), and Jiang and Lahiri (2006).

An early application of the parametric bootstrap method to obtain second-order unbiased MSE estimators of EBLUP's can be found in Butar (1997). The research on parametric boostrap MSE estimation has been followed up in different directions by a number of researchers, including Booth and Hobert (1998), Butar and Lahiri (2003), Pfeffermann and Glickman (2004), Pfeffermann and Tiller (2005), Hall and Maiti (2006b), and others. A comprehensive theory of second-order MSE estimation using jackknife method was put forth by Jiang et al. (2002), although similar jackknife methods were explored in special cases by Lahiri (1995) and Chattopadhyay et al. (1999). For alternative jackknife methods, see Rao (2003) and Lohr and Rao (2003). As noted by Bell, W. (2001), the jackknife estimator of Jiang et al. (2002) could take negative values. Chen and Lahiri (2003), however, noticed that this is not a severe problem and can be easily rectified. They also considered a weighted jackknife version of the Jiang–Lahiri–Wan jackknife, which improved the efficiency in certain situations. In the context of the normal linear mixed model, Chen and Lahiri (2008) proposed a linearized weighted jackknife, which cuts down the computations and performed well compared to the corresponding jackknife of Jiang et al. (2002) in their simulation. But, we note that the development of the Jiang-Lahiri-Wan jackknife method does not require any specific distributional assumptions other than the ones necessary to obtain the BP. In simulations, jackknife method has been found to be very robust in comparison to other MSE estimators (see, e.g., Fabrizi et al., 2007). There has been some effort to robustify the bootstrap method as well. See, for example, Pfeffermann and Glickman (2004) and Hall and Maiti (2006a).

We now explain some of the resampling methods for the Fay–Herriot model since this model has been studied extensively in the small area estimation literature. Lahiri (2003b) considered a comparison of different measures of uncertainty under this model. By the Kackar–Harville identity (Kackar and Harville, 1984), we have

$$\mathrm{MSE}\left\{\hat{\theta}_i(y_i;\hat{A})\right\} = g_{1i}(A) + g_{2i}(A) + G_{3i}(A), \tag{21}$$

where

$$g_{1i}(A) = \frac{AD_i}{A+D_i},$$

$$g_{2i}(A) = \frac{D_i^2}{(A+D_i)^2} x_i' \left(\sum_{j=1}^{m} \frac{1}{A+D_j} x_j x_j'\right)^{-1} x_i,$$

$$G_{3i}(A) = \mathrm{E}\left\{\hat{\theta}_i(y_i;\hat{A}) - \hat{\theta}_i(y_i;A)\right\}^2,$$

where $g_{1i}(A)+g_{2i}(A)$ is the MSE of the BLUP and $G_{3i}(A)$ is the additional uncertainty due to the estimation of the variance component A.

A naïve MSE estimator is obtained by estimating the MSE of BLUP and is given by:

$$\mathrm{mse}_i^{\mathrm{N}} = g_{1i}(\hat{A}) + g_{2i}(\hat{A}).$$

Intuitively, this naïve MSE estimator is likely to underestimate the true MSE because it fails to incorporate the additional uncertainty due to the estimation of A. In fact, Prasad and Rao (1990) showed that the order of this underestimation is $O(m^{-1})$ under certain regularity conditions. Interestingly, the naïve MSE estimator even underestimates the true MSE of the BLUP, the order of underestimation being $O(m^{-1})$.

Prasad and Rao (1990) proposed the following MSE estimator when A is estimated by the usual method of moments:

$$\mathrm{mse}_i^{\mathrm{PR}} = g_{1i}(\hat{A}) + g_{2i}(\hat{A}) + 2g_{3i}(\hat{A}),$$

where $g_{3i}(\hat{A}) = \left\{2D_i^2/m^2(\hat{A}+D_i)^3\right\} \sum_{j=1}^{m}(\hat{A}+D_j)^2$. Under the same regularity conditions, the bias of $\mathrm{mse}_i^{\mathrm{PR}}$ is of the order $o(m^{-1})$.

For the Fay–Herriot model, the jackknife MSE estimator, proposed by Jiang et al. (2002), reduces to:

$$\mathrm{mse}_i^{\mathrm{JLW}} = g_{1i}(\hat{A}) - \frac{m-1}{m}\sum_{u=1}^{m}\left\{g_{1i}(\hat{A}_{-u}) - g_{1i}(\hat{A})\right\} + \frac{m-1}{m}\sum_{u=1}^{m}\left\{\tilde{\theta}_i(y_i;\hat{A}_{-u}) - \hat{\theta}_i(y_i;\hat{A})\right\}^2,$$

where $\hat{A}_{-u}$ is obtained from $\hat{A}$ after deleting the uth small area data and

$$\tilde{\theta}_i(y_i;\hat{A}_{-u}) = \frac{D_i}{\hat{A}_{-u}+D_i} x_i'\hat{\beta}_{-u} + \frac{\hat{A}_{-u}}{\hat{A}_{-u}+D_i} y_i,$$

$$\hat{\beta}_{-u} = \left(\sum_{j\neq u} \frac{D_j}{\hat{A}_{-u}+D_j} x_j x_j'\right)^{-1} \sum_{j\neq u} \frac{D_j}{\hat{A}_{-u}+D_j} x_j y_j.$$

For the Fay–Herriot case, the weighted jackknife MSE estimator suggested by Chen and Lahiri (2003) is given by:

$$\begin{aligned}\text{mse}_i^{\text{CL}} &= g_{1i}(\hat{A}) + g_{2i}(\hat{A}) \\ &\quad - \sum_{u=1}^{m} w_u \left[g_{1i}(\hat{A}_{-u}) + g_{2i}(\hat{A}_{-u}) - \{g_{1i}(\hat{A}) + g_{2i}(\hat{A})\} \right] \\ &\quad + \sum_{u=1}^{m} w_u \{\hat{\theta}_i(y_i; \hat{A}_{-u}) - \hat{\theta}_i(y_i; \hat{A})\}^2.\end{aligned}$$

Chen and Lahiri (2003) suggested two choices of $w_u = (m-1)/m$ and $w_u = x_u' \left(\sum_{j=1}^{m} x_j x_j' \right) x_u$. Note that mse_i^{CL} is different from $\text{mse}_i^{\text{JLW}}$ in two respects. First, Chen and Lahiri (2003) used more exact calculations by exploiting the Kackar–Harville identity, which is valid for the normality assumption. Second, the method also adjusts the $g_{2i}(\hat{A})$ term for bias. Although in the standard second-order asymptotic sense this adjustment is not needed, we may not ignore this bias correction when the relative contribution from $g_{2i}(A)$ is significant.

Butar and Lahiri (2003) proposed the following parametric bootstrap MSE estimator:

$$\begin{aligned}\text{mse}_i^{\text{BL}} &= g_{1i}(\hat{A}) + g_{2i}(\hat{A}) - E_\star[g_{1i}(\hat{A}^\star) + g_{2i}(\hat{A}^\star) - \{g_{1i}(\hat{A}) + g_{2i}(\hat{A})\}] \\ &\quad + \text{E}_\star\{\hat{\theta}_i(y_i; \hat{A}^\star) - \hat{\theta}_i(y_i; \hat{A})\}^2.\end{aligned}$$

Similar parametric bootstrap methods can be found in Butar (1997) and Pfeffermann and Glickman (2004). In the aforementioned discussion, $\text{E}_\star$ is the bootstrap expectation, that is, the expectation with respect to the Fay–Herriot model with β and A replaced by $\hat{\beta}$ and $\hat{A}$, respectively. We obtain $\hat{A}^\star$ using the formula for $\hat{A}$ with the original sample replaced by the bootstrap sample. In practice, Monte Carlo methods are employed to approximate the bootstrap expectations. Note that because of the bias correction term the jackknife and parametric bootstrap methods could produce negative estimates. This was first observed by Bell, W. (2001) in the context of jackknife MSE estimator. But this can be easily corrected as noted by Chen and Lahiri (2003) who recommended the following MSE estimator in case mse_i^{CL} yields a negative value:

$$\text{mse}_i^{\text{ACL}} = g_{1i}(\hat{A}) + g_{2i}(\hat{A}) + \frac{D_i^2}{(\hat{A} + D_i)^3} v_{\text{WJ}} + \frac{D_i^2}{(\hat{A} + D_i)^4} r_i^2 v_{\text{WJ}}.$$

where $r_i = y_i - x_i' \hat{\beta}$, the residual for the ith area, and $v_{\text{WJ}} = \sum_{u=1}^{m} w_u (\hat{A}_{-u} - \hat{A})^2$, a weighted jackknife variance estimator of $\hat{A}$.

Alternatively, Chen and Lahiri (2008) suggested the linearized weighted jackknife $\text{mse}_i^{\text{ACL}}$ as a new MSE estimator. Similar corrections can be made to $\text{mse}_i^{\text{JLW}}$ and mse_i^{BL}. In the parametric bootstrap context, different nonnegative second-order unbiased MSE estimators were considered by Hall and Maiti (2006b) and Chatterjee and Lahiri (2007). All the MSE estimators, except the naïve one, has the same second-order bias property under the same regularity conditions. However, in different simulation studies, the performances of different MSE estimators varied depending on several factors such as the value of m, D_i/A, variations in D_i/A, presence of an outlier in D_i/A, leverage $h_i = x_i^T (X^T X)^{-1} x_i$, and the method of estimation of A.

For a special case of the Fay–Herriot model when $D_i = D$ and $x_i' \beta = \mu (i = 1, \ldots, m)$, Butar and Lahiri (2003) showed that their parametric bootstrap MSE estimator

is identical to a measure of uncertainty proposed by Morris (1983b) up to the order $O(m^{-1})$ if an unbiased estimator of $B = D/(A + D)$ is chosen in the EBLUP formula. This is also true for the Chen–Lahiri jackknife MSE estimator. Thus, the parametric bootstrap and the Chen–Lahiri jackknife MSE estimators are close to a Bayesian solution since Morris (1983b) obtained his uncertainty measure by approximating the posterior variance using flat uniform priors on μ and B. For the aforementioned model and the standard untruncated unbiased quadratic estimator of A, Lahiri (1995) appoximated $\text{mse}_i^{\text{JLW}}$ up to the order $O_{\text{P}}(m^{-1})$ and obtained the following result:

$$\text{mse}_i^{\text{JLW}} \doteq g_1(\hat{A}) + g_2(\hat{A}) + \frac{D^2}{m(\hat{A}+D)}(b_2 - 1) + \frac{D^2}{m(\hat{A}+D)^2}(b_2 - 1)r_i^2 - \frac{2D^2}{m(\hat{A}+D)^{3/2}}\sqrt{b_1}r_i,$$

where $b_1 = m_3^2/(\hat{A}+D)^3$, $b_2 = m_4/(\hat{A}+D)^2$, and $r_i = y_i - \overline{y}$. Here b_1 and b_2 can be viewed as estimated skewness and kurtosis for the marginal distribution of y_i. Under normality, $b_1 \approx 0$ and $b_2 \approx 3$ and so in this case $\text{mse}_i^{\text{JLW}}$ reduces to

$$\text{mse}_i^{\text{JLW}} \doteq g_1(\hat{A}) + g_2(\hat{A}) + \frac{2D^2}{m(\hat{A}+D)} + \frac{2D^2}{m(\hat{A}+D)^2}r_i^2.$$

It is reasonable to expect that, in this case, $\text{mse}_i^{\text{JLW}}$ is identical to mse_i^{CL} and mse_i^{BL} correct up to the order $O_{\text{P}}(m^{-1})$.

Chatterjee and Lahiri (2007) obtained second-order unbiased, nonnegative, conditional, and unconditional MSE estimators, and allowed the hyperparameter dimension to grow with the sample size, thus bringing in the dimension asymptotic effect of estimating the hyperparameters. Their method retains the basic simplicity of the bootstrap methodology in which laborious analytical calculations are replaced by computer-oriented simple techniques. The Chatterjee–Lahiri parametric bootstrap methods use a double bootstrap strategy as in Booth and Hobert (1998) and Hall and Maiti (2006b). However, apart from being applicable to a much broader collection of problems, the Chatterjee–Lahiri one-step (conditional or unconditional) MSE estimator methodology is not driven by stepwise calibration ideas. In each scenario, the resampling technique and the MSE estimate formula are exactly the same for all situations; the method does not require problem specific corrections.

Because of the excellent prospect of wide applicability of the Chatterjee–Lahiri parametric bootstrap method, we describe the method in detail. Let y_i be the $n_i \times 1$ vector of observations for the i^{th} area ($i = 1, \ldots, m$). The dimension of y_i can be arbitrary, and may or may not depend on m and i. Consider the following two-level model:

$$\begin{aligned} &\text{Level 1:} \quad y_i|\theta_i \sim f_i(\cdot;\theta_i,\xi), \quad i = 1,\ldots,m; \\ &\text{Level 2:} \quad \theta_i \sim g_i(\cdot;\xi), \quad i = 1,\ldots,m, \end{aligned}$$

where $\xi \in \Xi \subseteq R^d$ and the parameter space Ξ is an open set in R^d. The dimension of θ_i is arbitrary and can also depend on m and i. The pairs $\{(y_i, \theta_i),\ i = 1, \ldots, m\}$ are assumed to be independent. However, because there is no restriction on the dimension of y_i, the independence of the observations across different areas is not a strong assumption in most small area applications, and is a matter of nomenclature. As explained by Chatterjee and Lahiri (2007), the aforementioned model covers a wide variety of small area models.

The conditional mean squared prediction error (CMSE) of $\theta_i(y_i, \hat{\xi})$, an EBP of θ_i, is defined as the conditional expectation

$$\text{CMSE}(y_i, \xi) = E[\{\theta_i - \theta_i(y_i, \hat{\xi})\}^2 \mid y_i].$$

The unconditional mean squared prediction error (MSE) of $\theta_i(y_i, \hat{\xi})$ is defined as

$$\text{MSE}(\xi) = E[\theta_i - \theta_i(y_i, \hat{\xi})]^2,$$

where E denotes expectation with respect to the joint distribution of $y = (y_1, \ldots, y_m)$ and $\theta = (\theta_1, \ldots, \theta_m)$.

A two-level parametric bootstrap algorithm for generating resamples is given below:

1. Resample $y^* = \left(y_1^*, \ldots, y_m^*\right)$ using the following two-level model:

   $$\text{Level } 1^*: \; y_i^* \mid \theta_i^* \sim f_i\left(\cdot; \theta_i^*, \hat{\xi}\right);$$

   $$\text{Level } 2^*: \; \theta_i^* \sim g_i(\cdot; \hat{\xi}),$$

   $i = 1, \ldots, m$. The expectation at this step, which is conditional on y, is denoted by E^*.
2. Obtain $\hat{\xi}^* = \hat{\xi}(y^*)$, the estimator of ξ based on the resample y^*, using the same technique used to obtain $\hat{\xi}(y)$.
3. Resample $y^{**} = \left(y_1^{**}, \ldots, y_m^{**}\right)$ from y^* using the following two-level model:

   $$\text{Level } 1^{**}: \; y_i^{**} \mid \theta_i^{**} \sim f_i\left(\cdot; \theta_i^{**}, \hat{\xi}^*\right);$$

   $$\text{Level } 2^{**}: \; \theta_i^{**} \sim g_i\left(\cdot; \hat{\xi}^*\right),$$

   $i = 1, \ldots, m$. The expectation at this step, conditional on y and y^*, is denoted by E^{**}.
4. Define

   $$M_{1a} = E^*\left\{\theta_i^* - \theta_i\left(y_i, \hat{\xi}^*\right)\right\}^2,$$

   $$M_{2a} = E^* E^{**}\left\{\theta_i^{**} - \theta_i\left(y_i, \hat{\xi}^{**}\right)\right\}^2,$$

   $$M_{3a} = E^*\left\{\theta_i^* - \theta_i\left(y_i^*, \hat{\xi}^*\right)\right\}^2,$$

   $$M_{4a} = E^* E^{**}\left\{\theta_i^{**} - \theta_i\left(y_i^{**}, \hat{\xi}^{**}\right)\right\}^2,$$

$\hat{\xi}^{**} = \hat{\xi}(y^{**})$. The conditional and unconditional MSE estimators are given by:

$$\widehat{\text{MSE}}_a = H(M_{1a}, M_{2a} - M_{1a}),$$
$$\widetilde{\text{MSE}}_a = H(M_{3a}, M_{4a} - M_{3a}),$$

respectively. Chatterjee and Lahiri (2007) considered the following choice of $H(.)$:

$$H(x, b) = 2x/(1 + \exp\{2b/x\}).$$

Chatterjee and Lahiri (2007) examined the higher order asymptotic properties of both unconditional and conditional MSE estimators. They also studied the small sample properties of these MSE estimators by Monte Carlo simulations.

Let us now turn our attention to the prediction interval problem. Again, for illustration, we shall restrict ourselves to the Fay–Herriot model. Cox (1975) initiated the idea of developing the empirical Bayes confidence intervals. In the current context, his suggestion generates the following prediction interval:

$$I_i^C(\alpha) : \hat{\theta}_i(y_i; \hat{A}) \pm z_{\alpha/2}\sqrt{g_{1i}(\hat{A})}.$$

When m is large, under certain regularity conditions, $P(\theta_i \in I_i^C(\alpha)) = 1 - \alpha + O(m^{-1})$. Thus, this prediction interval attains the desired coverage probability asymptotically, but the coverage error is of order $O(m^{-1})$, not accurate enough in most small area applications. Intuitively, this could be due to the fact that the construction of the prediction interval does not take into account the additional errors incurred by the estimation of model parameters. See Jiang and Lahiri (2006) and Chatterjee et al. (2008) for a review of different methods for improving the coverage errors. We now review the parametric bootstrap method for constructing the prediction interval for θ_i.

The utility of parametric bootstrap prediction interval in small area estimation was first recognized by Chatterjee and Lahiri (2002) [also see Lahiri (2003a)] who developed the method for the Fay–Herriot model. The method was later extended to a general linear mixed model by Chatterjee et al. (2008). Their $100(1 - \alpha)\%$ prediction interval is given by:

$$I_i^{PB}(\alpha) : \hat{\theta}_i(y_i; \hat{A}) - b_{1i}\sqrt{g_{1i}(\hat{A})},\ \hat{\theta}_i(y_i; \hat{A}) + b_{2i}\sqrt{g_{1i}(\hat{A})},$$

where b_{1i} and b_{2i} are such that

$$P^*\left[\theta_i^* < \hat{\theta}_i(y_i^*; \hat{A}^\star) - b_{1i}\sqrt{g_{1i}(\hat{A}^\star)}\right] = \alpha_1$$

$$P^*\left[\theta_i^* > \hat{\theta}_i(y_i^*; \hat{A}^\star) + b_{2i}\sqrt{g_{1i}(\hat{A}^\star)}\right] = \alpha_2,$$

and $\alpha_1 + \alpha_2 = \alpha$. The method can provide both equal-tailed and the shortest length small area accurate prediction intervals.

Chatterjee et al. (2008) showed that, under certain mild regularity conditions,

$$P\left[\theta_i \in I_i^{\mathrm{PB}}(\alpha)\right] = 1 - \alpha + O(d^3 m^{-1.5}),$$

where $d = p + 1$ and p is the dimension of β. An alternative parametric bootstrap method was proposed by Hall and Maiti (2006b). As noted by Rao (2005b), the method proposed by Hall and Maiti (2006b), unlike Chatterjee and Lahiri (2002) or Chatterjee et al. (2008), does not utilize area specific data. For the prediction interval method of Chatterjee et al. (2008), it is important to use strictly positive estimates of the variance components. However, the standard variance component estimators such as method of moments, maximum likelihood, and residual likelihood methods are all subject to zero estimates for the variance components. A naïve solution is to take a small positive value of the variance component when the estimate is zero. Li (2007) noticed the importance of this truncation point in the performance of the prediction interval for the Fay–Herriot model and developed certain adjusted density maximization method, which improved the performance of the parametric bootstrap prediction interval to a great extent.

8. Discussion

We have attempted to cover various applications of resampling methods in sample surveys from both the design-based and model-based perspectives. Our emphasis on design-based variance estimation, prediction variance estimation, and mean square error estimation of EBP for small area estimation reflects the importance of these topics given in the survey literature. Certainly, resampling methods could also be used for more complex inferential problems such as the confidence interval problem which we reviewed to a much lesser extent. The development of the Sections 2–5 follows the traditional design-based approach in that the design-based variances are of interest and properties of the variance estimators are studied under the design-based framework, assuming the values of the study variable for the finite population to be nonstochastic. In these sections, asymptotics are the usual design-based asymptotics; for details on such asymptotics, we refer to Shao (1996) and Chapter 40.

In theory, the Taylor linearization and resampling methods are similar for large samples for full response. To compare different plans, Rao and Wu (1985) obtained second-order expansions of the different variance estimators. They found that different variants of the jackknife variance estimators are equivalent up to the second-order. The jackknife and a version of BHS are in general first-order equivalent and in the special case of two PSU's per stratum, the jackknife estimator is identical to the Taylor linearization variance estimator up to the second-order. These results suggest that the choice between the two estimators should depend more on the operational rather than statistical considerations. The empirical findings of Kovar et al. (1988) were in agreement with the analytical results of Rao and Wu (1985). For a detailed account on the asymptotic set-up and a review of asymptotic comparisons of different resampling methods, see Shao (1996).

Using data from the Current Population Survey conducted by the U.S. Census Bureau, Kish and Frankel (1974) described an empirical evaluation study that examines the relative performances of the Taylor linearization, BRR, and jackknife variance estimation methods. Each of the three methods was used to estimate variances of ratio means, simple correlations, and multiple regression coefficients. For partial and multiple correlation coefficients, only the two replication methods were applied. Relative biases, mean squared errors, and coverage properties were used as criteria for the evaluation. The study showed that the methods were equally good when used to estimate variances of ratio means, coefficients of regression, and of simple or partial correlation coefficients, whereas the results for coefficients of multiple correlation were poor on all criteria. The coverage error was reported to be the smallest for the BRR method, and the jackknife performed better than the linearization estimator; these results were especially noticeable for simple and partial correlation coefficients and could be associated with the negative relative biases of the mean squared errors of the jackknife and linearization methods. On the other hand, the variability was consistently the lowest for the linearization method and the highest for the BRR. Kish and Frankel (1974) concluded that none of the three methods was consistently better than the others; therefore, the choice of the method can be based on such practical criteria as relative cost and simplicity.

Using the UK Labor Force survey data, Canty and Davison (1999) conducted a Monte Carlo simulation study to demonstrate the usefulness of resampling-based methods in capturing variability due to calibration, in addition to the usual sampling variability. Their results show the following: (i) the traditional Taylor linearization design-based

method could severely underestimate the true variability; (ii) bootstrap and jackknife linearization methods provide more reliable standard estimates than do the jackknife and BRR; (iii) although the linearized jackknife has a slight edge over the bootstrap in terms of computational burden and reliability, in practice bootstrap may be preferred because it avoids all the analytical work that goes with jackknife linearization. Using a simulation study, Valliant (2004) demonstrated the utility of a suitable jackknife method in capturing the variability incurred due to calibration performed at different stages in the standard survey operation. Asymptotic theory for jackknife is mostly available for the basic design weights. One exception is the paper by Yung and Rao (1996) who showed that $v_{\mathrm{JK}}(\hat{Y}_{\mathrm{GREG}})$ is asymptotically equivalent to $\mathrm{var}(\hat{Y}_{\mathrm{GREG}})$.

In the small area estimation framework, the Taylor series, jackknife, and parameteric bootstrap methods for unconditional mean squared error estimation all enjoy the same second-order unbiasedness property. As explained by Chatterjee et al. (2008), it is difficult, if not impossible, to produce a purely nonparametric bootstrap satisfying the second-order unbiasedness property. This is primarily because of the difficulty in producing a consistent estimator of the conditional distribution function of the random effects given the data. Because of the scarcity of data at the small area level, the importance of parametric bootstrap in small area estimation problems cannot be overemphasized. Some versions of a semiparametric bootstrap were proposed by Pfeffermann and Glickman (2004) and Hall and Maiti (2006a). The stability of MSE estimators, in terms of the MSE of MSE estimators, has not been studied analytically for the Taylor series and the jackknife methods, although Chatterjee and Lahiri (2007) provided analytical results for the parametric bootstrap for a very general case.

Most of the simulation results are available for different particular cases of the Fay–Herriot model and the nested error model and often times the studies examine the performances of different MSE estimators in estimating the unconditional MSE, although there is a growing interest in evaluating different MSE estimators in estimating conditional MSEs. For the Fay–Herriot model, there is no conclusive evidence about the uniform superiority of one MSE estimator over another in estimating the unconditional MSE. The results are mixed and depend on various factors, including the relative magnitude of model variance to the sampling variances, magnitude and variability of the sampling variances, number of small areas, and the distribution of the random effects. The Prasad–Rao Taylor linearization estimator has a tendency of overestimating the unconditional MSE and sometimes the overestimation can be severe. The resampling methods, particularly the parametric bootstrap, have exhibited relatively robust results in terms of relative bias against variations of different factors, compared to the Taylor series method, at the cost of increasing the variance.

For the nested error model, Hall and Maiti (2006a) showed better performances of their nonparametric bootstrap over the naive estimator in terms of relative bias. They compared their method with the jackknife method of Jiang et al. (2002) in a simulation study, which showed that the jackknife method performed better in terms of the relative bias but inferior in terms of the coefficient of variation, relative to their nonparametric bootstrap method. Under the same simulation set-up, the simulation results of Ganesh (2007) and Tang (2008) show better performance of the normality-based Taylor series method compared to the nonparametric bootstrap results reported in Hall and Maiti (2006a), whereas Tang (2008) showed similar performance of the Taylor series method compared to the jackknife method of Jiang et al. (2002). The simulation results of

Pfeffermann and Glickman (2004) show good performance of parametric bootstrap and linearized jackknife method of Chen and Lahiri (2008). Lahiri and Rao (1995) showed the insensitivity of the normality-based Prasad–Rao Taylor series method against the variation of the distribution of the random effects when the variance components are estimated by the method of moments. However, this result does not extend to situations where sampling distribution is nonnormal or the variance components are estimated by the Fay–Herriot method of moments. Ganesh (2007) showed that the normality-based Prasad–Rao MSE estimator is not second-order unbiased against the violation of the normality assumption even when the sampling distribution is normal and the variance components are estimated by the method of moments. The parametric bootstrap of Chatterjee and Lahiri (2007) designed to estimate the conditional MSE performed well in their simulation study, compared to the other rival MSE estimators, in terms of bias even to estimate the unconditional MSE.

Results from different simulation studies have been documented in Prasad and Rao (1990), Lahiri and Rao (1995), Datta and Lahiri (2000), Jiang et al. (2002), Chen and Lahiri (2003), Hall and Maiti (2006a,b), Chatterjee and Lahiri (2007), Chen and Lahiri (2008), Li (2007), Tang (2008). Fabrizi et al. (2007) evaluated different MSE estimators using simulations with data from the European Social Survey. They concluded the jackknife method due to Jiang et al. (2002) performed better than the Taylor linearization method. Molina et al. (2007) conducted a simulation study using UK Labor force data to compare the Taylor linearization method and parametric bootstrap and concluded that parametric bootstrap outperformed the Taylor linearization method.

The literature on resampling methods in surveys is huge and is steadily growing. Hence, it is almost impossible to give a comprehensive review of the topic in such a limited space. For further readings on resampling methods in surveys, we refer the readers to the textbook by Wolter (1985) and excellent review papers by Rust (1985), Rust and Rao (1996), Shao (1996).

Acknowledgments

The authors would like to thank Professor S. Chatterjee for his comments on the section about small area estimation. Jiming Jiang's research is partially supported by NSF grants DMS - 0203676 and DMS - 0402824. Any opinions expressed in this chapter are those of the authors and do not constitute policy of the U.S. Bureau of Labor Statistics.

Sample Surveys: Inference and Analysis, Vol. 29B
ISSN: 0169-7161
© 2009 Published by Elsevier B.V.
DOI: 10.1016/S0169-7161(09)00229-6

Bayesian Developments in Survey Sampling

Malay Ghosh

1. Introduction

Sample surveys are widely used to gather information about various characteristics of a finite population such as the total or the mean of a response variable, or some other parameter of interest. One of the fundamental inference problems in survey sampling is to obtain estimate the total or the mean, and find also the associated measure of precision such as the variance, or more generally the mean squared error.

The classical approach towards this inferential problem is design-based, which includes the selection probabilities of the different sampling units. In contrast, there is a model-based approach that views the finite population as a sample from a hypothetical superpopulation, and inference for finite poulation parameters are model-based. On occasions, people have recommended methods that are hybrid of the two, that is, model-assisted design-based estimates or design-assisted model-based estimates (e.g., Prasad and Rao, 1999).

Both design- and model-based approaches can be frequentist, where such procedures do not make an explicit use of priors either for the finite population or the superpopulation parameters. In contrast, the Bayesian approach assumes that the response variable associated with any unit is the realization of a random variable following some specified distribution based on prior information.

Prior information always exists in survey sampling in the form of auxiliary variables, often found through administrative records. The Bayesian approach utilizes this auxiliary information explicitly through prior distributions for finite population parameters, distributions which relate these parameters and the auxiliary variables.

Little (2004), in a very elegant review article, has pointed out a sevenfold advantage of superpopulation or Bayesian models. First, such models enable one to integrate inference in finite population sampling with the mainstream statistics inference. Second, with noninformative priors, it is possible to match the design-based inference with model-based inference. Third, the Bayesian method is particularly well-suited to handle complex sampling designs such as those involving stratification and clustering. Fourth, often Bayesian methods provide better inferential procedures than their frequentist counterparts, especially for small samples. The reason behind is that though classical frequentist procedures often rely heavily on asymptotics requiring a very large sample size for its success, a Bayesian approach with an appropriate prior can yield

meaningful inference even in such situations. Fifth, because the Bayesian methods use two sources of information, namely, the likelihood and the prior, the resulting inference often provides greater precision than one which uses only one or the other. Sixth, the Bayesian procedures, unlike most frequentist procedures, do not violate the likelihood principle. Finally, even from asymptotic considerations, Bayesian methods enjoy the same efficiency as the maximum likelihood method.

Superpopulation modeling has been around for a long time in the survey sampling literature. Among others, we may refer to Royall (1970b), Thompson (1997), and Valliant, et al. (2000) (See the review article by Valliant in Chapter 23 of this book). The frequentist approach usually relies on estimation of superpopulation parameters. The Bayesian approach, on the other hand, assigns prior distributions to these parameters.

The objective of this chapter is to provide a review of Bayesian developments in survey sampling. In Section 2, we introduce the basic notation and a brief description of the "sufficiency principle" and the "likelihood principle" in the context of finite population sampling. Many view these principles as forming the cornerstone of any statistical inference. In finite population sampling, though the Bayesian paradigm obeys these principles, a design-based approach often violates the same. As a result, many find the Bayesian approach as an attractive alternative.

The Bayesian paradigm is introduced in Section 3 of this chapter, and is illustrated with the estimation of the finite population mean when a simple exchangibility assumption holds among the units in the population. Section 4 introduces linear Bayes estimators for estimation of the finite population mean. Section 5 addresses the same estimation problem for more complex models. Section 6 contains estimation of strata means and their application in domain estimation. Hierarchical Bayesian estimation for generalized linear models is discussed in Section 7. Some final remarks are made in Section 8.

2. Notation and preliminaries

In finite population sampling we deal with a population of N units labeled $1, 2, \ldots, N$. These units, for example, may be households in a city, or farms in a county, or schools in a certain school district, etc. We denote the finite population by $\mathcal{U}$ and assume that the population size N is known. Let y_i denote the unknown value of some characteristic of interest for unit i, $i = 1, \ldots, N$. For simplicity, we consider the y_i to be scalar, although they can be vector-valued as well. Often, in addition, a vector of auxiliary characteristics for unit i, say, $\boldsymbol{x}_i$ is also available. The components of $\boldsymbol{x}_i$, and their possible relationship to the y_i, summarize the prior information about the population. In finite population sampling, $\boldsymbol{y} = (y_1, \ldots, y_N)^T$ is regarded as an unknown parameter belonging to $\mathcal{Y}$, a suitable subset of $\mathcal{R}^N$, the N-dimensional Euclidean space.

To infer about $\boldsymbol{y}$, or some suitable function of it, we need to select a sample from the population $\mathcal{U}$. We will assume that the value of y on a sampled unit is known, that is, we will not consider here presence of any nonresponse, response bias, or measurement error. We closely follow Ghosh and Meeden (1997, Chapter 1) to define our notation later. We select a sample s which is a subset of $\mathcal{U}$. Let $n(s)$ denote the number of elements in s which consists of the units $i_1, \ldots, i_{n(s)}$, that is $s = \{i_1, \ldots, i_{n(s)}\}$. Let $\mathcal{S}$ denote the

(countable) set of all possible samples. A sampling design is given by a probability function $p(\cdot)$ defined on $\mathcal{S}$, that is, $p(s) \geq 0$ for all $s \in \mathcal{S}$ and $\sum_{s\in\mathcal{S}} p(s) = 1$. We assume sampling without replacement so that the labels of the units in a sample $s = \{i_1, \ldots, i_{n(s)}\}$ are ordered $1 \leq i_1 < \cdots < i_{n(s)} \leq N$. For a population vector $\boldsymbol{y} \in \mathcal{Y}$ and a sample s, we denote the sampled vector by $\boldsymbol{y}(s) = (y_{i_1}, \ldots, y_{i_{n(s)}})^T$.

Given a parameter space $\mathcal{Y}$ and design p, a typical sample point is the set of labels of the units in the selected sample along with their values of the characteristic of interest. For a sample $s = \{i_1, \ldots, i_{n(s)}\}$, we denote the data point by $z = (s, \boldsymbol{z}_s)$ where $\boldsymbol{z}_s$ is $\boldsymbol{y}(s)$, the vector of values of the characteristic of the units in s. For simplicity, we denote $\boldsymbol{z}_s$ by $(z_1, \ldots, z_{n(s)})^T$. For the remainder of this section, we closely follow Meeden (1992).

For a given sample s, let $s^c = \mathcal{U} - s$, be the set of labels associated with the unsampled units. The main objective of finite population sampling is inference about $\boldsymbol{y}(s^c)$ given z where $\boldsymbol{y}(s^c) = (y_{j_1}, \ldots, y_{j_{N-n(s)}})^T$ and $j_1 < \cdots < j_{N-n(s)}$ are such that $j_k \in s^c$, $k = 1, \ldots, N-n(s)$. The likelihood principle (see, e.g., Ghosh and Meeden, 1997, p. 7) says that in finite population sampling given the observed data z, one just learns the values of $\boldsymbol{z}_s$, and that $\boldsymbol{y}(s^c)$ must come from a $\boldsymbol{y}$ which is consistent with $\boldsymbol{z}_s$. Note that for a given design p the sample space is given by

$$Z(\mathcal{Y}, p) \equiv Z = \{(s, \boldsymbol{z}_s) : p(s) > 0 \text{ and } \boldsymbol{z}_s = \boldsymbol{y}(s) \text{ for some } \boldsymbol{y} \in \mathcal{Y}\}.$$

So for a fixed $\boldsymbol{y} \in \mathcal{Y}$ the probability function over Z is given by

$$\begin{aligned} P_{\boldsymbol{y}}(z) = P_{\boldsymbol{y}}(s, \boldsymbol{z}_s) &= p(s) \quad \text{if } \boldsymbol{z}_s = \boldsymbol{y}(s) \\ &= 0 \qquad \text{otherwise.} \end{aligned} \tag{1}$$

Define a subset of the parameter space $\mathcal{Y}$ by

$$\begin{aligned} \mathcal{Y}_z &= \{\boldsymbol{y} | P_{\boldsymbol{y}}(z) > 0\} \\ &= \{\boldsymbol{y} | \boldsymbol{y}(s) = \boldsymbol{z}_s\}, \end{aligned}$$

$\mathcal{Y}_z$ is determined by the sample z. The likelihood function $L_z(\boldsymbol{y})$ for $\boldsymbol{y}$ based on the data z is given by $L_z(\boldsymbol{y}) = P_{\boldsymbol{y}}(z)$. From (1) it follows that

$$L_z(\boldsymbol{y}) = \begin{cases} p(s) & \text{if } \boldsymbol{y} \in \mathcal{Y}_z \\ 0 & \text{elsewhere.} \end{cases} \tag{2}$$

The standardized likelihood function defined as $\bar{L}_z(\boldsymbol{y}) = L_z(\boldsymbol{y}) / \sup_{\boldsymbol{y}} L_z(\boldsymbol{y})$ is given by

$$\bar{L}_z(\boldsymbol{y}) = \begin{cases} 1 & \text{if } \boldsymbol{y} \in \mathcal{Y}_z \\ 0 & \text{elsewhere.} \end{cases} \tag{3}$$

Clearly, the standardized likelihood function $\bar{L}_z(\boldsymbol{y})$ is independent of the design probability $p(s)$.

The sufficiency and likelihood principles are two widely accepted principles in statistical inference. The sufficiency principle states that if two samples z and z' lead to the same value of a sufficient statistic, then the inference about $\boldsymbol{y}$ should be the same whether one obtains the sample z or z'. This principle does not say anything about the nature of the information or how to measure information supplied by the sample z.

It is the likelihood principle that states that the information supplied by z is measured by the standardized likelihood function $\bar{L}_z(\boldsymbol{y})$. Meeden (1992) noted that the mapping $z \to \bar{L}_z(\cdot)$ induces a minimal sufficient partition of Z.

Because $\bar{L}_z(\boldsymbol{y})$ is constant over $\mathcal{Y}_z$, maximum likelihood method is of no use to estimate $\boldsymbol{y}$. All we learn from the observed data $z = (s, \boldsymbol{z}_s)$ is that the true $\boldsymbol{y}$ must have no conflict with the observed data, that is, we must have $\boldsymbol{y}(s) = \boldsymbol{z}_s$.

The main objective of finite population sampling is drawing inference about $\boldsymbol{y}(s^c)$ given z. The two principles aforementioned state that given the observed data z one just learns about the values of $\boldsymbol{z}_s$ and that $\boldsymbol{y}(s^c)$ must come from a $\boldsymbol{y}$ which is consistent with $\boldsymbol{z}_s$. To be specific, for example, if our interest is the finite population total $\gamma(\boldsymbol{y}) = \sum_{i=1}^{N} y_i$, then we write $\gamma(\boldsymbol{y})$ as $\sum_{i\in s} z_i + \sum_{j\in s^c} y_j$ and it is enough to draw inference for the total of the unsampled (or unseen) units based on the sampled (or seen) units in s.

Obviously, one gains nothing about $\boldsymbol{y}(s^c)$ from $\boldsymbol{z}_s$ alone without further assumption relating these vectors. In the Bayesian approach to this problem, a statistician relates $\boldsymbol{y}(s^c)$ to $\boldsymbol{z}_s$ using a suitable prior distribution $\pi(\boldsymbol{y})$ on $\boldsymbol{y}$ and draws inference based on the posterior distribution of $\boldsymbol{y}(s^c)$ given the data $\boldsymbol{z}_s$. On the other hand, a frequentist achieves this by using the design p along with the unbiasedness requirement.

3. The Bayesian paradigm

As we have discussed in the last section, the main objective of finite population sampling is to infer about the unobserved units in the population given the observed sampled data. The Bayesian paradigm is particularly attractive to meet this goal. Let $\pi(\boldsymbol{y})$ denote the prior density or probability function of a Bayesian statistician to summarize prior beliefs about $\boldsymbol{y}$. Using (2) or (3), because the likelihood function for $\boldsymbol{y}$ is constant on $\mathcal{Y}_z$ the posterior density $\pi(\boldsymbol{y}|z)$ is given by

$$\pi(\boldsymbol{y}|z) = \begin{cases} \dfrac{\pi(\boldsymbol{y})}{\pi_{\boldsymbol{y}(s)}(\boldsymbol{z}_s)} & \text{for } \boldsymbol{y} \in \mathcal{Y}_z, \\ 0, & \text{otherwise,} \end{cases} \tag{4}$$

where $\pi_{\boldsymbol{y}(s)}(\boldsymbol{z}_s)$ is the marginal prior density of $\boldsymbol{y}(s)$ evaluated at the observed data $\boldsymbol{z}_s$.

REMARK 3.1. From (4), it is clear that the posterior density does not depend on the design p; it depends on the sample only through $\boldsymbol{z}_s$. Hence from now on we denote the posterior density by $\pi(\boldsymbol{y}|\boldsymbol{z}_s)$, which is just the prior density $\pi(\boldsymbol{y})$ with the sampled values $\boldsymbol{z}_s$ inserted in their appropriate places and renormalized, so that $\pi(\boldsymbol{y}|\boldsymbol{z}_s)$ integrates to one over $\mathcal{Y}_z$.

We note that in contrast with the Bayesian thinking, the frequentist approach uses the design probability and unbiasedness requirement in drawing inference for $\boldsymbol{y}$ based on the observed data z. It is demonstrated by Basu (1969) [see also Eq. (3) mentioned earlier] that the likelihood principle implies that the design probability should not be considered in analyzing the data after the sample has been observed. Thus, the frequentist approach requiring design unbiasedness violates the likelihood principle. Godambe (1966), a pioneer in statistical foundation of survey sampling, also noted this phenomenon.

In view of the sufficiency principle and the likelihood principle, one certainly finds the Bayesian paradigm obeying these principles attractive. However, in high-dimensional inference problems such as in finite population sampling it is often a formidable task to specify a sensible prior distribution to perform Bayesian analysis. For such problems one cannot carry out Bayesian analysis without some simplifying model assumptions. A variety of models can be employed to handle various amounts of prior knowledge. In a pioneering article on foundation of survey sampling, Hartley and Rao (1968) discussed a Bayesian approach to survey sampling. At about the same time Ericson (1969a) proposed in an important article a subjectivist Bayesian approach to finite population sampling. Some of this work is reviewed later.

In Bayesian approach to finite population sampling the goal is to obtain the conditional distribution of $\boldsymbol{y}$ given the data $\boldsymbol{z}_s$. This is tantamount to find the conditional distribution of unobserved values $\boldsymbol{y}(s^c)$ given the sampled values $\boldsymbol{z}_s$. This is really a prediction problem when the goal is to predict the unobserved $\boldsymbol{y}(s^c)$ based on its posterior density $\pi(\boldsymbol{y}(s^c)|\boldsymbol{z}_s)$. We consider some simple Bayesian models in the following sections.

3.1. A simple exchangeable model

The most important quantity of interest in finite population sampling is usually the population total or the population mean. Because the population total $\gamma(\boldsymbol{y})$ is the sum of the observed $\boldsymbol{z}_s$ values plus the sum of the unobserved $\boldsymbol{y}(s^c)$ values, under squared error loss the Bayes estimator of $\gamma(\boldsymbol{y})$ is given by

$$E_\pi[\gamma(\boldsymbol{y})|\boldsymbol{z}_s] = \sum_{i\in s} z_i + \sum_{j\in s^c} E_\pi[y_j|\boldsymbol{z}_s], \tag{5}$$

where $E_\pi[\cdot|\boldsymbol{z}_s]$ denotes the posterior expectation.

If one naively assumes a priori that $y_1, \ldots, y_N$ are independent with a mean ϕ, then from (5)

$$E_\pi[\gamma(\boldsymbol{y})|\boldsymbol{z}_s] = \sum_{i\in s} z_i + (N-n)\phi,$$

where n is the sample size. Clearly, the above is not a very good estimate because the observed units do not carry any information about the unobserved units. To relate the unobserved units to the observed units, the strong independence assumption is replaced by the exchangeability assumption of Hartley and Rao (1968), Hill (1968), and Ericson (1969a). The exchangeability assumption for the joint prior density, described later, is the Bayesian analog of simple random sampling without replacement in frequentist approach.

Suppose θ is a real-valued parameter. Assume that (i) $y_i|\theta$, $i = 1, \ldots, N$ are i.i.d. with a probability density function $g(\cdot|\theta)$, and (ii) θ has a prior density $h(\theta)$. Then the marginal density of $\boldsymbol{y}$ is given by

$$\pi(\boldsymbol{y}) = \int \prod_{i=1}^{N} g(y_i|\theta)h(\theta)d\theta. \tag{6}$$

Some people refer to (i) as a superpopulation model and (ii) as a prior, whereas others refer to (i) and (ii) as two stages of a hierarchical prior where θ is a hyperparameter. This

distinction, though often conceptually important, is not necessary from an operational point of view. The important point is that based on the given model, the conditional or the predictive distribution of $\boldsymbol{y}(s^c)$ given $\boldsymbol{z}_s$ is

$$\begin{aligned}\pi(\boldsymbol{y}(s^c)|\boldsymbol{z}_s) &= \frac{\pi(\boldsymbol{y}(s^c), \boldsymbol{z}_s)}{\pi(\boldsymbol{z}_s)} \\ &= \frac{\int \prod_{j\in s^c} g(y_j|\theta) \prod_{j\in s} g(z_j|\theta)h(\theta)d\theta}{\int \prod_{j\in s} g(z_j|\theta)h(\theta)d\theta} \\ &= \int \prod_{j\in s^c} g(y_j|\theta)h(\theta|\boldsymbol{z}_s)d\theta, \end{aligned} \tag{7}$$

where

$$h(\theta|\boldsymbol{z}_s) = \frac{\prod_{j\in s} g(z_j|\theta)h(\theta)}{\int \prod_{j\in s} g(z_j|\theta)h(\theta)d\theta} \tag{8}$$

denotes the posterior density of θ. Clearly the denominator of (8) is the marginal density of the sampled data.

As an illustration of the aforementioned setup we consider an important special case. Suppose that $y_1, \ldots, y_N|\theta$ are i.i.d. $N(\theta, \sigma^2)$ and $\theta \sim N(\phi, \tau^2)$, where $\sigma^2(>0)$, ϕ real, and $\tau^2(>0)$ are all known. The parameters σ^2, ϕ, and τ^2 are interpreted respectively as a guess at the amount of variability in the population, the prior guess about the mean of the population, and the measure of how certain one is about the choice of the prior mean. In this case, the posterior distribution of $\boldsymbol{y}(s^c)$ given $\boldsymbol{z}_s$ is multivariate normal. To determine the mean vector and the variance–covariance matrix of the predictive distribution, we first obtain the posterior distribution of θ. Note that elementary calculations yield

$$\begin{aligned} h(\theta|\boldsymbol{z}_s) &\propto \exp\left[-\frac{1}{2\sigma^2}\sum_{j=1}^{n}(z_j-\theta)^2 - \frac{(\theta-\phi)^2}{2\tau^2}\right] \\ &\propto \exp\left[-\frac{(\theta-\eta)^2}{2\delta^2}\right], \end{aligned} \tag{9}$$

where

$$\eta = E_\pi[\theta|\boldsymbol{z}_s] = (1-B)\bar{z}_s + B\phi, \tag{10}$$

$$\delta^2 = V_\pi[\theta|\boldsymbol{z}_s] = \frac{\sigma^2}{n}(1-B), \tag{11}$$

$\bar{z}_s = n^{-1}\sum_{i=1}^{n} z_i$ = sample mean and $B = n^{-1}\sigma^2/(n^{-1}\sigma^2 + \tau^2)$.

Using iterated formulas for expectation and variance, in conjunction with (10) and (11), we get for j, $j' \in s^c$ that

$$\begin{aligned} E_\pi[y_j|\boldsymbol{z}_s] &= E_\pi[E_\pi(y_j|\theta, \boldsymbol{z}_s)|\boldsymbol{z}_s] \\ &= E_\pi[\theta|\boldsymbol{z}_s] \\ &= (1-B)\bar{z}_s + B\phi, \end{aligned} \tag{12}$$

$$
\begin{aligned}
cov_\pi(y_j, y_{j'}|\boldsymbol{z}_s) &= E_\pi[cov_\pi(y_j, y_{j'}|\theta, \boldsymbol{z}_s)|\boldsymbol{z}_s] \\
&\quad + cov_\pi[E_\pi(y_j|\theta, \boldsymbol{z}_s),\ E_\pi(y_{j'}|\theta, \boldsymbol{z}_s)|\boldsymbol{z}_s] \\
&= \sigma^2\delta_{jj'} + \frac{\sigma^2}{n}(1-B), \qquad (13)
\end{aligned}
$$

where $\delta_{jj'} = 0$ if $j \neq j'$ and $\delta_{jj'} = 1$ if $j = j'$. We now use (5), (12), and (13) to find the posterior mean and the posterior variance of the finite population mean $\mu(\boldsymbol{y}) = N^{-1}\sum_{j=1}^{N} y_j$. By (5) and (12),

$$
\begin{aligned}
E_\pi[\mu(\boldsymbol{y})|\boldsymbol{z}_s] &= N^{-1}[n\bar{z}_s + (N-n)\{(1-B)\bar{z}_s + B\phi\}] \\
&= (1 - fB)\bar{z}_s + fB\phi, \qquad (14)
\end{aligned}
$$

where $f = (N-n)/N$ denotes the finite population correction. In a leading article that is most likely the first article on Bayesian approach to survey sampling, Hartley and Rao (1968, p. 552) obtained a similar representation of the Bayes estimator of the finite population mean. Note that the Bayes estimate of the finite population mean is the usual weighted average of the sample mean and the prior mean. A simple comparison between (10) and (14) reveals that the Bayes estimate of finite population mean assigns more weight to the sample mean than the Bayes estimate of the population mean θ does. However, for an infinite population, $f \to 1$ as $N \to \infty$, and the estimate in (14) approaches to the estimate in (10).

Using (13),

$$
\begin{aligned}
V_\pi\left[N^{-1}\sum_{i=1}^{N} y_i|\boldsymbol{z}_s\right] &= N^{-2}V_\pi\left[\sum_{j\in s^c} y_j|\boldsymbol{z}_s\right] \\
&= N^{-2}\left[\sum\sum_{j,j'\in s^c}\left\{\sigma^2\delta_{jj'} + \frac{\sigma^2}{n}(1-B)\right\}\right] \\
&= N^{-2}\left[(N-n)\sigma^2 + \frac{\sigma^2}{n}(1-B)(N-n)^2\right] \\
&= f\sigma^2 N^{-1}\left[1 + \frac{(N-n)\tau^2}{\sigma^2 + n\tau^2}\right] \\
&= f\sigma^2 N^{-1}\frac{\sigma^2 + N\tau^2}{\sigma^2 + n\tau^2} \\
&= fB\left(\tau^2 + \frac{\sigma^2}{N}\right) = fBV_\pi\left(N^{-1}\sum_{i=1}^{N} y_i\right) \qquad (15)
\end{aligned}
$$

because $V_\pi\left(N^{-1}\sum_{i=1}^{N} y_i\right) = \tau^2 + \frac{\sigma^2}{N}$ is the prior variance of $N^{-1}\sum_{i=1}^{N} y_i$.

REMARK 3.2. Note that if $\tau^2 \to \infty$, θ will have a uniform prior over $(-\infty, \infty)$ and since the shrinking factor $B \to 0$, the Bayes estimator of $\mu(\boldsymbol{y})$ approaches the classical estimator $\bar{z}_s$. In this case, the associated posterior variance approaches $f\sigma^2/n$, which is very similar to the measure of uncertainty associated with the classical estimator $\bar{z}_s$.

REMARK 3.3. We can rewrite the shrinking factor B as $B = \sigma^2/(\sigma^2 + n\tau^2) = n^{-1}\sigma^2(1 - B)/\tau^2 = V_\pi(\theta|z_s)/V_\pi(\theta)$. Thus, B is the ratio of the posterior and the prior variance of θ. The extent of shrinking depends on the prior variance. For finite population mean, the ratio of the posterior variance to the prior variance is fB; the appearance of f is due to finite population correction. Also, because of the finite population correction factor, the Bayes estimator of finite population mean shrinks less towards the prior mean than the Bayes estimator of θ does.

3.2. Generalizations

The results derived earlier under the normality assumptions can be generalized easily to the regular one-parameter exponential family with conjugate priors. Suppose that $y_1, \ldots, y_N|\theta$ are i.i.d. with a common pdf

$$g(y|\theta) = \exp\{\theta y - \psi(\theta)\}c(y), \tag{16}$$

while θ has a prior density given by

$$h(\theta) \propto \exp\{\alpha\theta - \nu\psi(\theta)\}. \tag{17}$$

This leads to the posterior

$$\pi(\theta|z_s) \propto \exp\{(n\bar{z}_s + \alpha)\theta - (n + \nu)\psi(\theta)\}. \tag{18}$$

Note that $E(y_i|\theta) = \psi'(\theta)$, and $V(y_i|\theta) = \psi''(\theta)$. Let $E_h(\cdot)$ denote expectation with respect to $h(\theta)$. Using $E_h[\partial \log h(\theta)/\partial\theta] = 0$, $E_h[-\partial^2 \log h(\theta)/\partial\theta^2] = V_h[\partial \log h(\theta)/\partial\theta]$, $\partial \log h(\theta)/\partial\theta = \alpha - \nu\psi'(\theta)$ and $\partial^2 \log h(\theta)/\partial\theta^2 = -\nu\psi''(\theta)$, we get $E_h[\psi\prime(\theta)] = \alpha/\nu$ and $V_h[\psi'(\theta)] = \nu^{-1}E_h[\psi''(\theta)]$. Also, note that $E_h[V(\bar{z}_s|\theta)] = n^{-1}E_h[\psi''(\theta)] = (\nu/n)V_h[\psi'(\theta)]$. Again using iterated expectation formula for $j \in s^c$ $E_\pi[y_j|z_s] = E_\pi[\psi'(\theta)|z_s]$ and $V_\pi[y_j|z_s] = E_\pi[\psi''(\theta)|z_s] + V_\pi[\psi'(\theta)|z_s]$. Noting the similarity between (17) and (18), we get in the same way, from (18) that

$$\begin{aligned} E_\pi[\psi'(\theta)|z_s] &= (n\bar{z}_s + \alpha)/(n + \nu), \\ &= (n\bar{z}_s + \nu\phi)/(n + \nu), \end{aligned} \tag{19}$$

where $\phi = \alpha/\nu$ is the prior mean of $\psi'(\theta)$. Also follows, by a similar comparison, that

$$V_\pi[\psi'(\theta)|z_s] = (n + \nu)^{-1}E_\pi[\psi''(\theta)|z_s]. \tag{20}$$

Note that the posterior expectation of $\psi'(\theta)$ in (19) is linear in $\bar{z}_s$, the sufficient statistic for θ based on the model given by (16). It is possible to express the posterior mean of $\psi'(\theta)$ alternatively as

$$\begin{aligned} &\frac{\bar{z}_s V_h[\psi'(\theta)] + (\alpha/\nu)(\nu/n)V_h[\psi'(\theta)]}{V_h[\psi'(\theta)] + (\nu/n)V_h[\psi'(\theta)]} \\ &\quad = \frac{\bar{z}_s V_h[\psi'(\theta)] + E_h[\psi'(\theta)]E_h[V(\bar{z}_s|\theta)]}{V_h[\psi'(\theta)] + E_h[V(\bar{z}_s|\theta)]}. \end{aligned} \tag{21}$$

The aforementioned result was obtained by Ericson (1969b) under less restrictive conditions. Ericson (1969b) did not use any distributional assumptions in deriving this result,

but instead used posterior linearity assumption. Posterior linearity is discussed in the next section.

Note that as in (10) the posterior mean of $\psi\prime(\theta)$ given by (19) can be expressed as $(1 - B)\bar{z}_s + B\phi$ with $1 - B = n/(\nu + n)$, that is, $B = \nu/(\nu + n)$. In this setup, it can be checked that as in (14), $E_\pi[\mu|\boldsymbol{z}_s] = (1 - fB)\bar{z}_s + fB\phi$ holds. However, in this case, the result given by (15) changes to $E_\pi\{V_\pi(\mu|\boldsymbol{z}_s)\} = fBV_\pi(\mu)$. The expectation on the left-hand side is necessary because in general the posterior variance, unlike in the normal model, depends on the sampled observations.

A very important subfamily of the regular one-parameter natural exponential family (NEF) is the natural exponential family with a quadratic variance function (NEF-QVF) introduced in Morris (1982, 1983a). For such a distribution,

$$\psi''(\theta) = v_0 + v_1\psi'(\theta) + v_2\{\psi'(\theta)\}^2, \tag{22}$$

where v_0, v_1, and v_2 are free from θ and not all zeroes such that the quadratic function $v_0 + v_1x + v_2x^2$ is nonnegative. Morris (1982) gives a complete characterization of the NEF-QVF family of distributions. He shows that there are only six families of distributions having the NEF-QVF structure. Among them, the normal, gamma, binomial, Poisson, and negative binomial are the most widely used distributions in statistical applications. It immediately follows that though the variance is a constant function of the mean for the normal distribution, it involves only the linear term for the Poisson distribution, only the quadratic term for the gamma distribution, and both the linear and the quadratic terms for the binomial and the negative binomial distributions. Note that while two of these members, namely, the normal and gamma are continuous distributions, the other three are examples of discrete distributions.

From (22), and the result $V_h[\psi'(\theta)] = \nu^{-1}E_h[\psi''(\theta)]$, it follows that $V_h[\psi'(\theta)] = (\nu - v_2)^{-1}[v_0 + v_1E_h\{\psi'(\theta)\} + v_2\{E_h(\psi'(\theta))\}^2]$ and $v_2 < \nu$. For the NEF-QVF subfamily, from (20)

$$\begin{aligned} V_\pi[\psi'(\theta)|\boldsymbol{z}_s] &= (n + \nu)^{-1}E_\pi[\psi''(\theta)|\boldsymbol{z}_s] \\ &= (n + \nu)^{-1}E_\pi[v_0 + v_1\psi'(\theta) + v_2\{\psi'(\theta)\}^2|\boldsymbol{z}_s] \\ &= (n + \nu)^{-1}[v_0 + v_1E_\pi(\psi'(\theta)|\boldsymbol{z}_s) \\ &\quad + v_2\{V_\pi(\psi'(\theta)|\boldsymbol{z}_s) + (E_\pi(\psi'(\theta)|\boldsymbol{z}_s))^2\}], \end{aligned}$$

which leads to

$$V_\pi[\psi'(\theta)|\boldsymbol{z}_s] = (n + \nu - v_2)^{-1}[v_0 + v_1E_\pi\{\psi'(\theta)|\boldsymbol{z}_s\} + v_2(E_\pi\{\psi'(\theta)|\boldsymbol{z}_s\})^2]. \tag{23}$$

It follows from (18) and (23) that with a one-parameter exponential likelihood, and an NEF-QVF conjugate prior, the posterior also preserves the NEF-QVF structure. Furthermore, for $n = 0$ (no data problem), the posterior variance (23) reduces to the prior variance.

4. Linear Bayes estimator

In the last section, the derivation of the Bayes estimator requires full specification of the prior distribution $\pi(\boldsymbol{y})$. One can also obtain the estimator of the finite population total

derived earlier through a linear Bayes approach. This approach relies on the specification of the lower order moments, typically the first two moments, of the prior distribution, and is often preferable to the fully Bayesian approach because the latter approach needs complete specification of the prior distribution, which may be difficult.

For the normal or natural exponential hierarchical models with conjugate priors, we noted that the posterior expectation of the population mean is a linear function of the sampled data. Actually, under the exchangeable model, the posterior expectation of the finite population mean is a linear function of the sample mean. One can extend the results of the last section by assuming *posterior linearity*, a concept introduced by Ericson (1969b) in deriving the linear Bayes estimator of the population mean. Indeed, the linear Bayes estimator has been independently discovered and rediscovered by many researchers; notably among them are Ericson (1969b) and Hartigan (1969). Linear Bayes idea was further developed in a series of articles by Goldstein (1975a,b). To derive the linear Bayes estimator we will first prove the following lemma, which is also important in deriving the best linear unbiased predictor. This lemma is a restatement of Result 3.1 of Ericson (1988).

LEMMA 4.1. *Let* $\boldsymbol{W}_1(n_1 \times 1)$ *and* $\boldsymbol{W}_2(n_2 \times 1)$ *be jointly distributed random vectors with mean vectors* $\boldsymbol{\mu}_1$ *and* $\boldsymbol{\mu}_2$*, respectively, and variance covariance matrix*

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}$$

which is finite and positive definite.

Then under a matrix loss the best linear approximation of $E(\boldsymbol{W}_1|\boldsymbol{W}_2)$ is given by $\boldsymbol{P}^*\boldsymbol{W}_2 + \boldsymbol{q}^*$ in the sense that

$$E\big[\{E(\boldsymbol{W}_1|\boldsymbol{W}_2) - \boldsymbol{P}\boldsymbol{W}_2 - \boldsymbol{q}\}\{\ \}^T\big] - E\big[\{E(\boldsymbol{W}_1|\boldsymbol{W}_2) - \boldsymbol{P}^*\boldsymbol{W}_2 - \boldsymbol{q}^*\}\{\ \}^T\big]$$

is nonnegative definite for any matrix $\boldsymbol{P}(n_1 \times n_2)$ and vector $\boldsymbol{q}(n_1 \times 1)$ not depending on $\boldsymbol{W}_2$ where

$$\boldsymbol{P}^* = \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1} \text{ and } \boldsymbol{q}^* = \boldsymbol{\mu}_1 - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\mu}_2. \tag{24}$$

Furthermore, $\boldsymbol{\Sigma}_{11.2} - E[V(\boldsymbol{W}_1|\boldsymbol{W}_2)]$ is n.n.d. It is a null matrix if and only if $E(\boldsymbol{W}_1|\boldsymbol{W}_2)$ is a linear function of $\boldsymbol{W}_2$ with probability 1. Here $\boldsymbol{\Sigma}_{11.2} = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$, and $\{\}^T$ indicates the transpose of the vector that appears immediately ahead.

PROOF. It is easy to verify that

$$\begin{aligned} &E[\{E(\boldsymbol{W}_1|\boldsymbol{W}_2) - \boldsymbol{P}\boldsymbol{W}_2 - \boldsymbol{q}\}\{\ \}^T] \\ &= \boldsymbol{\Sigma}_{11.2} - E\{V(\boldsymbol{W}_1|\boldsymbol{W}_2)\} + (\boldsymbol{P} - \boldsymbol{P}^*)\boldsymbol{\Sigma}_{22}(\boldsymbol{P} - \boldsymbol{P}^*)^T + (\boldsymbol{\mu}_1 - \boldsymbol{P}\boldsymbol{\mu}_2 - \boldsymbol{q})(\ )^T. \end{aligned} \tag{25}$$

By (24) and (25), it easily follows that

$$E\big[\{E(\boldsymbol{W}_1|\boldsymbol{W}_2) - \boldsymbol{P}^*\boldsymbol{W}_2 - \boldsymbol{q}^*\}\{\ \}^T\big] = \boldsymbol{\Sigma}_{11.2} - E\{V(\boldsymbol{W}_1|\boldsymbol{W}_2)\}. \tag{26}$$

From (25) and (26), it follows that

$$E\big[\{E(\boldsymbol{W}_1|\boldsymbol{W}_2) - \boldsymbol{P}\boldsymbol{W}_2 - \boldsymbol{q}\}\{\ \}^T\big] - E\big[\{E(\boldsymbol{W}_1|\boldsymbol{W}_2) - \boldsymbol{P}^*\boldsymbol{W}_2 - \boldsymbol{q}^*\}\{\ \}^T\big]$$
$$= (\boldsymbol{P} - \boldsymbol{P}^*)\boldsymbol{\Sigma}_{22}(\boldsymbol{P} - \boldsymbol{P}^*)^T + (\boldsymbol{\mu}_1 - \boldsymbol{P}\boldsymbol{\mu}_2 - \boldsymbol{q})(\ )^T$$

is nonnegative definite. The difference will be a null matrix provided $\boldsymbol{P} = \boldsymbol{P}^*$ and $\boldsymbol{\mu}_1 - \boldsymbol{P}\boldsymbol{\mu}_2 - \boldsymbol{q} = \boldsymbol{0}$, that is, $\boldsymbol{P} = \boldsymbol{P}^*$ and $\boldsymbol{q} = \boldsymbol{q}^*$. From (26), nonnegative definiteness of $\boldsymbol{\Sigma}_{11.2} - E\{V(\boldsymbol{W}_1|\boldsymbol{W}_2)\}$ follows. It follows further that the difference will be a null matrix if $E(\boldsymbol{W}_1|\boldsymbol{W}_2) = \boldsymbol{P}^*\boldsymbol{W}_2 + \boldsymbol{q}^*$, that is $E(\boldsymbol{W}_1|\boldsymbol{W}_2)$ is a linear function of $\boldsymbol{W}_2$ with probability 1. □

REMARK 4.2. If $E(\boldsymbol{W}_1|\boldsymbol{W}_2) = \boldsymbol{P}\boldsymbol{W}_2 + \boldsymbol{q}$ then $\boldsymbol{P}$ and $\boldsymbol{q}$ are given by (24). Note that in this case the expressions for $E(\boldsymbol{W}_1|\boldsymbol{W}_2)$ and $E[V(\boldsymbol{W}_1|\boldsymbol{W}_2)]$ are exactly the same as those under multivariate normal joint distribution.

COROLLARY 4.3. Under sum of squared error loss, the best linear approximation of $E(\boldsymbol{W}_1|\boldsymbol{W}_2)$ is given by $\boldsymbol{P}^*\boldsymbol{W}_2 + \boldsymbol{q}^*$, where $\boldsymbol{P}^*$ and $\boldsymbol{q}^*$ are as in (24).

PROOF. Proof follows easily by noting that for a vector $\boldsymbol{a}$, $\boldsymbol{a}^T\boldsymbol{a} = tr(\boldsymbol{a}\boldsymbol{a}^T)$.

We have pointed out earlier in this section that it is possible to derive the linear Bayes estimator of the finite population total using posterior linearity, which does not require completely specifying the prior distribution unlike in Section 3. The concept of posterior linearity is described later.

Suppose $E(y_i|\theta) = \kappa(\theta)$ for all i, and θ has a prior distribution under which

$$E[\kappa(\theta)|\boldsymbol{z}_s] = \alpha\bar{z}_s + \beta, \tag{27}$$

where α and β do not depend on the z_i. The aforementioned condition is referred to as the condition of posterior linearity. The posterior linearity holds even outside exponential models using conjugate priors. For example, in nonparametric estimation of a distribution function F of iid random variables $y_1, \ldots, y_N$ by assigning F a Dirichlet process prior, posterior linearity holds. Ericson (1969b) explicitly derived expressions for α and β to obtain the posterior mean in (27). He proved the following result. We omit the proof which can be derived from Lemma 4.1. □

RESULT 4.4. *Let $E(y_i|\theta) = \kappa(\theta)$ for all i, and $E[\kappa(\theta)|\boldsymbol{z}_s] = \alpha\bar{z}_s + \beta$ where α, β do not depend on y_i's. Also, let $0 < V(y_i) < \infty$ for all i. If $E[\kappa(\theta)] = \phi$ and $V[\kappa(\theta)] = \tau^2$, then*

$$E[\kappa(\theta)|\boldsymbol{z}_s] = \frac{\tau^2\bar{z}_s + \phi E[V(\bar{z}_s|\theta)]}{\tau^2 + E[V(\bar{z}_s|\theta)]}. \tag{28}$$

In Result 4.4 if it is assumed that $y_i|\theta$, $i = 1, \ldots, N$ are iid with $E[V(y_i|\theta)] = \sigma^2$, then (28) can be reexpressed as

$$E[\kappa(\theta)|\boldsymbol{z}_s] = (1 - B)\bar{z}_s + B\phi,$$

whereas in (11) $B = n^{-1}\sigma^2/(n^{-1}\sigma^2 + \tau^2)$. Moreover, from Lemma 4.1, because $E[\kappa(\theta)|\boldsymbol{z}_s]$ is a linear function of $\boldsymbol{z}_s$, it follows after some simplification that

$E[V\{\kappa(\theta)|z_s\}] = (1 - B)\sigma^2/n$. *For the finite population mean* $\mu(\boldsymbol{y}) = N^{-1}\sum_{i=1}^{N} y_i$, *it can be checked as in (14) and (15) that*

$$E[\mu(\boldsymbol{y})|z_s] = (1 - fB)\bar{z}_s + fB\phi$$

and

$$E[V(\mu(\boldsymbol{y})|z_s)] = fB\left(\tau^2 + \frac{\sigma^2}{N}\right). \tag{29}$$

REMARK 4.5. In the special case of a regular one-parameter exponential family with conjugate priors considered in Section 3, Ericson's formula (28) reduces to equation (19) obtained earlier. We may note here that the regular one-parameter exponential family along with natural conjugate priors leads to posterior linearity of the population mean. Diaconis and Ylvisaker (1979) showed under mild regularity conditions that the regular one-parameter exponential family along with posterior linearity implies a natural conjugate prior.

REMARK 4.6. It is well-known (cf. Basu, 1971) that Bayesian inference in finite population sampling is independent of the choice of the sampling design, because the posterior distribution remains invariant with respect to the choice of the sampling design [see Eq. (4)]. However, there is a Bayesian way for selecting sampling designs based on minimizing Bayes risk. Under squared error loss, the Bayes risk of the Bayes estimator of the finite population mean is given by $E[V(\mu(\boldsymbol{y})|z_s)]$. It follows from (15) or (29) or the similar expression for $E[V(\mu(\boldsymbol{y})|z_s)]$ in NEF setup (given in the paragraph following (21)), the Bayes risk is the same irrespective of the choice of sampling units. The intuitive reason for this is the basic exchangeability assumption. Thus, from a Bayesian point of view, it does not matter which units are selected. The sampling design could for example be a simple random sampling, or it could be purely purposive. Though in this context it is clear that randomization has no role, Basu (1980, p. 594) advocated pre-randomization of units to address the concern so that a statistician is not falsely accused of doctoring his data. We will see in the next section that in the presence of auxiliary information, the design resulting from the minimization of Bayes risk is different from a simple random sampling design.

5. Bayes estimators of the finite population mean under more complex models

We continue finding Bayes estimators of the finite population mean under more complex models. In particular, we consider appropriate Bayesian models to incorporate auxiliary information and to handle multistage sampling.

5.1. Bayesian models in the presence of auxiliary information

Bayesian models considered so far are mainly built on the idea of exchangeability among the different population units. However, these models, as such, are not appropriate in the presence of auxiliary information, and need to be suitably modified. Most sample surveys include auxiliary information, which when judiciously used, can lead to better estimates of the population characteristics.

We first consider a simple and yet fairly general Bayesian model which can be used to accommodate a number of interesting special cases. The model is described below.

(i) $y_i|\theta$ are independent $N(\theta a_i, \sigma_i^2)$;
(ii) $\theta \sim$ uniform $(-\infty, \infty)$,

where a_i and $\sigma_i^2 (> 0)$, $i = 1, \ldots, N$, are all known. A sample s of fixed size n is drawn. Our first objective is to find the posterior distribution of $\boldsymbol{y}(s^c)$ given $\boldsymbol{z}_s$. Let $\boldsymbol{a}(s^c) = (a_{j_1}, \ldots, a_{j_{N-n}})^T$. We state the following theorem without a proof which is straightforward.

THEOREM 5.1. *Under the model given in (i) and (ii), the joint posterior distribution of* $\boldsymbol{y}(s^c)$ *given* $\boldsymbol{z}_s$ *is an* $(N-n)$*-variate normal with mean vector* $\hat{\theta}\boldsymbol{a}(s^c)$*, and variance–covariance matrix*

$$Diag(\sigma_{j_1}^2, \ldots, \sigma_{j_{N-n}}^2) + d^{-1}\boldsymbol{a}(s^c)\boldsymbol{a}^T(s^c),$$

where

$$\hat{\theta} = \Sigma_s a_i \sigma_i^{-2} y_i / d, \; d = \Sigma_s a_i^2 \sigma_i^{-2}.$$

REMARK 5.2. It can be easily checked that the posterior distribution of θ given $\boldsymbol{z}_s$ is $N(\hat{\theta}, d^{-1})$.

REMARK 5.3. From Theorem 5.1, it can be derived that the posterior distribution of $\sum_{i=1}^N y_i$ given $\boldsymbol{z}_s$ is

$$N\left(\Sigma_s y_i + \hat{\theta}\sum_{k=1}^{N-n} a_{j_k}, \; \sum_{k=1}^{N-n}\sigma_{j_k}^2 + (\Sigma_s a_i)^2 d^{-1}\right). \tag{30}$$

Thus, the Bayes estimator of finite population total $\gamma(\boldsymbol{y}) = \sum_{i=1}^N y_i$ under squared error loss is given by

$$\hat{\gamma}_B = \Sigma_s y_i + \hat{\theta}\sum_{k=1}^{N-n} a_{j_k}. \tag{31}$$

A traditional estimator of the finite population total is given by the Horvitz–Thompson estimator $\hat{\gamma}_{\mathrm{HT}} = \sum_s y_i/\pi_i$ where π_i denotes the inclusion probability of the ith unit. We shall see now how this estimator can also be viewed as a Bayes estimator. Following Ghosh and Sinha (1990), taking $a_i = \pi_i$, and $\sigma_i^2 = \sigma^2\pi_i^2/(1-\pi_i)$, one finds from (31) that

$$\hat{\gamma}_B = \Sigma_s y_i + \frac{\sum_s (1-\pi_i)\pi_i^{-1} y_i}{\sum_s (1-\pi_i)}\Sigma_{s^c}\pi_j.$$

But since $\sum_{s^c}\pi_j = \sum_1^N \pi_i - \sum_s \pi_i = n - \sum_s \pi_i = \sum_s (1-\pi_i)$, it follows that

$$\hat{\gamma}_B = \Sigma_s y_i + \Sigma_s (1-\pi_i)\pi_i^{-1} y_i = \Sigma_s y_i/\pi_i = \hat{\gamma}_{\mathrm{HT}}.$$

Apart from motivating the Horvitz–Thompson estimator as a model-based estimator, the model-based approach provides also an interpretation of the inclusion probabilities π_i. To see this, note that the coefficient of variation of the ith unit $\sigma_i/\theta a_i \propto (1-\pi_i)^{-1/2}$, which is monotonically increasing in π_i. Thus, units with larger coefficients of variation have bigger probabilities of being included in the sample. This seems appropriate because these are the units that are more difficult to predict from the other observed sampled units.

Little (2004) provided an alternate model-based formulation under which $y_i/\pi_i|\theta$ are iid N$(0, \sigma^2)$. Then the MLE or UMVUE of θ is given by $\hat{\theta}_{\mathrm{L}} = n^{-1}\sum_{i\in s} y_i/\pi_i$, and the finite population total $\gamma(\boldsymbol{y})$ is now estimated by

$$\hat{\gamma}_{\mathrm{L}} = \hat{\gamma}_{\mathrm{HT}} + \sum_{i\in s}(y_i - \hat{\theta}_{\mathrm{L}}\pi_i)$$

which is not quite the same as $\hat{\gamma}_{\mathrm{HT}}$, but the difference between $\hat{\gamma}_{\mathrm{L}}$ and $\hat{\gamma}_{\mathrm{HT}}$ converges in probability to zero if the sampling fraction n/N converges to zero. Continuing with this idea, Little (2004) provided also a model-based interpretation of the generalized regression estimator.

We will now derive ratio-type estimators as special cases of the Bayes estimator in (31). Royall (1970b) discussed such estimators very extensively using a frequentist model-based approach. Taking $a_i = x_i$ and $\sigma_i^2 = \sigma^2 v(x_i)$, the Bayes estimator $\hat{\gamma}_{\mathrm{B}}$ takes the form

$$\hat{\gamma}_{\mathrm{B}} = \Sigma_s y_i + \big(\Sigma_s v^{-1}(x_i)x_i y_i / \Sigma_s v^{-1}(x_i)x_i^2\big)\Sigma_{s^c} x_i. \tag{32}$$

For the special choice $v(x_i) = x_i$, (32) simplifies to

$$E\left[\sum_1^N y_i | z_s\right] = (\Sigma_s y_i / \Sigma_s x_i)\sum_1^N x_i, \tag{33}$$

which is the ratio estimator. The other choice $v(x_i) = x_i^2$ leads to

$$E\left[\sum_1^N y_i | z_s\right] = \Sigma_s y_i + n^{-1}(\Sigma_s y_i/x_i)\Sigma_{s^c} x_i, \tag{34}$$

an estimator discussed very extensively in Basu (1971). This estimator is intuitively very appealing because it can be motivated by assuming that the unobserved ratios y_i/x_i behave very much like the corresponding observed ratios. The admissibility of this estimator was proved by Meeden and Ghosh (1983) using a stepwise Bayes approach.

5.2. *Choice of design*

In Remark 4.6 we noted that under the exchangeability assumption, the Bayes estimator of the finite population total has the same Bayes risk for any choice of a sample. Hence the sampling design plays no role. This is not so in the presence of auxiliary information as discussed below.

For a sampling design p, the Bayes risk (under squared error loss) of the finite population total based on the model described in Section 5.1 is given by

$$E\left[\left(\sum_{i=1}^{N} y_i - \Sigma_S y_i - \hat{\theta}\Sigma_{S^c} a_j\right)^2 |\theta\right] = E\Big[\{\Sigma_{S^c}(y_j - \theta a_j - (\hat{\theta} - \theta)a_j)\}^2 |\theta\Big]$$
$$= \sum_s p(s)\Big[\Sigma_{s^c}\sigma_j^2 + (\Sigma_{s^c} a_j)^2 \left(\Sigma_s a_i^2 \sigma_i^{-2}\right)^{-1}\Big]. \tag{35}$$

The aforementioned Bayes risk, fortunately, does not depend on any unknown parameter. One selects those units i that minimize the aforementioned Bayes risk with respect to the a_i.

We now study the special case of Horvitz–Thompson estimator taking $a_i = \pi_i$ and $\sigma_i^2 = \sigma^2\pi_i^2/(1-\pi_i)$. It can be checked that the Bayes risk, which is free from the parameter θ, simplifies to $\sigma^2 \sum_{\mathcal{S}} p(s) \sum_{s^c} \pi_j/(1-\pi_j)$. Clearly, this is minimized w.r.t. $p(s)$ by selecting those units for which $\pi_j/(1-\pi_j)$ is the largest, that is, π_j is the largest. Because the coefficient of variation, as noted earlier, is increasing in π_i, the units with the largest coefficient of variation should purposively be selected in the sample. Attaching arbitrary selection probabilities to the units in the population can often be disastrous as the following example shows. This hilarious example is due to Basu (1971).

Example 5.4. The circus owner is planning to ship his 50 adult elephants and so needs a rough estimate of the total weight of the elephants. As weighing an elephant is a cumbersome process, the owner wants to estimate the total weight by weighing just one elephant. Which elephant should he weigh? So the owner looks back on his records and discovers a list of the elephants' weights taken 3 years ago. He finds that 3 years ago Sambo, the middle-sized elephant was the average (in weight) elephant in the herd. He checks with the elephant trainer who reassures the owner that Sambo may still be considered to be the average elephant in the herd. Therefore, the owner plans to weigh Sambo and take $50y$ (where y is the present weight of Sambo) as an estimate of the total weight $\gamma \equiv \gamma(\boldsymbol{y}) = y_1 + \cdots + y_{50}$ of the 50 elephants. But the circus statistician is horrified when he learns of the owner's purposive sampling plan. "How can you get an unbiased estimate of γ this way?" protests the statistician. So, together they work out a compromise sampling plan. With the help of a table of random numbers they devise a plan that allots a selection probability of 99/100 to Sambo, and equal selection probabilities of 1/4900 to each of the other 49 elephants. Naturally, Sambo is selected and the owner is happy. "How are you going to estimate γ?" asks the statistician. "Why? The estimate ought to be $50y$ of course", says the owner. "Oh! No! That cannot possibly be right", says the statistician. "I recently read an article in the *Annals of Mathematical Statistics* where it is proved that the Horvitz–Thompson estimator is the unique hyperadmissible estimator in the class of all generalized polynomial unbiased estimators." "What is the Horvitz–Thompson estimate in this case?" asks the owner duly impressed. "Because the selection probability of Sambo in our plan was 99/100," says the statistician, the proper estimate of γ is $100y/99$ and not $50y$." "And, how would you have estimated γ," inquires the incredulous owner, "if our sampling made us select, say, the big elephant Jumbo?" "According to what I understand of the

> Horvitz–Thompson estimation method," says the unhappy statistician, "the proper estimate of γ would then have been $4900y$, where y is Jumbo's weight." That is how the statistician lost his circus job and perhaps became a teacher of statistics.

REMARK 5.5. Basu's example clearly demonstrates that unless one uses the design probabilities judiciously, the Horvitz–Thompson estimate can be meaningless. Indeed, our result suggests that if the circus owner has to select one elephant, he should select Jumbo, the big-sized elephant with probability 1. Even if one is hesitant to accept this extreme view of purposive sampling, assigning 1/4900 probability to Jumbo is clearly wrong, and it is no wonder that the statistician lost his circus job.

We will now consider the design issue for the ratio estimator of the total given by (32). In this case, $a_i = x_i$, $\sigma_i^2 = \sigma^2 v(x_i)$ and the expression of Bayes risk given in (35) simplifies to

$$\sigma^2 \sum_s p(s) \left[\Sigma_{s^c} v(x_j) + (\Sigma_{s^c} x_j)^2 / \Sigma_s x_i^2 v^{-1}(x_i) \right]. \tag{36}$$

When $v(x_i) = x_i$, that is, the Bayes estimator is the ratio estimator, the expression given in (36) reduces to

$$\sigma^2 \sum_s p(s) \left(\sum_1^N x_i \right) (\Sigma_{s^c} x_j)/(\Sigma_s x_i). \tag{37}$$

From (37), it is clear that one should select those units i with the largest x_i values. For $v(x_i) = x_i^2$, the expression given in (36) simplifies to

$$\sigma^2 \sum_s p(s) \left[\Sigma_{s^c} x_j^2 + n^{-1} (\Sigma_{s^c} x_j)^2 \right]. \tag{38}$$

Once again, select those units i with the largest x_i values. Clearly, the recommendation of selecting Jumbo in the elephant example is meaningful in light of (37) and (38).

REMARK 5.6. It is interesting to note that the results of Section 5.1 can also be obtained using a frequentist model-based approach. This idea has been put forward by Royall (1970b, 1971). Assume the model under which $y_1, \ldots, y_N$ are independent with $E(y_i) = \theta a_i$, $V(y_i) = \sigma_i^2$. Then the best linear unbiased estimator of θ is given by $\hat{\theta}$ defined in Theorem 5.1. Consequently, the best linear unbiased predictor of $\sum_1^N y_i$ is given by $\sum_s y_i + \hat{\theta} \sum_{s^c} a_j$. This is the same as the Bayes estimator in (31). Also, the frequentist risk is the same as the Bayes risk (conditional on θ) derived earlier. Thus, there appears to be a synthesis between the Bayesian and the frequentist methods of inference. Indeed, the frequentist model-based approach yields identical point estimates without involving any distributional assumptions. The Bayes procedure, however, has its advantages when one wants to construct credible sets for functions of $y_1, \ldots, y_N$ because one can then use the normal posterior distribution.

5.3. Multistage sampling

Multistage sampling is frequently used in sample surveys. For example, within a given state, at Stage 1, a sample of counties is chosen. At Stage 2, a sample of blocks is chosen

from each selected county, while in Stage 3, a sample of dwellings is selected from each chosen block. This is an example of three-stage sampling. On the other hand, for a fixed county, this is an example of two-stage sampling.

For simplicity of exposition, we shall consider only the two-stage sampling. To be specific, suppose there are M clusters or primary sampling units (PSU) labeled $1, \ldots, M$. The ith cluster contains N_i elements. The values associated with the N_i units in the ith PSU are denoted by $y_{i1}, \ldots, y_{iN_i}$, $i = 1, \ldots, M$. In the first stage, only a sample of m clusters is selected. In the second stage, a sample of n_i distinct elements is selected from the N_i elements in the ith sampled cluster. For simplicity of notation, we relabel the clusters and the units within the clusters so that we can denote the sampled clusters by $1, \ldots, m$, and the values of the sampled units within the ith selected cluster by $y_{i1}, \ldots, y_{in_i}$. It is well known that the special case $n_i = N_i$ for all $i = 1, \ldots, m$ corresponds to the cluster sampling.

The Bayesian analysis of two-stage sampling was first carried out by Scott and Smith (1969). They considered the following Bayesian model:

MODEL 5.1.

I. Conditional on $\theta_1, \ldots, \theta_M$, for $i = 1, \ldots, M$, $y_{i1}, \ldots, y_{iN_i}$ are independent and for each i $y_{i1}, \ldots, y_{iN_i}$ are iid $N\left(\theta_i, \delta_i^2\right)$;

II. $\theta_1, \ldots, \theta_M$ are iid $N(\nu, \tau^2)$.

Throughout this section $\delta_1^2, \ldots, \delta_m^2$ and τ^2 are assumed known.

Although the aforementioned model does not use an exchangeable prior (due to different δ_i^2) for the entire population, it uses an exchangeable prior for $\theta_1, \ldots, \theta_M$, and conditional on θ_i, an exchangeable prior for $y_{i1}, \ldots, y_{iN_i}$ within a PSU. Our objective is to infer about $y_{i,n_i+1}, \ldots, y_{iN_i}$ $(i = 1, \ldots, m)$ and $y_{i1}, \ldots, y_{iN_i}$ $(i = m+1, \ldots, M)$ given $\boldsymbol{y}(s) = (y_{11}, \ldots, y_{1n_1}, \ldots, y_{m1}, \ldots, y_{mn_m})^T$. In particular, we may be interested in the finite population mean $\mu_i(\boldsymbol{y}) = N_i^{-1} \sum_{j=1}^{N_i} y_{ij}$ of the ith PSU or in the overall population mean $\mu(\boldsymbol{y}) = N^{-1} \sum_{i=1}^{M} \sum_{j=1}^{N_i} y_{ij}$, where $N = \sum_{i=1}^{M} N_i$. The derivation of the predictive distribution proceeds as follows. Let $B_i = n_i^{-1}\delta_i^2 \big/ \left(n_i^{-1}\delta_i^2 + \tau^2\right)$, $i = 1, \ldots, m$. Now, it is immediate that $\theta_1, \ldots, \theta_M | \boldsymbol{y}(s)$ are mutually independent. As in (9) and (10), for $1 \le i \le m$,

$$\theta_i | \boldsymbol{y}(s) \sim N((1 - B_i)\bar{y}_{is} + B_i \nu, \tau^2 B_i), \tag{39}$$

where $\bar{y}_{is} = n_i^{-1} \sum_{j=1}^{n_i} y_{ij}$. For $m + 1 \le i \le M$,

$$\theta_i | \boldsymbol{y}(s) \sim N(\nu, \tau^2). \tag{40}$$

Hence, conditional on $\boldsymbol{y}(s)$, the joint posterior of $y_{i,n_i+1}, \ldots, y_{iN_i}$ $(i = 1, \ldots, m)$ and $y_{i1}, \ldots, y_{iN_i}$ $(i = m + 1, \ldots, M)$ is multivariate normal with the first two moments calculated as follows.

Let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)^T$. Then, for $j = n_i + 1, \ldots, N_i$, $1 \le i \le m$,

$$\begin{aligned} E[y_{ij} | \boldsymbol{y}(s)] &= E[E\{y_{ij} | \boldsymbol{\theta}, \boldsymbol{y}(s)\} | \boldsymbol{y}(s)] \\ &= (1 - B_i)\bar{y}_{is} + B_i \nu; \end{aligned} \tag{41}$$

$$V[y_{ij}|\boldsymbol{y}(s)] = V[E\{y_{ij}|\boldsymbol{\theta}, \boldsymbol{y}(s)\}|\boldsymbol{y}(s)] + E[V\{y_{ij}|\boldsymbol{\theta}, \boldsymbol{y}(s)\}|\boldsymbol{y}(s)]$$
$$= V[\theta_i|\boldsymbol{y}(s)] + E[\delta_i^2|\boldsymbol{y}(s)]$$
$$= \tau^2 B_i + \delta_i^2. \tag{42}$$

Also, for $n_i + 1 \le j \ne j' \le N_i$, $1 \le i \le m$,

$$\mathrm{Cov}[y_{ij}, y_{ij'}|\boldsymbol{y}(s)] = \mathrm{Cov}[E\{y_{ij}|\boldsymbol{\theta}, \boldsymbol{y}(s)\}, E\{y_{ij'}|\boldsymbol{\theta}, \boldsymbol{y}(s)\}|\boldsymbol{y}(s)]$$
$$+ E[\mathrm{Cov}\{y_{ij}, y_{ij'}|\boldsymbol{\theta}, \boldsymbol{y}(s)\}|\boldsymbol{y}(s)]$$
$$= V[\theta_i|\boldsymbol{y}(s)] + E[0|\boldsymbol{y}(s)]$$
$$= \tau^2 B_i. \tag{43}$$

Next, for $m + 1 \le i \le M$, $1 \le j \le N_i$,

$$E[y_{ij}|\boldsymbol{y}(s)] = E[y_{ij}] = \nu, \tag{44}$$
$$V[y_{ij}|\boldsymbol{y}(s)] = V(y_{ij}) = \tau^2 + \delta_i^2. \tag{45}$$

Also, for $m + 1 \le i \le M$, $1 \le j \ne j' \le N_i$,

$$\mathrm{Cov}[y_{ij}, y_{ij'}|\boldsymbol{y}(s)] = \mathrm{Cov}[\theta_i, \theta_i|\boldsymbol{y}(s)] = \tau^2. \tag{46}$$

Finally, for $m + 1 \le i \ne i' \le M$, $1 \le j \le N_i$, $1 \le j' \le N_{i'}$,

$$\mathrm{Cov}[y_{ij}, y_{i'j'}|\boldsymbol{y}(s)] = 0. \tag{47}$$

Hence, for the ith cluster ($1 \le i \le m$),

$$E\left[N_i^{-1}\sum_{j=1}^{N_i} y_{ij}|\boldsymbol{y}(s)\right] = N_i^{-1}\left[\sum_{j=1}^{n_i} y_{ij} + \sum_{j=n_i+1}^{N_i} E\{y_{ij}|\boldsymbol{y}(s)\}\right]$$
$$= N_i^{-1}[n_i\bar{y}_{is} + (N_i - n_i)\{(1 - B_i)\bar{y}_{is} + B_i\nu\}]$$
$$= (1 - f_i B_i)\bar{y}_{is} + f_i B_i \nu, \tag{48}$$

where, analogous to the definition of f, $f_i = (N_i - n_i)/N_i$ is the finite population correction for the ith cluster. Also, calculations similar to (15), for $1 \le i \le m$, lead to

$$V\left[N_i^{-1}\sum_{j=1}^{N_i} y_{ij}|\boldsymbol{y}(s)\right] = f_i B_i\left(N_i^{-1}\delta_i^2 + \tau^2\right). \tag{49}$$

On the other hand, for $m + 1 \le i \le M$,

$$E\left[N_i^{-1}\sum_{j=1}^{N_i} y_{ij}|\boldsymbol{y}(s)\right] = \nu, \tag{50}$$

$$V\left[N_i^{-1}\sum_{j=1}^{N_i} y_{ij}|\boldsymbol{y}(s)\right] = N_i^{-1}\delta_i^2 + \tau^2. \tag{51}$$

Note that (50) and (51) correspond, respectively, to (48) and (49) by taking $n_i = 0$ and noting that in that case $B_i = 1$ and $f_i = 1$. It easily follows that the posterior distribution of finite population mean $\mu(\boldsymbol{y}) = N^{-1} \sum_{i=1}^{M} \sum_{j=1}^{N_i} y_{ij}$ is given by normal with mean

$$\begin{aligned} E[\mu(\boldsymbol{y})|\boldsymbol{y}(s)] &= N^{-1} E\left[\sum_{i=1}^{m} \sum_{j=1}^{n_i} y_{ij} + \sum_{i=1}^{m} \sum_{j=n_i+1}^{N_i} y_{ij} + \sum_{i=m+1}^{M} \sum_{j=1}^{N_i} y_{ij} | \boldsymbol{y}(s) \right] \\ &= N^{-1} \left[\sum_{i=1}^{m} n_i \bar{y}_{is} + \sum_{i=1}^{m} (N_i - n_i)\{(1 - B_i)\bar{y}_{is} + B_i \nu\} + \sum_{i=m+1}^{M} N_i \nu \right] \\ &= N^{-1} \left[\sum_{i=1}^{m} N_i \{(1 - f_i B_i)\bar{y}_{is} + f_i B_i \nu\} + \sum_{i=m+1}^{M} N_i \nu \right]. \end{aligned} \tag{52}$$

Clearly, the posterior mean given earlier is a weighted average of the posterior means of the PSU means (cf. (48) and (50)), the weights being proportional to the sizes of the PSUs.

Finally, by (49) and (51), the posterior variance of the finite population mean is given by

$$\begin{aligned} V[\mu(\boldsymbol{y})|\boldsymbol{y}(s)] &= N^{-2} \left[\sum_{i=1}^{m} N_i^2 f_i B_i \left(N_i^{-1} \delta_i^2 + \tau^2 \right) + \sum_{i=m+1}^{M} N_i^2 \left(N_i^{-1} \delta_i^2 + \tau^2 \right) \right] \\ &= \tau^2 N^{-2} \left[\sum_{i=1}^{m} N_i^2 f_i B_i + \sum_{i>m}^{M} N_i^2 \right] + N^{-2} \left[\sum_{i=1}^{m} N_i f_i B_i \delta_i^2 + \sum_{i>m}^{M} N_i \delta_i^2 \right]. \end{aligned} \tag{53}$$

Remark 5.7. As in Section 4, it is possible to derive the posterior mean of $\mu(\boldsymbol{y})$ based only on the assumption of posterior linearity, that is retain the moment assumptions as given in I and II of Model 5.1, but instead of assuming normality, assume only that

$$E[\theta_i|\boldsymbol{y}(s)] = \sum_{j=1}^{n_i} a_{ij} y_{ij} + b_i, \quad i = 1, \ldots, m. \tag{54}$$

Then applying Lemma 4.1 one can obtain expressions for $E\left[N_i^{-1} \sum_{j=1}^{N_i} y_{ij} | \boldsymbol{y}(s) \right]$, $i = 1, \ldots, M$ or for $E[\mu(\boldsymbol{y})|\boldsymbol{y}(s)]$ that match corresponding expressions under the normality assumption. Also, it follows from this lemma that the expected value of the conditional variance $V[\mu(\boldsymbol{y})|\boldsymbol{y}(s)]$ will match the right-hand side of (53).

The results derived so far do not take into account any interrelationship among the clusters. In fact, we obtain separate estimates of the means from the different clusters, and take their weighted average to obtain an estimate of the finite population mean. However, Scott and Smith (1969) extended Model 5.1 and introduced a hierarchical Bayes model that builds correlations among the clusters. The hierarchical model of Scott and Smith (1969) is as follows.

MODEL 5.2.

I. Conditional on $\theta_1, \ldots, \theta_M$ *and* ν, *for* $i = 1, \ldots, M$, $y_{i1}, \ldots, y_{iN_i}$, *are independent and* $y_{i1}, \ldots y_{iN_i}$ *are iid* $N(\theta_i, \delta_i^2)$, $i = 1, \ldots, M$;

II. Conditional on ν, *PSU means* $\theta_1, \ldots, \theta_M$ *are iid* $N(\nu, \tau^2)$;

III. $\nu \sim$ *uniform* $(-\infty, \infty)$.

Here $\delta_1^2, \ldots, \delta_m^2$ *and* τ^2 *are assumed known.*

The objective is once again to find the joint posterior distribution of $y_{ij} (j = n_i + 1, \ldots, N_i, i = 1, \ldots, m)$ and $y_{ij} (j = 1, \ldots, N_i, i = m + 1, \ldots, M)$ given $\boldsymbol{y}(s)$. The joint distribution is once again multivariate normal with

$$\begin{aligned} E[y_{ij}|\boldsymbol{y}(s)] &= E[E\{y_{ij}|\boldsymbol{\theta}, \nu, \boldsymbol{y}(s)\}|\boldsymbol{y}(s)] \\ &= E[\theta_i|\nu, \boldsymbol{y}(s)] \\ &= (1 - B_i)\bar{y}_{is} + B_i E[\nu|\boldsymbol{y}(s)], \end{aligned} \tag{55}$$

for $j = n_i + 1, \ldots, N_i, i = 1, \ldots, m$, while for $j = 1, \ldots, N_i, i = m + 1, \ldots, M$,

$$\begin{aligned} E[y_{ij}|\boldsymbol{y}(s)] &= E[\theta_i|\nu, \boldsymbol{y}(s)] \\ &= E[\nu|\boldsymbol{y}(s)]. \end{aligned} \tag{56}$$

In fact to derive the posterior mean and posterior variance expressions, we need $E[\nu|\boldsymbol{y}(s)]$ and $V[\nu|\boldsymbol{y}(s)]$ in conjunction with results derived in (48)–(53). Note that $\boldsymbol{y}(s)|\nu$ is multivariate normal with $E[\boldsymbol{y}(s)|\nu] = \nu\big(\mathbf{1}_{n_1}^T, \ldots, \mathbf{1}_{n_m}^T\big)^T$ and $V[\boldsymbol{y}(s)|\nu] = \oplus_{i=1}^m \big(\delta_i^2 \boldsymbol{I}_{n_i} + \tau^2 \mathbf{J}_{n_i}\big)$, is a block diagonal matrix and $\mathbf{J}_{n_i} = \mathbf{1}_{n_i}\mathbf{1}_{n_i}^T$. Now, since in our case $p(\nu|\boldsymbol{y}(s)) \propto p(\boldsymbol{y}(s)|\nu)$, it implies that the posterior distribution of ν is normal with

$$E[\nu|\boldsymbol{y}(s)] = \bar{y}_{ws} \text{ and } V[\nu|\boldsymbol{y}(s)] = v^*, \tag{57}$$

where

$$\bar{y}_{ws} = \frac{\sum_{i=1}^m (1 - B_i)\bar{y}_{is}}{\sum_{i=1}^m (1 - B_i)} \quad \text{and} \quad v^* = \tau^2 \left[\sum_{i=1}^m (1 - B_i)\right]^{-1}.$$

Using (48)–(53) and (55)–(57), it can be checked that for $1 \le i \le m$

$$E\left[N_i^{-1} \sum_{j=1}^{N_i} y_{ij}|\boldsymbol{y}(s)\right] = (1 - f_i B_i)\bar{y}_{is} + f_i B_i \bar{y}_{ws}, \tag{58}$$

$$V\left[N_i^{-1} \sum_{j=1}^{N_i} y_{ij}|\boldsymbol{y}(s)\right] = f_i B_i \big(N_i^{-1}\delta_i^2 + \tau^2\big) + f_i^2 B_i^2 v^*, \tag{59}$$

and for $m + 1 \le i \le M$,

$$E\left[N_i^{-1} \sum_{j=1}^{N_i} y_{ij}|\boldsymbol{y}(s)\right] = \bar{y}_{ws}, \tag{60}$$

$$V\left[N_i^{-1}\sum_{j=1}^{N_i} y_{ij}|\boldsymbol{y}(s)\right] = N_i^{-1}\delta_i^2 + \tau^2 + v^*, \tag{61}$$

and

$$E[\mu(\boldsymbol{y})|\boldsymbol{y}(s)] = N^{-1}\left[\sum_{i=1}^{m} N_i\{(1-f_iB_i)\bar{y}_{is} + f_iB_i\bar{y}_{ws}\} + \sum_{i=m+1}^{M} N_i\bar{y}_{ws}\right], \tag{62}$$

$$V[\mu(\boldsymbol{y})|\boldsymbol{y}(s)] = \text{Right side of (53)} + N^{-2}\left[\sum_{i=1}^{m} N_i f_i B_i + \sum_{i=m+1}^{M} N_i\right]^2 v^*. \tag{63}$$

The estimators of $\mu_i(\boldsymbol{y})$ or $\mu(\boldsymbol{y})$ in (58) or (62) are known as HB estimators.

REMARK 5.8. The only difference between the formula given in (62) and the one given in (52) is that ν is replaced by $\bar{y}_{ws}$. Thus, units in the unsampled clusters are estimated by a weighted average of the sample means instead of the second stage prior mean. The estimator given in (62) is, therefore, more adaptive than the one given in (52). Also, as expected, the estimator in (62) is more variable than the one in (52). This is evident from the posterior variance expression in (63).

REMARK 5.9. The unknown mean ν of the population of cluster means in Model 5.1 was given a uniform $(-\infty, \infty)$ prior in Model 5.2. Instead of assigning a uniform prior to ν, one can estimate it from the data based on the marginal distribution $p(\boldsymbol{y}(s)|\nu)$ of $\boldsymbol{y}(s)$, which is multivariate normal given earlier. It easily follows based on this distribution that $\bar{y}_{ws}$ is the MLE as well as the UMVUE of ν. If we replace the unknown ν appearing in the Bayes estimates (48) or (52) by this estimate, the resulting estimates are known as empirical Bayes (EB) estimates (see Berger, 1985, Section 4.3; or Carlin and Louis, 1996, Section 3.1). In this example and for this estimate of ν, the resulting EB estimates of $\mu_i(\boldsymbol{y})$ or $\mu(\boldsymbol{y})$ are identical to the corresponding HB estimates given by (58) or (62). Although such exact coincidence of EB and HB estimates occurs only for normal hierarchical models, in general, these estimates are very similar. However, if we use the quantities in (49) or (53) as the associated measure of uncertainty for the EB estimator, we can clearly see from (59) or (63) that such measures will result in the underestimation of the true uncertainty. Posterior variance (49) or (53) are often termed as naive EB measures of uncertainty. Though the posterior variance formulas in (59) or (63) for the HB estimators automatically account for the uncertainty due to estimation of ν, this is not so for the corresponding EB measures. This is definitely a clear advantage of the HB procedure over the EB procedure. To derive an accurate measure of uncertainty of an EB estimator, we need to ascertain in a nonnaive way the contribution of the unknown ν. An early application is due to Morris (1983b). This point is further elaborated in the sequel.

REMARK 5.10. The extension of the hierarchical model of Scott and Smith (1969) to three-stage sampling is given in Malec and Sedransk (1985). Interpretation of the estimator

of Scott and Smith (1969) using a model-based prediction approach is given in Royall (1976). The latter also provides alternative model-based estimators of the finite population mean.

6. Stratified sampling and domain estimation

A useful concept in finite population sampling is stratification. In stratified sampling, the population of N units is first divided into subpopulations of $N_1, \ldots, N_m$ units respectively, where $\sum_{i=1}^{m} N_i = N$. These subpopulations are called strata. Samples of sizes $n_1, \ldots, n_m$ are available from these m strata or domains. Often the N_i and n_i are known in advance. In other instances, a sample of size n is drawn from the entire population, and then one finds out the sample sizes $n_1, \ldots, n_m$ for the m strata. This is known as poststratification. In this chapter, we will not consider poststratification.

Stratified sampling is often used to facilitate the administration of a survey. There may be several field offices located in different regions that will conduct the survey. From cost and other considerations it is meaningful to subdivide the entire population into several regions and draw independent samples from each of them. Another reason for using stratified sampling is to improve the precision of the estimator of the population total by dividing a heterogeneous population into several homogeneous strata.

In certain surveys, one may be required to produce estimates of known precision for some or all of the subdivisions. In such case, it is better to treat each such subdivisions as a population. This is similar to small area estimation problem treated in a sequel to this chapter.

In Bayesian approach to survey sampling, stratification is a useful concept in specifying prior information for a finite population where the exchangeability assumption is less tenable for the whole population. In such case, one stratifies the population such that the exchangeability assumption holds, at least approximately, within each stratum. This is similar to stratification in traditional sampling where the goal is to reach within stratum homogeneity.

We will assume that there are m strata $\mathcal{U}_k$, $k = 1, \ldots, m$, partitioning the population $\mathcal{U}$. Let h_i denote the stratum membership of the ith unit. Here we will assume that the vector $\boldsymbol{h} = (h_1, \ldots, h_N)^T$ and the strata sizes $N_1, \ldots, N_m$ and the corresponding sample sizes $n_1, \ldots, n_m$ are completely known and positive. For discussion of the case where $\boldsymbol{h}$ is not completely known, one may refer to Chapter 3 of Ghosh and Meeden (1997).

Let y_{kj}, $j = 1, \ldots, N_k$, be the value of the variable of interest for the jth unit in the kth stratum, $k = 1, \ldots, m$. Also, let $\boldsymbol{y}_k = (y_{k1}, \ldots, y_{kN_k})^T$, and the finite population mean and finite population variance of the kth stratum be μ_k and V_k where

$$\mu_k = N_k^{-1} \sum_{j=1}^{N_k} y_{kj} \quad \text{and} \quad V_k = N_k^{-1} \sum_{j=1}^{N_k} (y_{kj} - \mu_k)^2, \quad k = 1, \ldots, m.$$

Obviously, the overall finite population mean μ is given by $\mu = \sum_{k=1}^{m} N_k \mu_k / N$. For simplicity of notation, we denote the sample values from the kth stratum by z_{kj}, $j = 1, \ldots, n_k$, $k = 1, \ldots, m$.

Before we start our discussion on EB and HB prediction of stratum means in the following subsections, we would like to mention that we will not consider the choice of Bayesian optimal designs in stratified sampling. For interesting discussion on this issue we refer to Ericson (1965) and Rao and Ghangurde (1972).

6.1. EB estimation in stratified sampling

Ghosh and Meeden (1997) discussed EB and HB estimation of the stratified mean vector $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_m)^T$. Their normal theory EB approach for solving this problem is based on the model given in the following sections.

MODEL 6.1.

I. Conditional on $\theta_1, \ldots, \theta_m$, *for* $j = 1, \ldots, N_k, k = 1, \ldots, m$, y_{kj} *are independently normally distributed with* $E(y_{kj}|\theta_1, \ldots, \theta_m) = \theta_k$ *and* $V(y_{kj}|\theta_1, \ldots, \theta_m) = \sigma^2$;

II. $\theta_1, \ldots, \theta_m$ *are iid* $N(\nu, \tau^2)$.

Let $\boldsymbol{z}$ denote the sample vector and $\bar{z}_k = n_k^{-1} \sum_{j=1}^{n_k} z_{kj}$, the kth stratum sample mean. Denote the ratio σ^2/τ^2 of the variance components by λ and let $B_k = \lambda/(\lambda + n_k)$. As in Section 5, one can derive that the joint posterior distribution of y_{kj}, $j = n_k + 1, \ldots, N_k$, $k = 1, \ldots, m$ is multivariate normal. From this result, it easily follows that

$$E[\mu_k|z] = N_k^{-1}[n_k\bar{z}_k + (N_k - n_k)\{(1 - B_k)\bar{z}_k + B_k\nu\}], \tag{64}$$

$$V[\mu_k|z] = N_k^{-1}(N_k - n_k)B_k\left(\tau^2 + N_k^{-1}\sigma^2\right), \tag{65}$$

$$\mathrm{Cov}[\mu_k, \mu_k'|z] = 0 \quad \text{for} \quad k \neq k' = 1, \ldots, m. \tag{66}$$

Because both λ and ν are unknown, in an EB analysis they are estimated from the marginal distribution of the data. As mentioned earlier, it can be seen that marginally $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_m$ are independent with $\boldsymbol{y}_k$ distributed as $N\left(\nu\mathbf{1}_{N_k}, \sigma^2\left(\boldsymbol{I}_{N_k} + \lambda\mathbf{1}_{N_k}\mathbf{1}_{N_k}^T\right)\right)$.

Let $n_T = \sum_{k=1}^m n_k$ and $\bar{z} = n_T^{-1} \sum_{k=1}^m n_k\bar{z}_k$. Define

$$\mathrm{MSB} = (m - 1)^{-1} \sum_{k=1}^{m} n_k(\bar{z}_k - \bar{z})^2, \tag{67}$$

$$\mathrm{MSW} = (n_T - m)^{-1} \sum_{k=1}^{m} \sum_{j=1}^{n_k} (z_{kj} - \bar{z}_k)^2. \tag{68}$$

Also, let $h = n_T - \sum_{k=1}^m n_k^2/n_T$. Note that to estimate $B_k = (1 + \lambda^{-1}n_k)^{-1}$, appearing in the Bayes estimator (64), we need to estimate λ^{-1}. Ghosh and Meeden (1997) used the following ANOVA estimator of λ^{-1} given by

$$\hat{\lambda}^{-1} = \max[0, \{(m - 1)\mathrm{MSB}/((m - 3)\mathrm{MSW}) - 1\}(m - 1)h^{-1}], \tag{69}$$

assuming $m \geq 4$.

To estimate ν, for known λ, Ghosh and Meeden (1997) have shown that the MLE of ν is given by

$$\tilde{\nu}(\lambda) = \sum_{k=1}^{m}(1 - B_k)\bar{z}_k \sum_{k=1}^{m}(1 - B_k)$$
$$= \sum_{k=1}^{m} n_k\big(1 + \lambda^{-1}n_k\big)^{-1}\bar{z}_k \Big/ \sum_{k=1}^{m} n_k\big(1 + \lambda^{-1}n_k\big)^{-1}. \tag{70}$$

By (69) and (70), an estimator of ν is given by

$$\hat{\nu} = \begin{cases} \sum_{k=1}^{m}(1 - \hat{B}_k)\bar{z}_k / \sum_{k=1}^{m}(1 - \hat{B}_k) & \text{if} \quad \hat{\lambda}^{-1} > 0 \\ \sum_{k=1}^{m} n_k\bar{z}_k/n_T & \text{if} \quad \hat{\lambda}^{-1} = 0. \end{cases} \tag{71}$$

This estimator of ν is slightly different from the one proposed by Ghosh and Meeden (1997); see their Eq. (4.19). They defined for $\hat{\lambda}^{-1} = 0$, $\hat{\nu} = m^{-1}\sum_{k=1}^{m}\bar{z}_k$. Substituting the estimators of $\hat{\nu}$ and $\hat{B}_k$ respectively for ν and B_k, it follows from (64) that an EB predictor of μ_k is given by

$$\hat{\mu}_{k,\mathrm{EB}} = N_k^{-1}[n_k\bar{z}_k + (N_k - n_k)\{(1 - \hat{B}_k)\bar{z}_k + \hat{B}_k\hat{\nu}\}]. \tag{72}$$

A naive measure of uncertainty associate with (72) can be obtained from

$$\hat{V}(\mu_k|\boldsymbol{z}) = N_k^{-1}(N_k - n_k)\hat{B}_k\hat{\sigma}^2(\hat{\lambda}^{-1} + N_k^{-1}),$$

where $\hat{\sigma}^2 = \mathrm{MSW}$. This usually underestimates the true measure of uncertainty because it ignores the estimation error associated with the parameters σ^2, τ^2, and ν. The above measure of uncertainty is accurate only to the first order. In contrast, we will see in subsection 6.3 that HB measures of uncertainly take into account all sources of error in estimating the hyperparameters. More accurate (second-order accurate) measures of uncertainty associated with the EB predictors as the number of strata increases are available. We will not discuss them here because these problems are very much similar to accurate estimation of the mean squared error of EBLUPs in small area estimation treated in a sequel of this chapter. This similarity is due to the fact that one can treat a stratum as a small domain. However, such mean squared error estimation will not be valid unless the number of strata is large.

In this context, it is worthwhile to mention tht the EB predictors of the strata means are optimal in certain sense. In Section 4.3, Ghosh and Meeden (1997) have shown that under average squared error loss and the Model 6.1, the difference of the Bayes risks of the EB predictors $\hat{\mu}_{k,\mathrm{EB}}$ and the Bayes predictors in (69) goes to zero as m goes to infinity.

In stratified sampling, one is usually interested in the finite population mean μ. An EB predictor of μ is given by $\hat{\mu}_{\mathrm{EB}} = N^{-1}\sum_{k=1}^{m} N_k\hat{\mu}_{k,\mathrm{EB}}$ along with a naive measure of uncertainty given by $N^{-2}\sum_{k=1}^{m} N_k^2 V(\mu_k|\boldsymbol{z})$. Note that to the first order of approximation of the estimated measure of uncertainty, the EB predictors $\hat{\mu}_{k,\mathrm{EB}}$ are uncorrelated.

Little (2004) considered a model similar to Ghosh and Meeden (1986) except that in Part I of the model, he allowed different error variances σ_i^2 $(i = 1, \ldots, m)$ for different strata. He considered independent vague priors $\pi\big(\theta_i, \sigma_i^2\big) = \sigma_i^{-2}$, $i = 1, \ldots, m$. Although the resulting posterior does not come out in a closed form, it is easy to draw

samples from this posterior by the standard Markov chain Monte Carlo numerical integration technique, and the resulting inference can easily be based on these Monte-Carlo samples.

6.2. Linear Bayes estimation of stratum means

In the last section, we have considered EB estimation of the finite population stratum means assuming a normal superpopulation model. In this section, we will relax the normality assumption made in Model 6.1, and replace it by the posterior linearity assumption introduced in Section 4. We assume the following model.

MODEL 6.2.

I. *Conditional on* $\theta_1, \ldots, \theta_m$, *for* $j = 1, \ldots, N_k$, $k = 1, \ldots, m$, y_{kj} *are mutually independent with* $E(y_{kj}|\theta_1, \ldots, \theta_m) = \theta_k$ *and* $V(y_{kj}|\theta_1, \ldots, \theta_m) = \mu_2(\theta_k)$, $k = 1, \ldots, m$;

II. θ_k *are iid with mean* ν *and variance* τ^2;

III. $0 < \sigma^2 = E[\mu_2(\theta_k)] < \infty$.

We also assume the posterior linearity that says that

$$E[\theta_k|z] = \sum_{j=1}^{n_k} a_{kj} z_{kj} + b_k, \quad k = 1, \ldots, m, \tag{73}$$

where the a_{kj} and the b_k are constants not depending on the y. Because conditionally y_{kj}, $j = 1, \ldots, n_k$ are iid for given $\theta_1, \ldots, \theta_m$, it follows from Goldstein (1975a) or from Result 4.4 that (73) leads to

$$E(\theta_k|z) = a_k \bar{z}_k + b_k, \quad k = 1, \ldots, m, \tag{74}$$

where a_k are constants. In fact, by Resut 4.4, we get that

$$a_k = 1 - B_k, \quad b_k = B_k \nu, \quad \text{and} \quad B_k = n_k^{-1}\sigma^2 \big/ \left(n_k^{-1}\sigma^2 + \tau^2\right).$$

Hence, as in Ghosh and Meeden (1997, p. 173) under the average squared error loss, the Bayes estimator of $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_m)^T$ is given by $\hat{\mu}_{\mathrm{B}} = (\hat{\mu}_{1,\mathrm{B}}, \ldots, \hat{\mu}_{m,\mathrm{B}})^T$, where

$$\begin{aligned}
\hat{\mu}_{k,\mathrm{B}} &= E[\mu_k|z] \\
&= N_k^{-1}\left[n_k \bar{z}_k + \sum_{j=n+1}^{N_k} E(y_{kj}|z)\right] \\
&= N_k^{-1}\left[n_k \bar{z}_k + \sum_{j=n_k+1}^{N_k} E\{E(y_{kj}|\boldsymbol{\theta}, z)|z\}\right] \\
&= N_k^{-1}[n_k \bar{z}_k + (N_k - n_k)E(\theta_k|z)] \\
&= N_k^{-1}[n_k \bar{z}_k + (N_k - n_k)\{(1 - B_k)\bar{z}_k + B_k \nu\}] \\
&= (1 - f_k B_k)\bar{z}_k + f_k B_k \nu \\
&= \bar{z}_k - f_k B_k (\bar{z}_k - \nu),
\end{aligned} \tag{75}$$

where $f_k = (N_k - n_k)/N_k$ denotes the finite population correction factor for stratum k.

6.3. HB estimation in stratified sampling

In this section, we will consider a hierarchical model to develop HB estimates of stratum means. The hierarchical model, given later, adds another stage to the Model 6.1 by assigning a prior distribution to the unknown mean ν and variance parameters σ^2 and τ^2. We reparameterize $\sigma^2 = r^{-1}$ and $\tau^2 = (r\lambda)^{-1}$. We assign a uniform prior distribution to ν and independent inverse gamma distributions to σ^2 and τ^2. We use the notation $Gamma(a/2, b/2)$ to denote a gamma distribution with mean b/a and variance b/a^2. We also note that a p-variate t-ditribution with location vector $\boldsymbol{a}$, positive definite scale matrix $\boldsymbol{\Sigma}$ and degrees of freedom d is given by the pdf

$$f(\boldsymbol{x}) \propto [1 + (\boldsymbol{x} - \boldsymbol{a})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{a})/d]^{-(d+p)/2}, \quad \boldsymbol{x} \in \mathcal{R}^p.$$

MODEL 6.3.

I. Conditional on $\theta_1, \ldots, \theta_m$, ν, σ^2 and τ^2, for $j = 1, \ldots, N_k$, $k = 1, \ldots, m$, y_{kj} are independently normally distributed with $E(y_{kj}|\theta_1, \ldots, \theta_m) = \theta_k$ and $V(y_{kj}|\theta_1, \ldots, \theta_m) = \sigma^2$;

II. Conditional on ν, σ^2, and τ^2, $\theta_1, \ldots, \theta_m$ are iid $N(\nu, \tau^2)$;

III. Marginally, ν, σ^2, and τ^2 are independently distributed with $\nu \sim$ uniform $(-\infty, \infty)$, $(\sigma^2)^{-1} \sim Gamma(a/2, b/2)$, and $(\tau^2)^{-1} \sim Gamma(c/2, d/2)$.

The following theorem provides the predictive distribution of $\boldsymbol{y}(s^c)$ given $\boldsymbol{z}$. A proof of this theorem, omitted here, is available in Ghosh and Meeden (1997, pp.227–235) or Datta and Ghosh (1991, p. 1754).

THEOREM 6.1. *Under the hierarchical model mentioned earlier, the predictive distribution of $\boldsymbol{y}(s^c)$ given $\boldsymbol{z}$ is given in two steps.*

(i) conditional on λ and $\boldsymbol{z}$, the joint distribution of $\boldsymbol{y}(s^c)$ is multivariate-t with location vector

$$\left[\oplus_{i=1}^m \mathbf{1}_{N_i - n_i}\right] [(1 - B_1(\lambda))\bar{z}_1 + B_1(\lambda)\nu(\lambda), \ldots, (1 - B_m(\lambda))\bar{z}_m + B_m(\lambda)\nu(\lambda)]^T,$$

degrees of freedom $n_T + b + d - 1$ and scale matrix

$$(n_T + b + d - 1)^{-1}[a + c\lambda + Q_0(\lambda)]\boldsymbol{G}(\lambda),$$

where

$$Q_0(\lambda) = \sum_{k=1}^m \sum_{j=1}^{n_k} (z_{kj} - \bar{z}_k)^2 + \lambda \sum_{k=1}^m (1 - B_k(\lambda))(\bar{z}_k - \nu(\lambda))^2,$$

and

$$\boldsymbol{G}(\lambda) = \oplus_{i=1}^m \left[\mathbf{I}_{N_i - n_i} + (\lambda + n_i)^{-1} \mathbf{J}_{N_i - n_i}\right] + \left[\sum_{i=1}^m (1 - B_i(\lambda))\right]^{-1} \left[B_1(\lambda)\mathbf{1}^T_{N_1 - n_1}, \ldots, B_m(\lambda)\mathbf{1}^T_{N_m - n_m}\right]^T \times \left[B_1(\lambda)\mathbf{1}^T_{N_1 - n_1}, \ldots, B_m(\lambda)\mathbf{1}^T_{N_m - n_m}\right];$$

(ii) the conditional pdf of λ given $\mathbf{z}$ is given by the pdf

$$f(\lambda|z) \propto \prod_{k=1}^{m} B_k(\lambda)^{1/2} \left\{ \sum_{k=1}^{m} (1 - B_k(\lambda)) \right\}^{-1/2} \times \{a + c\lambda + Q_0(\lambda)\}^{-\frac{1}{2}(n_T+b+d-1)}.$$

From Theorem 6.1, the HB predictor of $\boldsymbol{\mu}$ is given by $\hat{\boldsymbol{\mu}}_{\mathrm{HB}} = (\hat{\mu}_{1,\mathrm{HB}}, \ldots, \hat{\mu}_{m,\mathrm{HB}})^T$, where

$$\hat{\mu}_{k,\mathrm{HB}} = \bar{z}_k - f_k E[B_k(\lambda)(\bar{z}_k - \nu(\lambda))|z], \tag{76}$$

$k = 1, \ldots, m$. Also, the posterior variance of μ_k is given by

$$\begin{aligned} V(\mu_k|z) &= E[V(\mu_k|\lambda, z)|z] + V[E(\mu_k|\lambda, z)|z] \\ &= E\big[(n_T + b + d - 3)^{-1}\{a + c\lambda + Q_0(\lambda)\}G_{kk}(\lambda)|z\big] \\ &\quad + f_k^2 V[B_k(\lambda)(\bar{z}_k - \nu(\lambda))|z]. \end{aligned} \tag{77}$$

Similarly, the posterior covariance between μ_i and μ_k for $i \neq k$ is given by

$$\mathrm{Cov}(\mu_i, \mu_k|z) = f_i f_k \mathrm{cov}[B_i(\lambda)(\bar{z}_i - \nu(\lambda)), B_k(\lambda)(\bar{z}_k - \nu(\lambda))|z]. \tag{78}$$

Using (76)–(78), an HB predictor of μ is given by $\hat{\mu}_{\mathrm{HB}} = N^{-1} \sum_{k=1}^{m} N_k \hat{\mu}_{k,\mathrm{HB}}$ and the associated posterior variance is given by

$$V(\mu|z) = N^{-2} \left[\sum_{k=1}^{m} N_k^2 V(\mu_k|z) + \sum_{i \neq k} N_i N_k \mathrm{Cov}(\mu_i, \mu_k|z) \right].$$

To evaluate the HB estimates and the posterior variances we usually need numerical integration method.

7. Generalized linear models

The hierarchical and empirical Bayes estimation techniques discussed in the previous chapters has mainly concentrated on continuous-valued variates. Often the survey data are discrete or categorical, for which the HB or EB analysis suitable for continuous variates is not appropriate. In the past few years, work has begun to appear on the Bayesian analysis of discrete survey data. Dempster and Tomberlin (1980) and MacGibbon and Tomberlin (1989) obtained small area estimates of proportions via EB techniques, whereas Malec et al. (1993) found the predictive distributions of a linear combination of binary random variables using a HB technique. Stroud (1991) developed a general HB methodology for binary data, whereas Nandram and Sedransk (1993) suggested Bayesian predictive inference for binary data from a two-stage cluster sample. Subsequently, Stroud (1994) provided a comprehensive treatment of binary survey data encompassing simple random, stratified, cluster and two-stage sampling, as well as two-stage sampling within strata.

The binary models constitute a subclass of generalized linear models that are often used for a unified analysis of both discrete and continuous data. Ghosh et al. (1998) and Ghosh and Natarajan (1999) developed HB generalized linear models with applications

to small area estimation. On the other hand, Raghunathan (1993) suggested a quasi-empirical Bayes method to address small area problems.

The HB model considered by Ghosh et al. (1998) is as follows. Suppose there are m strata or local areas. Let Y_{ik} denote the minimal sufficient statistic (discrete or continuous) for the kth unit within the ith stratum ($k = 1, \ldots, n_i; i = 1, \ldots, m$). The Y_{ik} are assumed to be conditionally independent with pdf

$$f(y_{ik}|\theta_{ik}, \phi_{ik}) = exp\big[\phi_{ik}^{-1}(y_{ik}\theta_{ik} - \psi(\theta_{ik})) + \rho(y_{ik}; \phi_{ik})\big] \tag{79}$$

($k = 1, \ldots, n_i; i = 1, \ldots, m$). Such a model is referred to as a generalized linear model (McCullagh and Nelder, 1989, p. 28). The density (79) is parameterized with respect to the canonical parameters θ_{ik} and the scale parameters $\phi_{ik} (> 0)$. It is assumed that the scale parameters ϕ_{ik} are known.

The natural parameters θ_{ik} are first modeled as

$$h(\theta_{ik}) = \boldsymbol{b}_{ik}^T\boldsymbol{\beta} + u_i + \epsilon_{ik} \quad (k = 1, \ldots, n_i;\ i = 1, \ldots, m), \tag{80}$$

where h is a strictly increasing function; the $\boldsymbol{x}_{ik}$ $(p \times 1)$ are known design vectors, $\boldsymbol{\beta}$ $(p \times 1)$ is the unknown regression coefficient, u_i are the random effects, and ϵ_{ik} are the errors. It is assumed that the u_i and the ϵ_{ik} are mutually independent with $u_i \overset{iid}{\sim} N\big(0, \sigma_u^2\big)$ and $\epsilon_{ik} \overset{iid}{\sim} N(0, \sigma^2)$.

It is possible to represent (79) and (80) in a hierarchical framework. Let $r_u = \sigma_u^{-2}$ and $r = \sigma^{-2}$. Also, let $\boldsymbol{\theta} = (\theta_{11}, \ldots, \theta_{1n_1}, \ldots, \theta_{m1}, \ldots, \theta_{mn_m})^T$ and $\boldsymbol{u} = (u_1, \ldots, u_m)^T$. Then the hierarchical model is given by

(I) conditional on $\boldsymbol{\theta}$, $\boldsymbol{\beta}$, $\boldsymbol{u}$, r_u, and r, Y_{ik} are independent with densities given in (79);

(II) conditional on $\boldsymbol{\beta}$, $\boldsymbol{u}$, r_u, and r, $h(\theta_{ik}) \overset{ind}{\sim} \mathrm{N}(\boldsymbol{x}_{ik}^T\boldsymbol{\beta} + u_i\ ,\ r^{-1})$;

(III) conditional on $\boldsymbol{\beta}$, r_u, and r, $u_i \overset{ind}{\sim} \mathrm{N}(0, r_u^{-1})$.

To complete the hierarchical model, we assign the following prior to $\boldsymbol{\beta}$, r_u and r:

(IV) $\boldsymbol{\beta}$, r_u, and r are mutually independent with $\boldsymbol{\beta} \sim$ uniform$(\boldsymbol{R}^p)$, $(p < m)$, $r_u \sim$ Gamma$\big(\frac{1}{2}a, \frac{1}{2}b\big)$, and $r \sim$ Gamma$\big(\frac{1}{2}c, \frac{1}{2}d\big)$.

The main objective is to find the joint posterior distribution of $g(\boldsymbol{\theta}_{ik})$ where g is a strictly increasing function, given the data $\boldsymbol{y} = (y_{11}, \ldots, y_{1n_1}, \ldots, y_{m1}, \ldots, y_{mn_m})^T$, and in particular in finding the posterior means, variances, and covariances of these parameters. In typical applications, $g(\theta_{ik}) = \psi'(\theta_{ik}) = E(Y_{ik}|\theta_{ik})$.

First, however, one needs to ensure that the joint posterior distribution of θ_{ik} given $\boldsymbol{y}$ is proper. A theorem is proved to this effect. In what follows, the support of θ_{ik} is the open interval $(\underline{\theta}_{ik}, \overline{\theta}_{ik})$, where the lower endpoint of the interval can be $-\infty$, the upper endpoint can be $+\infty$, or both.

A few notations are needed before stating the theorem. Let $I_{ik} = \int_{\underline{\theta}_{ik}}^{\overline{\theta}_{ik}} f(y_{ik}|\theta_{ik})h'(\theta_{ik})d\theta_{ik}$, and $S = \{(i, k)|I_{ik} < \infty\}$. We denote by $\boldsymbol{\theta}_*$ the vector of θ_{ik} which belong to S. The cardinality of S is denoted by s. Also, let $\boldsymbol{X}_*^T$ denote the matrix consisting of the column vectors $\boldsymbol{x}_{ik}$, where $(i, k) \in S$. Further, let m_* denote the number of local areas i that have at least one unit k for which $(i, k) \in S$. We now have the following theorem.

THEOREM 7.1. *Assume (i)* $a > 0,\ c > 0$, *(ii)* $s \geq p,\ s + b > p$, *(iii)* $\text{rank}(\boldsymbol{X}_*) = p$, *(iv)* $m_* + b > 0$, *and (v)* $f(y_{ik}|\theta_{ik})$ *is bounded for all* (i, k). *Then the joint posterior of* θ_{ik}*'s given* $\boldsymbol{y}$ *is proper.*

This theorem is a stronger version of the one given in Ghosh et al. (1998).

Two special cases are of interest. In the first case, $Y_{ik}|\theta_{ik} \sim Bin(n_{ik}, \exp(\theta_{ik})/(1+\exp(\theta_{ik})))$. Suppose now h is the identity function, that is, the link is canonical. Then writing $w_{ik} = \exp(\theta_{ik})/[1 + \exp(\theta_{ik})]$, the condition that $I_{ik} < \infty$ reduces to $\int_0^1 w_{ik}^{y_{ik}-1}(1 - w_{ik})^{n-y_{ik}-1} dw_{ik} < \infty$. As long as this happens for p pairs (i, k), and the other conditions of the theorem hold, that is the prior is not too ill-behaved, one has the propriety of the posterior. In particular, when $p = 1$, $I_{ik} < \infty$ for one pair (i, k) amounts only to the requirement that not all outcomes are either successes or failures. In the second case, $Y_{ik}|\theta_{ik} \sim \text{Poisson}(\exp(\theta_{ik}))$. Then, if h is the canonical link, the condition that $I_{ik} < \infty$ reduces to $\int_0^\infty \zeta_{ik}^{y_{ik}-1} \exp(-\zeta_{ik}) d\zeta_{ik} < \infty$. This condition needs to hold once again only for p pairs (i, k). For $p = 1$, all one requires is that y_{ik} is not zero for at least one (i, k).

Direct evaluation of the joint posterior distribution of $g(\theta_{ik})$ given $\boldsymbol{y}$ involves high-dimensional numerical integration, and is not computationally feasible. Instead, we use the Gibbs sampler (Gelfand and Smith, 1990). Its implementation requires generating samples from certain conditional posterior distributions. Write $\boldsymbol{h}(\boldsymbol{\theta}) = (h(\theta_{11}), \ldots, h(\theta_{1n_1}), \ldots, h(\theta_{m1}), \ldots, h(\theta_{mn_m}))^T$, and $\boldsymbol{X} = (\boldsymbol{x}_{11}, \ldots, \boldsymbol{x}_{1n_1}, \ldots, \boldsymbol{x}_{m1}, \ldots, \boldsymbol{x}_{mn_m})^T$. Assume $\boldsymbol{X}^T\boldsymbol{X}$ is nonsingular. The necessary conditional distributions based on (I)–(IV) are as follows:

(i) $\boldsymbol{\beta}|\boldsymbol{\theta}, \boldsymbol{u}, r_u, r, \boldsymbol{y} \sim N\big((\boldsymbol{X}^T\boldsymbol{X})^{-1}\big(\boldsymbol{X}^T\boldsymbol{h}(\boldsymbol{\theta}) - \sum_i u_i \sum_k \boldsymbol{x}_{ik}\big),\ r^{-1}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\big)$;

(ii) $u_i|\boldsymbol{\theta}, \boldsymbol{\beta}, r_u, r, \boldsymbol{y} \overset{ind}{\sim} N\left((rn_i + r_u)^{-1} r \sum_k \left(h(\theta_{ik}) - \boldsymbol{x}_{ik}^T\boldsymbol{\beta}\right),\ (rn_i + r_u)^{-1}\right)$;

(iii) $r|\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{u}, r_u, \boldsymbol{y} \sim \text{Gamma}\Big(\frac{1}{2}\Big(c + \sum_i \sum_k \left(h(\theta_{ik}) - \boldsymbol{x}_{ik}^T\boldsymbol{\beta} - u_i\right)^2\Big),\ \frac{1}{2}\big(d + \sum_1^m n_i\big)\Big)$;

(iv) $r_u|\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{u}, r, \boldsymbol{y} \sim \text{Gamma}\big(\frac{1}{2}\big(a + \sum_i u_i^2\big),\ \frac{1}{2}\big(b + \sum_1^m n_i\big)\big)$;

(v) $\pi(\theta_{ik}|\theta_{jl},\ (j, l) \neq (i, k), \beta, \boldsymbol{u}, r_u, r, \boldsymbol{y})$

$$\propto \exp\left[(y_{ik}\theta_{ik} - \psi(\theta_{ik}))\phi_{ik}^{-1} - \frac{r}{2}(h(\theta_{ik}) - \boldsymbol{x}_{ik}^T\boldsymbol{\beta} - u_i)^2\right] h'(\theta_{ik}).$$

It is easy to generate samples from the normal and gamma distributions given in (i)–(iv). On the other hand, as evidenced in (v), the posterior distribution of θ_{ik} given $\boldsymbol{\beta}, \boldsymbol{u}, r_u, r$, and $\boldsymbol{y}$ is known only up to a multiplicative constant, and accordingly one has to use a general accept–reject algorithm to generate samples from this pdf. In the special case where h is the identity function, the task becomes much simpler due to the following lemma that establishes log-concavity of $\pi(\theta_{ik}|\boldsymbol{\beta}, \boldsymbol{u}, r_u, r, \boldsymbol{y})$. In such cases, one can use the adaptive rejection sampling scheme of Gilks and Wild (1992).

LEMMA 7.2. *When* $h(z) = z$ *for all* z, $\log \pi(\theta_{ik}|\boldsymbol{\beta}, \boldsymbol{u}, r, r_u, \boldsymbol{y})$ *is a concave function of* θ_{ik}.

PROOF. $\log \pi(\theta_{ik}|\boldsymbol{\beta}, \boldsymbol{u}, r, r_u, \boldsymbol{y}) = -\phi_{ik}^{-1}\psi''(\theta_{ik}) - \frac{r}{2}$, and the result follows since $V(Y_{ik}|\theta_{ik}) = \phi_{ik}\psi''(\theta_{ik})$ and $\phi_{ik} > 0$. □

Inference for $\boldsymbol{\theta}$ will be based on (i)–(v). Indeed, based on (v), one can also find $E(\theta_{ik}|\boldsymbol{y})$, $V(\theta_{ik}|\boldsymbol{y})$, and $\mathrm{Cov}(\theta_{ik}, \theta_{i'k'}|\boldsymbol{y})$ $(i,k) \neq (i',k')$ based on Monte Carlo integration techniques and formulas for iterated conditional expectations and variances.

The aforementioned method is different from that of Albert (1988), applied to binary survey data by Stroud (1994). Albert's method when applied to the present setting first uses independent conjugate priors

$$\pi(\theta_{ik}|m_{ik}, \zeta) \;=\; \exp\left[\zeta(m_{ik}\theta_{ik} - \psi(\theta_{ik})) + g(m_{ik}; \zeta)\right] \tag{81}$$

for the θ_{ik}. Next, he assumes $h(m_{ik}) = \boldsymbol{x}_{ik}^T\boldsymbol{\beta}$ for some known monotone function h. Subsequently, he assigns distributions (possibly diffuse) to the hyperparameters $\boldsymbol{\beta}$ and ζ. In contrast, the present HB model does not need the conjugacy of the prior, and models monotone functions of θ_{ik} instead of monotone functions of $m_{ik} = \mathrm{E}[\psi'(\theta_{ik})]$. Moreover, Albert (1988) suggests approximation to the Bayes procedure by one or the other of the following methods: (i) Laplace's method, (ii) quasi-likelihood approaches. These approximations are not utilized here. Instead, the MCMC numerical integration technique will be used.

The log-concavity idea is used slightly differently in Dellaportas and Smith (1993) where the prime objective is inference about $\boldsymbol{\beta}$ in generalized linear models, and θ_{ik} are modeled as functions of $\boldsymbol{\beta}$ without any error. In addition, their method does not include the ϵ_{ik}, the uncertainty in specifying the model.

We now examine how the previous results can be generalized for the analysis of multicategory data. Consider m strata labeled $1, \ldots, m$. Within each stratum, several units are selected, and suppose that the responses of individuals within each selected unit are independent, and can be classified into J categories. For the kth selected unit within the ith stratum, let p_{ijk} denote the probability that an individual's response falls in the jth category ($j = 1, \ldots, J$; $k = 1, \ldots, n_i$). Then within the kth selected unit within the ith stratum, Z_{ijk} $(j = 1, \ldots, J)$ have a joint multinomial $(t_{ik};\ p_{i1k}, \ldots, p_{iJk})$ distribution, where $t_{ik} = \sum_j Z_{ijk}$. Using the well-known relationship between the multinomial and Poisson distributions, $(Z_{i1k}, \ldots, Z_{iJk})$ has the same distribution as the joint conditional distribution of $(Y_{i1k}, \ldots, Y_{iJk})$ given $\sum_{j=1}^J Y_{ijk} = t_{ik}$ where the Y_{ijk} $(j = 1, \ldots, J)$ are independent Poisson(ζ_{ijk}) and $p_{ijk} = \zeta_{ijk} / \sum_{j=1}^J \zeta_{ijk}$ $(j = 1, \ldots, J)$.

Let $\theta_{ijk} = \log \zeta_{ijk}$, and let $\boldsymbol{\theta}$ be the vector whose elements are θ_{ijk}. One can also model θ_{ijk} as

$$h(\theta_{ijk}) \;=\; \boldsymbol{x}_{ijk}^T\boldsymbol{\beta} \;+\; u_{ij} \;+\; \epsilon_{ijk}. \tag{82}$$

Also, it is assumed that u_{ij} and the ϵ_{ijk} are mutually independent with $u_{ij} \overset{iid}{\sim} N\big(0, \sigma_u^2\big)$ and $\epsilon_{ijk} \overset{iid}{\sim} N(0, \sigma^2)$.

Then the hierarchical model (which is closely related to (I)–(IV)) is given by

(A) $Y_{ijk}|\boldsymbol{\theta}, \boldsymbol{u}, \boldsymbol{\beta}, r_u, r$ are independent with

$$f(y_{ijk}|\boldsymbol{\theta}, \boldsymbol{u}, \boldsymbol{\beta}, r_u, r) = \exp\left[\phi_{ijk}^{-1}(y_{ijk}\theta_{ijk} - \psi(\theta_{ijk})) + \rho(y_{ijk}; \phi_{ijk})\right];$$

(B) $h(\theta_{ijk})|\boldsymbol{u}, \boldsymbol{\beta}, r_u, r \overset{ind}{\sim} \mathrm{N}\Big(\boldsymbol{x}_{ijk}^T\boldsymbol{\beta} + u_{ij}, r^{-1}\Big)$;

(C) $u_{ij}|\boldsymbol{\beta}, r_u, r \overset{ind}{\sim} \mathrm{N}\big(0,\ r_u^{-1}\big)$;

(D) $\boldsymbol{\beta}, r_u$, and r are mutually independent with $\boldsymbol{\beta} \sim \text{uniform}(\boldsymbol{R}^p)$, $r_u \sim \text{Gamma}\left(\frac{1}{2}a, \frac{1}{2}b\right)$, and $r \sim \text{Gamma}\left(\frac{1}{2}c, \frac{1}{2}d\right)$.

We are interested in the posterior means, variance, and covariances of the $p_{ijk} = exp(\theta_{ijk})/\sum_{j=1}^{J} \exp(\theta_{ijk})$ $(k = 1, \ldots, n_i;\ i = 1, \ldots, m;\ j = 1, \ldots, J)$. The necessary posterior distributions for doing these calculations are given by

(a) $\boldsymbol{\beta}|\boldsymbol{\theta}, \boldsymbol{u}, r_u, r, \boldsymbol{y} \sim$
$\mathrm{N}\left(\left(\sum_{i,j,k} \boldsymbol{x}_{ijk}\boldsymbol{x}_{ijk}^T\right)^{-1}\left(\sum_{i,j,k} \boldsymbol{x}_{ijk}(h(\theta_{ijk}) - u_{ij})\right), r^{-1}\left(\sum_{i,j,k} \boldsymbol{x}_{ijk}\boldsymbol{x}_{ijk}^T\right)^{-1}\right)$;

(b) $u_{ij}|\boldsymbol{\theta}, \boldsymbol{\beta}, r_u, r, \boldsymbol{y} \overset{ind}{\sim} \mathrm{N}\left((rn_i + r_u)^{-1} r \sum_k \left(h(\theta_{ijk}) - \boldsymbol{x}_{ijk}^T\boldsymbol{\beta}\right),\ (rn_i + r_u)^{-1}\right)$;

(c) $r|\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{u}, r_u, \boldsymbol{y} \sim \text{Gamma}\left(\frac{1}{2}\left(c + \sum_{i,j,k}\left(h(\theta_{ijk}) - \boldsymbol{x}_{ijk}^T\boldsymbol{\beta} - u_{ij}\right)^2\right),\ \frac{1}{2}\left(d + J\sum_i n_i\right)\right)$;

(d) $r_u|\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{u}, r, \boldsymbol{y} \sim \text{Gamma}\left(\frac{1}{2}\left(a + \sum_i \sum_j u_{ij}^2\right), \frac{1}{2}(b + mJ)\right)$;

(e) $\pi(\theta_{ijk}|\boldsymbol{\beta}, \boldsymbol{u}, r_u, r, \boldsymbol{y})$
$\propto \exp\left[(y_{ijk}\theta_{ijk} - \psi(\theta_{ijk}))\phi_{ijk}^{-1} - \frac{r}{2}\left(h(\theta_{ijk}) - \boldsymbol{x}_{ijk}^T\boldsymbol{\beta} - u_{ij}\right)^2\right] h'(\theta_{ijk})$.

Once again posterior inference about $g(\theta_{ijk})$'s is performed using (e) and iterated formulas for posterior moments.

The discussion so far has concentrated on small area estimation based on specific HB generalized linear models. Often, there are situations where there is no clearcut choice among several models. In such situations, one can find the posterior probabilities of the different models, pick the one with the highest posterior probability and find small area estimates and standard errors based on that model. Another option is not to report estimates and standard errors based on a single model, but report estimates which are weighted averages of estimates based on the different contemplated models, the respective weights being proportional to the posterior probabilities of these models. Similar views are expressed, for example, in Raftery (1996). This method, being adaptive in nature, has intrinsic appeal, especially in situations when one particular model does not outperform the rest. Moreover, in finding the standard errors associated with the small area estimates, there is an extra layer of uncertainty due to the choice of models. This results in larger standard errors associated with the estimates, but the procedure seems worthwhile especially when none of the contemplated models emerges as a clearcut winner.

To be specific, suppose there are K contemplated models labeled $M_1, \cdots, M_K$. Suppose the data is $\boldsymbol{y}$ and the parameter of interest is Δ. Then

$$P(\Delta|\boldsymbol{y}) = \sum_{k=1}^{K} P\left(\Delta|\boldsymbol{y}, M_k\right) P(M = M_k|\boldsymbol{y}). \tag{83}$$

This leads to

$$E(\Delta|\boldsymbol{y}) = \sum_{k=1}^{K} E(\Delta|\boldsymbol{y}, M_k) P(M = M_k|\boldsymbol{y}); \tag{84}$$

$$
\begin{aligned}
V(\Delta|\boldsymbol{y}) &= E(\Delta^2|\boldsymbol{y}) - (E(\Delta|\boldsymbol{y}))^2 \\
&= \sum_{k=1}^{K} E(\Delta^2|\boldsymbol{y}, M_k) P(M = M_k|\boldsymbol{y}) - [E(\Delta|\boldsymbol{y})]^2 \\
&= \sum_{k=1}^{K} V(\Delta|\boldsymbol{y}, M_k) P(M = M_k|\boldsymbol{y}) \\
&\quad + \sum_{k=1}^{K} (E(\Delta|\boldsymbol{y}, M_k))^2 P(M = M_k|\boldsymbol{y}) - E(\Delta|\boldsymbol{y})^2. \qquad (85)
\end{aligned}
$$

Clearly, the first term in the right-hand side of (85) represents the expectation of the conditional variance of Δ given the data and the model, whereas the second and the third terms when combined represents the variance of the conditional expectation of Δ given the data and the model. Raftery (1996) contains a discussion of multiple models.

We present here the idea of mixing the models using the general description of the HB GLMs and later illustrate this idea with an example. For simplicity of the discussion we assume there are two possible models labeled M_1 and M_2.

Let Y_{ik} denote the minimal sufficient statistic (discrete or continuous) for the kth unit within the ith stratum($k = 1, \ldots, n_i$; $i = 1, \ldots, m$) and Y_{ik} are assumed to be conditionally independent with pdf

$$
f(y_{ik} \mid \theta_{ik}) \;=\; \exp\left[\phi_{ik}^{-1}(y_{ik}\theta_{ik} - \psi(\theta_{ik})) + \rho(y_{ik}; \phi_{ik})\right] \qquad (86)
$$

($k = 1, \ldots, n_i$; $i = 1, \ldots, m$). Under model M_1 the canonical parameter θ_{ik} is modeled as

$$
\theta_{ik} = \boldsymbol{x}_{ik}^{(1)^T} \boldsymbol{b}_1^{(M_1)} + \boldsymbol{x}_{ik}^{(2)^T} \boldsymbol{b}_2^{(M_1)} + u_i + \epsilon_{ik}, \qquad (87)
$$

and under model M_2, θ_{ik} is modeled as

$$
\theta_{ik} = \boldsymbol{x}_{ik}^{(1)^T} \boldsymbol{b}_1^{(M_2)} + u_i + \epsilon_{ik}, \qquad (88)
$$

where u_i and ϵ_{ik} are mutually independent with u_i iid $N(0, r_u^{-1})$, while ϵ_{ik} iid $N(0, r^{-1})$.

Notice that it is important to distinguish between $\boldsymbol{b}_1^{(M_1)}$ and $\boldsymbol{b}_1^{(M_2)}$ in the two models, as they carry different interpretations. However, either model will cause estimates of the θ_{ik} to borrow strength from other strata, as well as other cells within a given stratum.

Since, this case involves finding posterior probabilities or the Bayes factor of the two models along with small area estimates, if one assigns a diffuse prior to the regression coefficients, then the posterior distribution of M (the indicator variable for the model) given the data becomes improper. Hence, one needs to assign a proper prior to $\boldsymbol{b}$.

The full hierarchical model is described in the following sections. For model M_1,

(I) conditional on θ, $\boldsymbol{b}_1^{(M_1)}$, $\boldsymbol{b}_2^{(M_1)}$, $\boldsymbol{u}$, $R_u = r_u$, and $R = r$, Y_{ik} are independent with pdf (86);

(II) conditional on $\boldsymbol{b}_1^{(M_1)}$, $\boldsymbol{b}_2^{(M_1)}$, $\boldsymbol{u}$, r_u, and r,

$$
\theta_{ik} \overset{ind}{\sim} N\left(\boldsymbol{x}_{ik}^{(1)^T} \boldsymbol{b}_1^{(M_1)} + \boldsymbol{x}_{ik}^{(2)^T} \boldsymbol{b}_2^{(M_1)} + u_i, r^{-1}\right);
$$

(III) conditional on $\boldsymbol{b}_1^{(M_1)}$, $\boldsymbol{b}_2^{(M_1)}$, r_u, and r, $u_i \overset{ind}{\sim} N\left(0, r_u^{-1}\right)$;

(IV) $\boldsymbol{b}_1^{(M_1)}, \boldsymbol{b}_2^{(M_1)}, r_u$, and r are mutually independent with

$$\boldsymbol{b}^{(M_1)} = \begin{pmatrix} \boldsymbol{b}_1^{(M_1)} \\ \boldsymbol{b}_2^{(M_1)} \end{pmatrix} \sim N\left(\begin{pmatrix} \boldsymbol{0} \\ \boldsymbol{0} \end{pmatrix}, \begin{pmatrix} \eta_1^2 \boldsymbol{I} & \boldsymbol{0} \\ \boldsymbol{0} & \eta_2^2 \boldsymbol{I} \end{pmatrix}\right),$$

$r_u \sim \text{Gamma}\left(\frac{1}{2}a, \frac{1}{2}g\right)$ and $\text{r} \sim \text{Gamma}\left(\frac{1}{2}c, \frac{1}{2}d\right)$.

We have chosen η_1^2 to be large and η_2^2 to be small, to reflect the strong belief that $\boldsymbol{b}_2^{(M_1)}$ is close to 0 but attach a small amount of uncertainity to this belief. The near diffuseness of the prior on $\boldsymbol{b}_1^{(M_1)}$ reflects the vagueness in its choice in conformity with earlier models.

Model M_2 sets $\boldsymbol{b}_2^{(M_2)} = \boldsymbol{0}$. Other than that, the remainder of the Model M_2 remains the same as in Model M_1. Also, we assign $P(M = M_1) = \pi$.

On notations, superscripts $M_j,\ j = 1, 2$ for $\boldsymbol{u},\ r_u$, and r indicates that samples were observed from their respective full conditional distributions based on model M_j, j=1,2.

The full conditional distributions based on model M_1 are given by

(i) $\boldsymbol{b}^{(M_1)}|\boldsymbol{\theta}^{(M_1)}, \boldsymbol{u}^{(M_1)}, r_u^{(M_1)}, r^{(M_1)}$,
$\boldsymbol{y} \sim \text{N}\left(\left(r^{(M_1)}\boldsymbol{X}^T\boldsymbol{X} + \begin{pmatrix} \eta_1^{-2}\boldsymbol{I} & \boldsymbol{0} \\ \boldsymbol{0} & \eta_2^{-2}\boldsymbol{I} \end{pmatrix}\right)^{-1} r^{(M_1)}\boldsymbol{X}^T\left(\boldsymbol{\theta}^{(M_1)} - \Sigma_i u_i^{(M_1)}\Sigma\boldsymbol{x}_{ik}\right),\right.$
$\left.\left(r^{(M_1)}\boldsymbol{X}^T\boldsymbol{X} + \begin{pmatrix} \eta_1^{-2}\boldsymbol{I} & \boldsymbol{0} \\ \boldsymbol{0} & \eta_2^{-2}\boldsymbol{I} \end{pmatrix}\right)^{-1}\right)$;

(ii) $u_i^{(M_1)}|\boldsymbol{\theta}^{(M_1)}, \boldsymbol{b}^{(M_1)}, r_u^{(M_1)}, r^{(M_1)}, \boldsymbol{y} \sim \text{N}\left(\left(r^{(M_1)}n_i + r_u^{(M_1)}\right)^{-1} r^{(M_1)}\sum_k\left(\theta_{ik}^{(M_1)} - \boldsymbol{x}_{ik}^T\boldsymbol{b}^{(M_1)}\right), \left(r^{(M_1)}n_i + r_u^{(M_1)}\right)^{-1}\right)$;

(iii) $r^{(M_1)}|\boldsymbol{\theta}^{(M_1)}, \boldsymbol{b}^{(M_1)}, \boldsymbol{u}^{(M_1)}, r_u^{(M_1)}, \boldsymbol{y} \sim \text{Gamma}\left(\frac{1}{2}\left(c + \sum_i\sum_k\left(\theta_{ik}^{(M_1)} - \boldsymbol{x}_{ik}^T\boldsymbol{b}^{(M_1)} - u_i^{(M_1)}\right)^2\right), \frac{1}{2}\left(d + \sum_1^m n_i\right)\right)$;

(iv) $r_u^{(M_1)}|\boldsymbol{\theta}^{(M_1)}, \boldsymbol{b}^{(M_1)}, \boldsymbol{u}^{(M_1)}, r^{(M_1)}, \boldsymbol{y} \sim \text{Gamma}\left(\frac{1}{2}\left(a + \boldsymbol{u}^{(M_1)^T}\boldsymbol{u}^{(M_1)}\right), \frac{1}{2}\left(g + \sum_1^m n_i\right)\right)$;

(v) $\theta_{ik}^{(M_1)}|\boldsymbol{b}^{(M_1)}, \boldsymbol{u}^{(M_1)}, r_u^{(M_1)}, r^{(M_1)}, \boldsymbol{y} \overset{ind}{\sim} \Pi_{i,k}(\theta_{ik}^{(M_1)}|\boldsymbol{b}^{(M_1)}, \boldsymbol{u}^{(M_1)}, r_u^{(M_1)}, r^{(M_1)}, \boldsymbol{y}) \propto \exp\left[\left\{y_{ik}\theta_{ik}^{(M_1)} - \psi\left(\theta_{ik}^{(M_1)}\right)\right\}\phi_{ik}^{-1} - \frac{1}{2}r^{(M_1)}\left(\theta_{ik}^{(M_1)} - \boldsymbol{x}_{ik}^T\boldsymbol{b}^{(M_1)} - u_i^{(M_1)}\right)^2\right]$;

For the full conditionals based on M_2, replace M_1 by M_2, $\boldsymbol{X}$ by $\boldsymbol{X}^{(1)}$, $\boldsymbol{x}_{ik}$ by $\boldsymbol{x}_{ik}^{(1)}$, and $\begin{pmatrix} \eta_1^{-2}\boldsymbol{I} & \boldsymbol{0} \\ \boldsymbol{0} & \eta_2^{-2}\boldsymbol{I} \end{pmatrix}$ by $\eta_1^{-2}\boldsymbol{I}$.

Finally, the full conditional for the model indicator variable M is given by

$$P\Big(M = M_1|\boldsymbol{\theta}^{(M_1)}, \boldsymbol{\theta}^{(M_2)}, \boldsymbol{b}_1^{(M_1)}, \boldsymbol{b}_2^{(M_1)}, \boldsymbol{b}_1^{(M_2)}, \boldsymbol{u}^{(M_1)}, r_u^{(M_1)}, r^{(M_1)}, \boldsymbol{u}^{(M_2)}, r_u^{(M_2)}, r^{(M_2)}, \boldsymbol{y}\Big)$$

$$= \pi \exp\left[-\frac{r^{(M_1)}}{2} \sum_i \sum_k \left(\theta_{ik}^{(M_1)} - \boldsymbol{x}_{ik}^T \boldsymbol{b}^{(M_1)} - u_i^{(M_1)} \right)^2 - \frac{\eta_1^{-2}}{2} ||\boldsymbol{b}_1^{(M_1)}||^2 - \frac{\eta_2^{-2}}{2} ||\boldsymbol{b}_2^{(M_1)}||^2 \right]$$

$$\div \left\{ \pi \exp\left[-\frac{r^{(M_1)}}{2} \sum_i \sum_k \left(\theta_{ik}^{(M_1)} - \boldsymbol{x}_{ik}^T \boldsymbol{b}^{(M_1)} - u_i^{(M_1)} \right)^2 - \frac{\eta_1^{-2}}{2} ||\boldsymbol{b}_1^{(M_1)}||^2 - \frac{\eta_2^{-2}}{2} ||\boldsymbol{b}_2^{(M_1)}||^2 \right] + (1-\pi) \exp\left[-\frac{r^{(M_2)}}{2} \sum_i \sum_k \left(\theta_{ik}^{(M_2)} - \boldsymbol{x}_{ik}^{(1)T} \boldsymbol{b}_1^{(M_2)} - u_i^{(M_2)} \right)^2 - \frac{\eta_1^{-2}}{2} ||\boldsymbol{b}_1^{(M_2)}||^2 \right] \right\}.$$

Ghosh et al. illustrated this idea with a dataset related to job satisfaction based on 1981 survey of employees of a large national corporation.

Raghunathan (1993), on the other hand, advocated a quasi-likelihood approach that does not require any distributional assumptions, but involves specification of the first two sample moments in terms of certain parameters. A quasi-likelihood can be constructed based on these parameters. The next step is to model the means and variances of these prior means in terms of some other parameters, and subsequently generate a quasi prior density. Estimation of the prior parameters now takes place on the basis of the score functions constructed from the marginal quasi-likelihood.

8. Summary

This chapter revisits the Bayesian developments in survey sampling, and traces much of the earlier history beginning with the sufficiency and likelihood principles. The Bayesian methods are illustrated primarily through the normal examples and its variants. Some duality between model- and design-based estimators commonly used in finite population sampling are also pointed out. This chapter discusses also at some length Bayesian methods in the presence of auxiliary information, stratification, and multistage sampling. Also, Bayesian inference based on generalized linear models is discussed in a hierarchical Bayesian setting.

There are many issues that are not addressed much in this chapter. A very important one is the role of sampling weights in survey inference. Other than a brief account in the beginning of Section 5 of viewing some of the popular estimators such as the Horvitz–Thompson estimator and the ratio estimator from a Bayesian angle, the chapter has not gone into an in-depth study of design-based estimators that make explicit use of survey weights. Although Bayesian inference typically does not recognize the need for these weights, their importance cannot but be underscored, especially when one is seeking some degree of robustness in inference, especially for protection against

model misspecification (see for example, Dumouchel and Duncan, 1983; Korn and Graubard, 1999; Little, 1991; Pfeffermann, 1993; Smith, 1988, among others). Little (1991, 2004) has demonstrated how to incorporate these weights in a model-based framework. More recently, there are attempts for design-assisted model-based small area estimation, (Prasad and Rao, 1999; Ghosh and Maiti, 2004), which can accomodate survey weights for inferential purposes. Pfeffermann et al. (1998b) have demonstrated how to use survey weights in multilevel models through Horvitz–Thompsonization of score functions in a very natural manner (see Chapters 23 and 24 of this volume).

Another very important issue is that of nonresponse. If the nonreponse is ignorable, then the Bayesian methods with some simple modifications can be directly used. However, nonignorable missingness requires further modeling, and the Bayesian approach is very natural in this case as is found in the classic text of Little and Rubin (2002) (see Chapter 8 of this Volume).

A final thing which is gaining momentum in recent years is confidentiality and disclosure of survey data (see Chapter 15 of this Volume). Bayesian methods have just started being applied in this context, but much remains to be done.

Acknowledgments

This research was partially supported by NSF Grants SES-9911485 and SES-0631426. I appreciate very much the reviewers' comments.

Sample Surveys: Inference and Analysis, Vol. 29B
ISSN: 0169-7161
© 2009 Elsevier B.V. All rights reserved
DOI: 10.1016/S0169-7161(09)00230-2

Empirical Likelihood Methods

J.N.K. Rao and Changbao Wu

1. Likelihood-based approaches

Let $U = \{1, 2, \ldots, N\}$ be the set of units in the finite population and y_i and $\boldsymbol{x}_i$ be, respectively, the values of the study variable y and the vector of the auxiliary variables $\boldsymbol{x}$ attached to the ith unit. In this chapter, we restrict our discussion to the estimation of the population total $Y = \sum_{i=1}^{N} y_i$, the population mean $\overline{Y} = Y/N$, or the population distribution function $F_N(t) = N^{-1} \sum_{i=1}^{N} I(y_i \le t)$ using a survey sample $\{(y_i, \boldsymbol{x}_i), i \in s\}$, where s is the set of sample units selected by the probability sampling design, $p(s)$, and $I(y \le t)$ is the indicator function. The population totals $\boldsymbol{X} = \sum_{i=1}^{N} \boldsymbol{x}_i$ or means $\overline{\boldsymbol{X}} = \boldsymbol{X}/N$ may also be available and can be used at the estimation stage.

Likelihood-based estimation methods in survey sampling do not follow as special cases from classical parametric likelihood inferences. Under the conventional design-based framework, values of the study variable for the finite population, $\{y_1, y_2, \ldots, y_N\}$, are viewed as fixed. The only randomization is induced by the probability sampling selection of units. In the design-based setup, an unbiased minimum variance estimator or even an unbiased minimum variance linear estimator of Y does not exist (Godambe, 1955; Godambe and Joshi, 1965). If we consider a class of linear estimators of Y in the form of $\sum_{i \in s} c_i y_i$ where the weight c_i depends only on i, then the unique unbiased estimator in the class is the well-known Horvitz–Thompson (HT) estimator $\hat{Y}_{\mathrm{HT}} = \sum_{i \in s} y_i/\pi_i$, where $\pi_i = P(i \in s)$ is the first-order inclusion probability for unit i. The HT estimator, therefore, is often treated as a baseline estimator for inferences concerning Y.

One of the early attempts in formulating a likelihood-based approach was the flat likelihood function (Godambe, 1966). The population vector of parameters is specified as $\tilde{\boldsymbol{y}} = (\tilde{y}_1, \tilde{y}_2, \ldots, \tilde{y}_N)'$, where the tilde indicates that each y_i is treated as an unknown parameter. The likelihood function of $\tilde{\boldsymbol{y}}$ is the probability of observing the sample data $\{y_i, i \in s\}$ for the given $\tilde{\boldsymbol{y}}$. For a given sampling design, we can write down the likelihood function as

$$L(\tilde{\boldsymbol{y}}) = P(y_i, i \in s | \tilde{\boldsymbol{y}}) = \begin{cases} p(s) & \text{if } \ y_i = \tilde{y}_i \ \text{ for } \ i \in s, \\ 0 & \text{otherwise,} \end{cases} \tag{1}$$

where $p(s)$ denotes the probability of selecting the sample s under the design. Although the likelihood function $L(\tilde{\boldsymbol{y}})$ is well defined, it is uninformative in the sense that all possible nonobserved values y_i, $i \notin s$ lead to the same likelihood. This difficulty arises because of the distinct labels i associated with the units in the sample data that make the sample unique.

To circumvent the difficulty associated with Godambe's flat likelihood, one possible resolution is to take a Bayesian route (Ericson, 1969). Given a joint N-dimensional prior on $\tilde{\boldsymbol{y}}$ with probability density function $g(\tilde{\boldsymbol{y}})$ and assume that the sampling design is independent of $\tilde{\boldsymbol{y}}$, the posterior density is given by

$$h(\tilde{\boldsymbol{y}}|y_i, i \in s) = \begin{cases} g(\tilde{\boldsymbol{y}})/g(\tilde{\boldsymbol{y}}_s) & \text{if } y_i = \tilde{y}_i \text{ for } i \in s, \\ 0 & \text{otherwise,} \end{cases}$$

where $\tilde{\boldsymbol{y}}_s = \{\tilde{y}_i, i \in s\}$ and $g(\tilde{\boldsymbol{y}}_s)$ is the marginal prior density of $\tilde{\boldsymbol{y}}_s$. Any informative prior will lead to an informative posterior distribution of $\tilde{\boldsymbol{y}}$ given the sample data. A popular choice of $g(\cdot)$ is the so-called exchangeable prior which basically states that the labels i carry no information regarding the associated y_i and the finite population is effectively randomized. However, in addition to the difficulty of choosing a prior, inferences under the Bayesian formulation are independent of the sampling design, an undesirable feature under the design-based framework.

Hartley and Rao (1968) took a different route in searching for a likelihood-based approach. In their proposed *scale-load* approach, some aspects of the sample data are ignored to make the sample nonunique and in turn the likelihood informative. The basic feature of the Hartley–Rao approach is to assume that the variable y is measured on a scale with a finite set of known scale points y_t^*, $t = 1, 2, \ldots, T$. The number of scale points, T, is only conceptual and inferences do not require the specification of T.

Let N_t be the number of units in U having the value y_t^*. It follows that $N = \sum_{t=1}^T N_t$ and $Y = \sum_{t=1}^T N_t y_t^*$, which is completely specified by the population "scale-loads" $\boldsymbol{N} = (N_1, N_2, \ldots, N_T)'$. Let n be the total sample size and n_t be the number of units in the sample having the value y_t^*. The sample data is effectively reduced to the observed scale-loads $\boldsymbol{n} = (n_1, n_2, \ldots, n_T)'$, with $n_t \geq 0$ and $n = \sum_{t=1}^T n_t$. Under simple random sampling without replacement, the likelihood based on the reduced sample data is given by the multi-hypergeometric distribution that depends on the population parameter $\boldsymbol{N}$, unlike the flat likelihood of (1) based on the full sample data. If the sampling fraction is negligible, the likelihood may be approximated by using the multinomial distribution with the log likelihood given by $l(\boldsymbol{p}) = \sum_{t=1}^T n_t \log(p_t)$, where $\boldsymbol{p} = (p_1, \ldots, p_T)'$ and $p_t = N_t/N$. Without using any auxiliary information, the maximum likelihood estimator of $\overline{Y} = \sum_{t=1}^T p_t y_t^*$ is the sample mean $\overline{y} = \sum_{t=1}^T \hat{p}_t y_t^*$, where $\hat{p}_t = n_t/n$.

The scale-load approach also provides an effective method for using known population mean $\overline{X}$ of an auxiliary variable x in estimating $\overline{Y}$. Denoting the scale points of x as x_j^*, $j = 1, \ldots, J$, and the scale load of (y_t^*, x_j^*) as N_{tj}, we have $\overline{Y} = \sum_{t=1}^T \sum_{j=1}^J p_{tj} y_t^*$ and $\overline{X} = \sum_{t=1}^T \sum_{j=1}^J p_{tj} x_j^*$, where $p_{tj} = N_{tj}/N$ and $\sum_{t=1}^T \sum_{j=1}^J p_{tj} = 1$. The sample data reduces to the observed frequencies n_{tj} for the scale points (y_t^*, x_j^*) such that $\sum_{t=1}^T \sum_{j=1}^J n_{tj} = n$. The scale-load estimator of $\overline{Y}$ is computed as $\hat{\overline{Y}} = \sum_{t=1}^T \sum_{j=1}^J \hat{p}_{tj} y_{tj}^*$, where $\hat{p}_{tj}$ maximize the log likelihood $\sum_{t=1}^T \sum_{j=1}^J n_{tj} \log(p_{tj})$ subject to constraints

$\sum_{t=1}^{T}\sum_{j=1}^{J} p_{tj} = 1$ and $\sum_{t=1}^{T}\sum_{j=1}^{J} p_{tj}x_{tj}^{*} = \overline{X}$. Hartley and Rao (1968) showed that $\hat{\overline{Y}}$ is asymptotically equivalent to the customary regression estimator of $\overline{Y}$. This result was later "rediscovered" when Chen and Qin (1993) applied Owen's 1988 formulation of the empirical likelihood (EL) method to the same settings. Section 2 provides more details.

Hartley and Rao (1969) generalized the scale-load approach to unequal probability sampling with replacement where selection probability is proportional to size (PPS). If y_i is approximately proportional to the size x_i, then it is reasonable to consider the scale points of $r_i = y_i/x_i$, say r_t^{*}, and the resulting scale-load estimator of Y is equal to the customary unbiased estimator in PPS sampling with replacement. Extending the scale-load approach to unequal probability sampling without replacement does not seem to be straightforward, and confidence intervals based on the likelihood ratio function were not studied under the scale-load approach.

2. Empirical likelihood method under simple random sampling

The scale-load approach of Hartley and Rao (1968, 1969) based on a multinomial distribution has the same spirit as EL proposed later by Owen (1988). Let $y_1, y_2, \ldots, y_n$ be an independent and identically distributed (*iid*) random sample from y with cumulative distribution function $F(\cdot)$. Let $p_i = P(y = y_i) = F(y_i) - F(y_i-)$ be the probability mass assigned to y_i. The EL function defined by Owen (1988) is $L(\boldsymbol{p}) = \prod_{i=1}^{n} p_i$. Maximizing $l(\boldsymbol{p}) = \log\{L(\boldsymbol{p})\} = \sum_{i\in s}\log(p_i)$ subject to $p_i > 0$ and $\sum_{i=1}^{n} p_i = 1$ leads to $\hat{p}_i = 1/n$, the maximum empirical likelihood (MEL) estimator of $F(u)$ is given by $\hat{F}(u) = \sum_{i=1}^{n}\hat{p}_i I(y_i \le u) = F_n(u)$, where $I(\cdot)$ is the indicator function and $F_n(u) = n^{-1}\sum_{i=1}^{n} I(y_i \le u)$ is the empirical distribution function based on the *iid* sample.

There have been many important contributions to the development of the EL method in mainstream statistics since Owen's 1988 paper on the asymptotic χ^2 distribution of the EL ratio statistic for the mean $\mu = E(y)$. This is evident from Owen's (2001) monograph on EL. Among other results, the work by Qin and Lawless (1994), which showed that side information in the form of a set of estimating equations can be used to improve the maximum EL estimators and the EL ratio confidence intervals, is particularly appealing for inference from survey data in the presence of auxiliary information.

The first formal use of the EL method in survey sampling was presented by Chen and Qin (1993) under simple random sampling with or without replacement. The sampling fraction is assumed to be negligible in the case of without replacement sampling so that Owen's EL function for *iid* cases can be directly used. Let $\{(y_i, x_i), i \in s\}$ be the sample data and $\theta_0 = N^{-1}\sum_{i=1}^{N} g(y_i)$ be the population parameter with $g(\cdot)$ being a known function. The known population auxiliary information is in the form of $E\{w(x)\} = 0$ for some known $w(\cdot)$. The log-likelihood function is given by $l(\boldsymbol{p}) = \log\{L(\boldsymbol{p})\} = \sum_{i\in s}\log(p_i)$. The MEL estimator of θ is defined as $\hat{\theta} = \sum_{i\in s}\hat{p}_i g(y_i)$, where $\hat{p}_i$ maximizes $l(\boldsymbol{p})$ subject to $p_i > 0$, $\sum_{i\in s} p_i = 1$, and $\sum_{i\in s} p_i w(x_i) = 0$. The MEL estimator has no closed form expression. It can be shown that the solution is given by $\hat{p}_i = \{n(1 + \lambda w(x_i)\}^{-1}$, where the Lagrange multiplier λ is the solution to $\sum_{i\in s} w(x_i)/\{1 + \lambda w(x_i)\} = 0$. We give further computational details in Section 5. For now, we note that the choice of

$g(y_i) = y_i$ gives $\theta_0 = \overline{Y}$ and incorporating the known population mean $\overline{X}$ translates into $w(x_i) = x_i - \overline{X}$ and $\sum_{i\in s} p_i x_i = \overline{X}$. The MEL estimator is uniquely defined when $\overline{X}$ is an inner point of the convex hull formed by $\{x_i, i \in s\}$. This happens with probability approaching one as $n \to \infty$. For cases where x is univariate, the convex hull becomes $(x_{(1)}, x_{(n)})$, where $x_{(1)} = \min_{i\in s} x_i$ and $x_{(n)} = \max_{i\in s} x_i$. Let $F_x(t)$ be the distribution function of x and assume simple random sampling with replacement, then $P\{x_{(1)} < \overline{X} < x_{(n)}\} = 1 - \{1 - F_x(\overline{X}-)\}^n - \{F_x(\overline{X})\}^n$, which goes to one at an exponential rate. The MEL estimator of $\overline{Y}$ is equivalent to the scale-load estimator of Hartley and Rao (1968).

By letting $g(y_i) = I(y_i \le t)$ for a fixed t, we get the MEL estimator of the population distribution function $F_N(t) = N^{-1}\sum_{i=1}^N I(y_i \le t)$ as $\hat{F}_N(t) = \sum_{i\in s} \hat{p}_i I(y_i \le t)$. The estimator $\hat{F}_N(t)$ is a genuine distribution function, that is, it is monotone nondecreasing and is confined within the range [0, 1]. Consequently, MEL estimators of population quantiles can be obtained through direct inversion of $\hat{F}_N(t)$.

The EL approach provides nonparametric confidence intervals through the profiling of EL ratio statistics, similar to the parametric case. For $\theta_0 = \overline{Y}$ and in the absence of any auxiliary information, the log EL ratio function is given by $r(\theta) = -2\sum_{i\in s} \log(n\hat{p}_i)$, where $\hat{p}_i$ maximizes the EL function $l(\boldsymbol{p})$ subject to constraints $p_i > 0$, $\sum_{i\in s} p_i = 1$, and $\sum_{i\in s} p_i y_i = \theta$ for a fixed value of θ. It can be shown that under some moment conditions on y for the finite population and a suitable asymptotic framework that allows n and N simultaneously go to infinity while n/N goes to zero, the EL ratio function $r(\theta)$ converges in distribution to a χ^2 random variable with *one* degree of freedom when $\theta = \theta_0$. A $1-\alpha$ level EL confidence interval for $\theta_0 = \overline{Y}$ is then given by $\mathcal{C}_{\text{el}} = \{\ \theta \mid r(\theta) \le \chi_1^2(\alpha)\}$, where $\chi_1^2(\alpha)$ is the upper α-quantile of the χ^2 distribution with one degree of freedom. Finding such an interval involves profiling. Section 5 again contains the computational detail. Unlike the symmetric interval based on the normal approximation (NA) to the Z-statistic $(\hat{\theta}_0 - \theta_0)/\{\text{var}(\hat{\theta}_0)\}^{1/2}$, the orientation of the EL interval $\mathcal{C}_{\text{el}}$ is determined by the data and the range of the parameter space is fully preserved.

Section 4.3 contains results from a limited simulation study on the EL interval for $\theta_0 = F_N(t)$. The results demonstrate several advantages of the EL intervals. One of them is that the upper and lower bounds of the EL intervals are always within the range of [0, 1], which is not the case for the conventional normal theory intervals.

Under simple random sampling, Chen et al. (2003) compared EL intervals with several alternatives for the population mean of populations containing many zero values. Such populations are encountered, for instance, in audit sampling, where the response variable y denotes the amount of money owned to the government and the population mean $\overline{Y}$ is the average amount of excessive claims. Most of the claims are legitimate, with corresponding y being zeros, but a small portion of claims may be excessive. The lower bound (LB) of the 95% confidence interval on $\overline{Y}$ is often used to compute the total amount of money owned to the government. Since the total number of claims selected for auditing is usually not large, the NA confidence intervals perform poorly in terms of the lower tail error rate and the average LB. Parametric likelihood ratio intervals based on parametric mixture distributions for the y variable have been used in auditing, but the performance of such intervals depends heavily on the validity of the assumed parametric model. The EL intervals exhibit behavior similar to intervals based on a correctly specified mixture model. More importantly, they perform better than the intervals based on incorrectly specified mixture models. That is to say, the EL intervals

Table 1
95% Confidence intervals for $\overline{Y}$ under PPS sampling

ρ	Zeros (%)	CI	CP	L	U	LB
0.30	95	NA	86.0	0.3	13.7	−0.01
		EL	91.0	2.4	6.6	0.09
	90	NA	87.3	0.5	12.2	0.13
		EL	91.4	2.5	6.1	0.26
	80	NA	92.0	1.0	7.0	0.58
		EL	93.6	2.5	3.9	0.72
	70	NA	93.0	1.6	5.4	1.11
		EL	94.5	2.7	2.8	1.23
0.80	95	NA	84.1	0.4	15.5	0.00
		EL	92.8	3.1	4.1	0.10
	90	NA	89.6	1.3	9.1	0.17
		EL	92.1	3.3	4.6	0.27
	80	NA	92.9	1.6	5.5	0.65
		EL	94.3	2.6	3.1	0.73
	70	NA	93.8	2.1	4.1	1.18
		EL	94.9	2.8	2.3	1.24

provide lower error rates at least as close to the nominal values while the intervals based on incorrectly specified mixture models lead to lower error rates much smaller than the nominal values. As a result, EL intervals have larger LB, and methods that respect nominal error rates and at the same time provide larger LBs are regarded as desirable ones.

Under unequal probability sampling, the use of any parametric mixture model for such populations becomes difficult to justify. The EL intervals, however, are still available using the pseudo EL formulation described in Section 4.3. Table 1 reports the performance of the pseudo EL confidence intervals for the population mean from a simulation study. The finite population is first generated through Model I used by Wu and Rao (2006), with the correlation coefficient between the design variable z_i and the response variable y_i indicated by ρ, and a random portion of the y_is is then set to be zeros. Information on the design variable z is not further used in constructing the EL confidence intervals. The interval based on NA is included for comparison. The reported results on coverage probability (CP), lower (L) and upper (U) tail error rates, and the average LB are based on 1000 simulated samples of size $n = 60$, selected by the PPS sampling method of Rao (1965) and Sampford (1967). While intervals based on NAs are clearly inappropriate, the EL interval maintains the same desirable performance observed under simple random sampling, with the lower tail error rates close to the nominal value and larger LB for all cases considered.

3. Stratified simple random sampling

Stratified simple random sampling is commonly used when list frames within strata are available, as in many business surveys. Let $\{(y_{hi}, \boldsymbol{x}_{hi}), i \in s_h,\ h = 1, \ldots, L\}$ be a stratified simple random sample, where y_{hi} and $\boldsymbol{x}_{hi}$ are, respectively, the values of the

study variable y and the vector of auxiliary variables $\boldsymbol{x}$ associated with the ith element in stratum h, L is the total number of strata in the population, and s_h is the set of n_h sampled units from stratum h. Let N_h be the stratum size, $W_h = N_h/N$ be the stratum weight, and $N = \sum_{h=1}^{L} N_h$ is the overall population size. Zhong and Rao (1996, 2000) studied EL inferences on $\overline{Y}$ when the vector-valued population mean $\overline{\boldsymbol{X}}$ is known but the stratum means $\overline{\boldsymbol{X}}_h$ are unknown. Assuming negligible sampling fractions within strata and noting that samples from different strata are independent, the log EL function under stratified simple random sampling is given by $l(\boldsymbol{p}_1, \ldots, \boldsymbol{p}_L) = \sum_{h=1}^{L} \sum_{i \in s_h} \log(p_{hi})$, where $\boldsymbol{p}_h = (p_{h1}, \ldots, p_{hn_h})'$ and p_{hi} is the probability mass assigned to y_{hi}, $i \in s_h$, $h = 1, \ldots, L$. The MEL estimator of $\overline{Y}$ is defined as $\hat{\overline{Y}} = \sum_{h=1}^{L} W_h \sum_{i \in s_h} \hat{p}_{hi} y_{hi}$, where the $\hat{p}_{hi}$ maximize $l(\boldsymbol{p}_1, \ldots, \boldsymbol{p}_L)$ subject to $p_{hi} > 0$, $\sum_{i \in s_h} p_{hi} = 1$, $h = 1, \ldots, L$, and $\sum_{h=1}^{L} W_h \sum_{i \in s_h} p_{hi} \boldsymbol{x}_{hi} = \overline{\boldsymbol{X}}$. For fixed L and large sample sizes n_h with negligible sampling fraction within strata, the MEL estimator is asymptotically equivalent to the randomization optimal linear regression estimator (Kott, Chapter 25; Zhong and Rao, 2000). While most design-based estimators can only be justified under unconditional repeated sampling, the optimal estimator leads to valid conditional inferences, with negligible conditional relative bias given the stratified mean $\overline{\boldsymbol{x}}_{\mathrm{st}} = \sum_{h=1}^{L} W_h \overline{\boldsymbol{x}}_h$, where $\overline{\boldsymbol{x}}_h = n_h^{-1} \sum_{i \in s_h} \boldsymbol{x}_{hi}$ (Rao, 1994). Zhong and Rao (2000) also studied the EL ratio confidence intervals on $\overline{Y}$. An efficient computational algorithm was proposed by Wu (2004b) with simple R/SPLUS functions and codes available in Wu (2005). We will describe some of the details in Sections 4.3 and 5. Note that the constrained maximization problem here is more involved than in the case of simple random sampling (Section 2) because it is not possible to impose separate constraints of the form $\sum_{i \in s_h} p_{hi} \boldsymbol{x}_{hi} = \overline{\boldsymbol{X}}_h$, $h = 1, \ldots, L$ when the strata means $\overline{\boldsymbol{X}}_h$ are unknown. The case of deep stratification (large L and small n_h within each stratum h) is not covered by Zhong and Rao (2000) and it had not been studied in the EL literature.

4. Pseudo empirical likelihood method

One of the major difficulties for the EL inferences under general unequal probability sampling designs is to obtain an informative EL function for the given sample. The likelihood depends necessarily on the sampling design and a complete specification of the joint probability function of the sample is usually not feasible under any without replacement sampling. Because of this difficulty, Chen and Sitter (1999) proposed a *pseudo* EL approach using a two-stage argument. Suppose the finite population $\{y_1, \ldots, y_N\}$ can be viewed as an *iid* sample from a superpopulation, then the population (or "census") log EL would be $l_N(\boldsymbol{p}) = \sum_{i=1}^{N} \log(p_i)$, the population total of the $\log(p_i)$. For a given sample, the design-based Horvitz–Thompson estimator of $l_N(\boldsymbol{p})$ is given by $l_{\mathrm{HT}}(\boldsymbol{p}) = \sum_{i \in s} d_i \log(p_i)$, where $d_i = 1/\pi_i$, and $\pi_i = P(i \in s)$ are the first-order inclusion probabilities. Chen and Sitter termed $l_{\mathrm{HT}}(\boldsymbol{p})$ the pseudolog EL. Under simple random sampling and ignoring a multiplying constant, $l_{\mathrm{HT}}(\boldsymbol{p})$ reduces to $l(\boldsymbol{p})$ used by Chen and Qin (1993).

The pseudo EL function $l_{\mathrm{HT}}(\boldsymbol{p})$ involves only the first-order inclusion probabilities and does not catch the design effects under general unequal probability sampling without replacement. Wu and Rao (2006) defined the pseudo empirical log-likelihood (PELL)

function under nonstratified (ns) sampling designs as

$$l_{\mathrm{ns}}(\boldsymbol{p}) = n \sum_{i \in s} \tilde{d}_i(s) \log(p_i), \tag{2}$$

where $\tilde{d}_i(s) = d_i / \sum_{i \in s} d_i$. The pseudo EL function (2) has likelihood-based motivation but is directly related to the "backward" Kullback–Leibler distance (DiCiccio and Romano, 1990) between $\boldsymbol{p} = (p_1, \ldots, p_n)'$ and $\boldsymbol{d}(s) = (\tilde{d}_1(s), \ldots, \tilde{d}_n(s))'$ in the form of $D(\boldsymbol{d}(s), \boldsymbol{p}) = \sum_{i \in s} \tilde{d}_i(s) \log(\tilde{d}_i(s)/p_i)$. Since $D(\boldsymbol{d}(s), \boldsymbol{p}) = \sum_{i \in s} \tilde{d}_i(s) \log(\tilde{d}_i(s)) - l_{\mathrm{ns}}(\boldsymbol{p})/n$, minimizing the Kullback–Leibler distance with respect to p_i subject to a set of constraints is equivalent to maximizing the pseudo EL function subject to the same set of constraints.

The PELL function given by (2) differs from $l_{\mathrm{HT}}(\boldsymbol{p})$ used by Chen and Sitter (1999) in the sense that the normalized weights $\tilde{d}_i(s)$, also called the Hajek weights (Hajek, 1971), are used instead of d_i. But maximizing (2) subject to a set of constraints on the p_i is equivalent to maximizing $l_{\mathrm{HT}}(\boldsymbol{p})$ subject to the same set of constraints, and the resulting maximum pseudo empirical likelihood (MPEL) estimators remain the same. However, it is shown in Section 4.3 that $l_{\mathrm{ns}}(\boldsymbol{p})$ allows for simple adjustment for the design effect in constructing pseudo EL ratio confidence intervals.

For stratified (st) sampling with an arbitrary sampling design within each stratum, the PELL function of Wu and Rao (2006) is defined as

$$l_{\mathrm{st}}(\boldsymbol{p}_1, \ldots, \boldsymbol{p}_L) = n \sum_{h=1}^{L} W_h \sum_{i \in s_h} \tilde{d}_{hi}(s_h) \log(p_{hi}), \tag{3}$$

where $\tilde{d}_{hi}(s_h) = d_{hi} / \sum_{i \in s_h} d_{hi}$ are the normalized weights within each stratum with $d_{hi} = \pi_{hi}^{-1}$ denoting the design weights, $\pi_{hi} = P(i \in s_h)$ the hth stratum unit inclusion probabilities, $n = \sum_{h=1}^{L} n_h$, and n_h the stratum sample sizes. Note that $l_{\mathrm{st}}(\boldsymbol{p}_1, \ldots, \boldsymbol{p}_L)$ does *not* reduce to the empirical log-likelihood function $\sum_{h=1}^{L} \sum_{i \in s_h} \log(p_{hi})$ under stratified simple random sampling (Zhong and Rao, 2000) unless $n_h = n W_h$, that is, the stratum sample sizes are proportionally allocated.

In the absence of auxiliary information, maximizing $l_{\mathrm{ns}}(\boldsymbol{p})$ subject to $p_i > 0$ and $\sum_{i \in s} p_i = 1$ gives $\hat{p}_i = \tilde{d}_i(s)$. The resulting MPEL estimator of $\overline{Y}$, defined as $\hat{\overline{Y}}_{\mathrm{EL}} = \sum_{i \in s} \hat{p}_i y_i$, is given by the Hajek estimator $\hat{\overline{Y}}_{\mathrm{H}} = \sum_{i \in s} \tilde{d}_i(s) y_i$, and the MPEL estimator of the distribution function $F_N(t)$ is given by $\hat{F}_{\mathrm{H}}(t) = \sum_{i \in s} d_i I(y_i \le t) / \sum_{i \in s} d_i$.

The Hajek estimator of $\overline{Y}$, however, can be less efficient than the Horvitz–Thompson estimator $\hat{\overline{Y}}_{\mathrm{HT}} = N^{-1} \sum_{i \in s} d_i y_i$ under PPS sampling without replacement when the response variable y is highly correlated with the size variable. For designs with fixed sample size, Wu and Rao (2006) suggested a more efficient estimator by imposing the constraint

$$\sum_{i \in s} p_i \pi_i = \frac{n}{N}. \tag{4}$$

Equation (4) is the same as $\sum_{i \in s} p_i z_i = \overline{Z}$, where z_i is the size variable and $\overline{Z}$ is the population mean. The resulting MPEL estimator, which is a special case of the estimators discussed in Section 4.1, is equivalent to a regression type estimator with variance

depending on the residuals. The Hajek estimator of $F_N(t)$ at a fixed t, on the other hand, is very efficient since the indicator variable $I(y_i \le t)$ is weakly correlated with the size variable, and $\hat{F}_{\mathrm{H}}(t)$ itself is a genuine distribution function.

4.1. Pseudo empirical likelihood approach to calibration

Suppose the population mean $\overline{\boldsymbol{X}}$ of a vector of auxiliary variables $\boldsymbol{x}$ is known. In this case, the MPEL estimator of $\overline{Y}$ is given by $\hat{\overline{Y}}_{\mathrm{EL}} = \sum_{i\in s} \hat{p}_i y_i$, where $\hat{p}_i$ maximize $l_{\mathrm{ns}}(\boldsymbol{p})$ subject to $p_i > 0$, $\sum_{i\in s} p_i = 1$, and

$$\sum_{i\in s} p_i \boldsymbol{x}_i = \overline{\boldsymbol{X}}. \tag{5}$$

Using the Lagrange multiplier method, we can show that $\hat{p}_i = \tilde{d}_i(s)/(1 + \boldsymbol{\lambda}'\boldsymbol{u}_i)$, where $\boldsymbol{u}_i = \boldsymbol{x}_i - \overline{\boldsymbol{X}}$, and $\boldsymbol{\lambda}$ is the solution to

$$g(\boldsymbol{\lambda}) = \sum_{i\in s} \frac{\tilde{d}_i(s)\boldsymbol{u}_i}{1 + \boldsymbol{\lambda}'\boldsymbol{u}_i} = \boldsymbol{0}. \tag{6}$$

Constraints such as (5) are often referred to as benchmark constraints or calibration equations. A calibration estimator for $\overline{Y}$ can be defined as $\hat{\overline{Y}}_{\mathrm{C}} = N^{-1}\sum_{i\in s} w_i y_i$, where the calibrated weights w_i minimize a distance measure $\Phi(\boldsymbol{w}, \boldsymbol{d})$ between $\boldsymbol{w} = (w_1, \ldots, w_n)'$ and the basic design weights $\boldsymbol{d} = (d_1, \ldots, d_n)'$ subject to $\sum_{i\in s} w_i \boldsymbol{x}_i = \boldsymbol{X}$. The simple chi-squared distance $\Phi(\boldsymbol{w}, \boldsymbol{d}) = \sum_{i\in s}(w_i - d_i)^2/(d_i q_i)$ with prespecified q_i provides closed-form solutions to w_i, leading to a generalized regression (GREG) estimator of Y (Särndal et al., 1992), but the resulting w_i can take negative values under unbalanced sample configurations. Other distance measures that force the weights to be positive are available, but most of them suffer computational inefficiencies or other undesirable features.

There are several attractive features with the pseudo empirical likelihood approach to calibration estimation. First, the weights $\hat{p}_i$ are intrinsically positive and normalized, that is, $\hat{p}_i > 0$ and $\sum_{i\in s} \hat{p}_i = 1$. This is particularly appealing for the MPEL estimator of $F_N(t)$ computed as $\hat{F}_{\mathrm{EL}}(t) = \sum_{i\in s} \hat{p}_i I(y_i \le t)$. Like the MEL estimator in the *iid* setting, it is a genuine distribution function, and quantile estimates can be obtained through a direct inversion of $\hat{F}_{\mathrm{EL}}(t)$. Second, for the major computational task of finding the Lagrange multiplier $\boldsymbol{\lambda}$ as the solution to (6), a modified Newton–Raphson algorithm (Chen et al., 2002), which guarantees fast convergence, is available as we show in Section 5. Third, the pseudo EL ratio confidence intervals, described in Section 4.3, have several advantages over the conventional normal theory intervals.

If confidence intervals are not of major interest and the focus is on reporting standard errors, the approximate design-based variance and a variance estimator on $\hat{\overline{Y}}_{\mathrm{EL}}$ are also readily available. Under regularity conditions C1–C3 described in Section 4.3, we have $\boldsymbol{\lambda} = \left(\sum_{i\in s} d_i \boldsymbol{u}_i \boldsymbol{u}_i'\right)^{-1} \sum_{i\in s} d_i \boldsymbol{u}_i + o_p\left(n^{-1/2}\right)$ and $\hat{p}_i \doteq \tilde{d}_i(s)(1 - \boldsymbol{\lambda}'\boldsymbol{u}_i)$, which lead to

$$\hat{\overline{Y}}_{\mathrm{EL}} = \hat{\overline{Y}}_{\mathrm{H}} + \hat{\boldsymbol{B}}'\left(\overline{\boldsymbol{X}} - \hat{\overline{\boldsymbol{X}}}_{\mathrm{H}}\right) + o_p\left(n^{-1/2}\right), \tag{7}$$

where $\boldsymbol{u}_i = \boldsymbol{x}_i - \hat{\overline{\boldsymbol{X}}}_{\mathrm{H}}$, $\hat{\boldsymbol{B}} = (\sum_{i\in s} d_i \boldsymbol{u}_i \boldsymbol{u}_i')^{-1} \sum_{i\in s} d_i \boldsymbol{u}_i y_i$, $\hat{\overline{\boldsymbol{X}}}_{\mathrm{H}} = \sum_{i\in s} \tilde{d}_i(s)\boldsymbol{x}_i$, and $\hat{\overline{Y}}_{\mathrm{H}} + \hat{\boldsymbol{B}}'\left(\overline{\boldsymbol{X}} - \hat{\overline{\boldsymbol{X}}}_{\mathrm{H}}\right)$ is a GREG estimator of $\overline{Y}$. It follows from (7) that linearization variance estimation techniques for GREG estimators can be applied to $\hat{\overline{Y}}_{\mathrm{EL}}$. Similarly, for $\hat{F}_{\mathrm{EL}}(t)$ by changing y_i in (7) to $I(y_i \le t)$.

In practice, one might wish to restrict the range of the calibrated weights so that $c_1 \le w_i/d_i \le c_2$ for some prespecified $0 < c_1 < 1 < c_2$. Under the pseudo EL approach, this amounts to imposing

$$c_1 \le p_i/\tilde{d}_i(s) \le c_2\,,\ i \in s\,. \tag{8}$$

Chen et al. (2002) suggested a simple computational procedure to achieve (8) through a minimal relaxation of the benchmark constraints (5). If the MPEL solutions $\hat{p}_i$ using (5) do not satisfy (8), we replace (5) by

$$\sum_{i\in s} p_i \boldsymbol{x}_i = \overline{\boldsymbol{X}} + \delta(\hat{\overline{\boldsymbol{X}}}_H - \overline{\boldsymbol{X}})$$

for some $\delta \in [0, 1]$ in finding the MPEL solutions $\hat{p}_i$. The choice of $\delta = 0$ (i.e., no relaxation from (5)) corresponds to the the initial MPEL solution. At the other end, the value of $\delta = 1$ gives $\hat{p}_i = \tilde{d}_i(s)$, which always satisfy (8). For any prechosen $c_1 < 1 < c_2$, the smallest value of δ with the corresponding MPEL solutions satisfying (8) can be found through a simple bisection search method (Chen et al., 2002).

There are two major motivations behind any type of calibration estimation method, including the MPEL method: (i) internal consistency, achieved through the benchmark constraints (5); (ii) efficiency, due to the asymptotic equivalence of $\hat{\overline{Y}}_{\mathrm{EL}}$ to the GREG estimator, as shown from (7). But using (5) for the estimation of $\overline{Y}$ may not be very efficient when the underlying relationship between y and $\boldsymbol{x}$ does not approximate a linear model. Wu and Sitter (2001a) proposed a model-calibrated pseudo EL method when the underlying superpopulation model, linear or nonlinear, is specified as $E_\xi(y_i|\boldsymbol{x}_i) = \mu(\boldsymbol{x}_i, \boldsymbol{\theta})$ and $V_\xi(y_i|\boldsymbol{x}_i) = v(\boldsymbol{x}_i)\sigma^2$, where ξ denotes the superpopulation model, $\boldsymbol{\theta}$ and σ^2 are model parameters, $\mu(\cdot,\cdot)$ and $v(\cdot)$ are known functions. The traditional constraints (5) are replaced by

$$\sum_{i\in s} p_i \hat{\mu}_i = N^{-1} \sum_{i=1}^{N} \hat{\mu}_i\,, \tag{9}$$

where the calibration variable $\hat{\mu}_i = \mu(\boldsymbol{x}_i, \hat{\boldsymbol{\theta}})$ is the predicted value of y_i based on the model, and $\hat{\boldsymbol{\theta}}$ is a design-based estimator of $\boldsymbol{\theta}$. The PELL function (2) is maximized subject to the calibration constraint (9), leading to the model-calibrated MPEL estimator of the mean $\overline{Y}$.

Note that when $\mu(\boldsymbol{x}_i, \boldsymbol{\theta})$ has a nonlinear form, constraint (9) requires that complete information on $\boldsymbol{x}$, that is, $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$, be known. Under the linear model $\mu(\boldsymbol{x}_i, \boldsymbol{\theta}) = \boldsymbol{x}_i'\boldsymbol{\theta}$, the "population mean" $N^{-1}\sum_{i=1}^{N} \hat{\mu}_i$ reduces to $\overline{\boldsymbol{X}}'\hat{\boldsymbol{\theta}}$, so only $\overline{\boldsymbol{X}}$ is needed in (9) in this case. The resulting model-calibrated MPEL estimator of $\overline{Y}$ using (9) is asymptotically equivalent to the MPEL estimator using (5), under the linear model.

The model-calibrated MPEL estimator of $\overline{Y}$ is asymptotically *optimal* in the class of MPEL estimators $\hat{\overline{Y}}_u = \sum_{i\in s} p_i y_i$ satisfying the calibration constraint $\sum_{i\in s} p_i u(\boldsymbol{x}_i) = N^{-1}\sum_{i=1}^N u(\boldsymbol{x}_i)$, where $u(\cdot)$ is an arbitrary function satisfying finite moment conditions and $\mu(\cdot, \boldsymbol{\theta})$ is assumed to belong to this class (Wu, 2003). That is, the choice $u(\boldsymbol{x}_i) = \mu(\boldsymbol{x}_i, \boldsymbol{\theta})$ minimizes the anticipated asymptotic variance $E_\xi \mathrm{AV}(\hat{\overline{Y}}_u)$, where AV denotes the asymptotic design variance. One immediate application of this result is the optimal calibration estimator of the distribution function $F_N(t)$ at a fixed t. The optimal calibration variable that should be used in (9) is given by $E_\xi\{I(y_i \le t)\} = P(y_i \le t)$. Chen and Wu (2002) compared the efficiency of the optimal model-calibrated MPEL estimator of $F_N(t)$ to several alternative estimators and also discussed the related quantile estimation problem.

The model-calibrated MPEL estimation method can be extended to cover quadratic population parameters in the form of $T = \sum_{i=1}^N \sum_{j=i+1}^N \phi(y_i, y_j)$, which includes the population variance, the covariance, and the variance of a linear estimator as special cases (Sitter and Wu, 2002). The basic idea is to view T as a total over a synthetic finite population, that is, $T = \sum_{\alpha=1}^{N^*} t_\alpha$, where $\alpha = (ij)$ is relabelled from 1 to $N^* = N(N-1)/2$ and $t_\alpha = \phi(y_i, y_j)$. The synthetic sample consists of all the pairs from the original sample and the "first-order" inclusion probabilities under this setting are $\pi_{ij} = P(i, j \in s)$. The "basic design weights" are $d_{ij} = 1/\pi_{ij}$. The extended pseudolog EL function for quadratic parameters is defined as

$$l^*(\boldsymbol{p}) = \sum_{i\in s}\sum_{j>i} d_{ij}\log(p_{ij}),$$

where p_{ij} is the probability mass assigned to the pair (i, j). The model-calibrated MPEL estimator of T is defined as

$$\hat{T}_{EL} = N^* \sum_{i\in s}\sum_{j>i} \hat{p}_{ij}\phi(y_i, y_j),$$

where the $\hat{p}_{ij}$ maximize $l^*(\boldsymbol{p})$ subject to

$$\sum_{i\in s}\sum_{j>i} p_{ij} = 1 \quad \text{and} \quad \sum_{i\in s}\sum_{j>i} p_{ij}u_{ij} = \frac{1}{N^*}\sum_{i=1}^N\sum_{j=i+1}^N u_{ij}.$$

The optimal calibration variable u_{ij} is given by $u_{ij} = E_\xi\{\phi(y_i, y_j)\}$. For the population variance $S^2 = (N-1)^{-1}\sum_{i=1}^N (y_i - \overline{Y})^2$, which can be expressed as $\{N(N-1)\}^{-1}\sum_{i=1}^N\sum_{j=i+1}^N (y_i - y_j)^2$, the optimal calibration variable is given by $u_{ij} = E_\xi\{(y_i - y_j)^2\} = \{\mu(\boldsymbol{x}_i, \boldsymbol{\theta}) - \mu(\boldsymbol{x}_j, \boldsymbol{\theta})\}^2 + \{v(\boldsymbol{x}_i) + v(\boldsymbol{x}_j)\}\sigma^2$ under the assumed model. In applications, the unknown model parameters $\boldsymbol{\theta}$ and σ^2 will have to be replaced by suitable design-based estimators. The resulting model-calibrated MPEL estimator of T remains consistent under mild conditions.

4.2. Pseudo empirical likelihood alternative to raking

Raking ratio estimation can be viewed as a special application of the calibration method, where the auxiliary information is in the form of known marginal totals of a contingency table of two or more dimensions. Unfortunately, the number of benchmark constraints

involved is often very large and the related computational procedures can be problematic. The pseudo EL alternative to raking offers a major advantage in computational efficiency and stability. Like raking, and unlike standard linear calibration approaches, the successful completion of pseudo EL will always produce positive weights. Moreover, convergence is guaranteed and fast. Confidence intervals can be constructed using either the pseudo EL ratio function to be described in Section 4.3 or the linearization variance estimator and NA. Auxiliary population information other than the marginal totals and features of complex sampling designs can also be incorporated into the estimation procedure.

We first describe the EL alternative to raking under the classical setting of Deming and Stephan (1940). Suppose the finite population is cross-classified into $r \times c$ cells with a total number of N_{ij} units in the (i, j)th cell, $i = 1, \ldots, r$, $j = 1, \ldots, c$. Let $N = \sum_{i=1}^{r} \sum_{j=1}^{c} N_{ij}$ be the total population size. The marginal totals $N_{i\cdot} = \sum_{j=1}^{c} N_{ij}$, $i = 1, \ldots, r$ and $N_{\cdot j} = \sum_{i=1}^{r} N_{ij}$, $j = 1, \ldots, c$ are known. The cell totals N_{ij} are unknown and need to be estimated. Let n be the size of a simple random sample drawn from the population and n_{ij} be the sample frequency for the (i, j)th cell. Noting that $N_{i\cdot}$, $N_{\cdot j}$, and N are all known, we could estimate N_{ij} by $n_{ij}(N/n)$ or $n_{ij}(N_{i\cdot}/n_{i\cdot})$ or $n_{ij}(N_{\cdot j}/n_{\cdot j})$, where $n_{i\cdot} = \sum_{j=1}^{c} n_{ij}$ and $n_{\cdot j} = \sum_{i=1}^{r} n_{ij}$. But none of these estimators will necessarily match the known marginal totals in both dimensions (i.e., equal $N_{i\cdot}$ when summed across the rows and $N_{\cdot j}$ when summed across the columns).

The classical raking ratio estimator of N_{ij} in the form of $m_{ij}(N/n)$, obtained through the so-called iterative proportional fitting procedure (IPFP) (Deming and Stephan, 1940), was initially conceived to minimize the least square distance $\Phi = \sum_{i=1}^{r} \sum_{j=1}^{c} (m_{ij} - n_{ij})^2 / n_{ij}$ subject to the set of constraints

$$\sum_{j=1}^{c} m_{ij} = N_{i\cdot} n / N, \quad i = 1, \ldots, r, \tag{10}$$

$$\sum_{i=1}^{r} m_{ij} = N_{\cdot j} n / N, \quad j = 1, \ldots, c - 1. \tag{11}$$

Although the m_{ij} obtained through the IPFP satisfy (10) and (11), they do not minimize the least square distance Φ (Stephan, 1942). Ireland and Kullback (1968) showed that the estimates $\hat{p}_{ij} = m_{ij}/n$ in fact minimize the "forward" discrimination information (also called the "forward" Kullback–Leibler distance by DiCiccio and Romano, 1990)

$$I(\boldsymbol{p}, \boldsymbol{q}) = \sum_{i=1}^{r} \sum_{j=1}^{c} p_{ij} \log \left(\frac{p_{ij}}{q_{ij}} \right)$$

with respect to p_{ij} subject to (10) and (11), where $q_{ij} = n_{ij}/n$ are the observed cell proportions.

Our proposed EL alternative to raking, under simple random sampling, is to estimate the cell proportions N_{ij}/N by the $\hat{p}_{ij}$ that maximize the EL function

$$l_0(\boldsymbol{p}) = \sum_{i=1}^{r} \sum_{j=1}^{c} n_{ij} \log(p_{ij})$$

subject to

$$\sum_{i=1}^{r}\sum_{j=1}^{c} p_{ij} = 1, \tag{12}$$

$$\sum_{j=1}^{c} p_{ij} = N_{i\cdot}/N, \quad i = 1, \ldots, r-1, \tag{13}$$

$$\sum_{i=1}^{r} p_{ij} = N_{\cdot j}/N, \quad j = 1, \ldots, c-1. \tag{14}$$

The EL function $l_0(\boldsymbol{p})$ is related to the "backward" discrimination information

$$I(\boldsymbol{q}, \boldsymbol{p}) = \sum_{i=1}^{r}\sum_{j=1}^{c} q_{ij} \log\left(\frac{q_{ij}}{p_{ij}}\right) = \sum_{i=1}^{r}\sum_{j=1}^{c} q_{ij}\log(q_{ij}) - \frac{1}{n} l_0(\boldsymbol{p}).$$

It is apparent that minimizing $I(\boldsymbol{q}, \boldsymbol{p})$ with respect to p_{ij} is equivalent to maximizing $l_0(\boldsymbol{p})$ with respect to p_{ij}. The EL function $l_0(\boldsymbol{p})$ is indeed the true multinomial likelihood function under simple random sampling with replacement.

The use of "forward" discrimination information $I(\boldsymbol{p}, \boldsymbol{q})$ for classical raking ratio estimation is equivalent to the *multiplicative* method described in Deville et al. (1993) and Deville and Särndal (1992). The resulting $\hat{p}_{ij}$ are guaranteed to be positive but often contain some extremely large values compared to q_{ij}. Deville et al. (1993) also discussed other alternative distance measures that force the ratio p_{ij}/q_{ij} to be confined within certain range. The major challenge in using these alternative methods, as well as the multiplicative method, is the computational implementation. Efficient algorithms are not available and the convergence of the involved iterative procedures is not guaranteed.

The most important feature of the EL approach, however, is the availability of a simple and efficient algorithm for the constrained maximization problem. Let $x_{(1)ij}, \ldots, x_{(r-1)ij}$ be the first $r-1$ row indicator variables and $x_{ij(1)}, \ldots, x_{ij(c-1)}$ be the first $c-1$ column indicator variables. For instance, $x_{(1)ij} = 1$ if $i = 1$ and zero otherwise. Let

$$\boldsymbol{x}_{ij} = (x_{(1)ij}, \ldots, x_{(r-1)ij}, x_{ij(1)}, \ldots, x_{ij(c-1)})'$$

and

$$\overline{\boldsymbol{X}} = \left(\frac{N_{1\cdot}}{N}, \ldots, \frac{N_{(r-1)\cdot}}{N}, \frac{N_{\cdot 1}}{N}, \ldots, \frac{N_{\cdot(c-1)}}{N}\right)'.$$

The two sets of constraints (13) and (14) can be rewritten as

$$\sum_{i=1}^{r}\sum_{j=1}^{c} p_{ij}\boldsymbol{x}_{ij} = \overline{\boldsymbol{X}}. \tag{15}$$

Using the standard Lagrange multiplier method, it can be shown that the $\hat{p}_{ij}$ which maximize the pseudo EL function $l_0(\boldsymbol{p})$ subject to the normalization constraint (12) and the benchmark constraint (15) are given by $\hat{p}_{ij} = n_{ij}/\{n(1+\boldsymbol{\lambda}'\boldsymbol{u}_{ij})\}$, where $\boldsymbol{u}_{ij} = \boldsymbol{x}_{ij} - \overline{\boldsymbol{X}}$,

and the vector-valued Lagrange multiplier $\boldsymbol{\lambda}$ is the solution to

$$g_0(\boldsymbol{\lambda}) = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{n_{ij}\boldsymbol{u}_{ij}}{n(1+\boldsymbol{\lambda}'\boldsymbol{u}_{ij})} = \mathbf{0}. \tag{16}$$

A unique solution to (16) exists if none of the observed sample marginal totals $n_{i\cdot}$ or $n_{\cdot j}$ is zero. If a particular sample cell frequency $n_{ij} = 0$, we set $\hat{p}_{ij} = 0$ for that cell, and this does not change the existence and uniqueness of the solution. On the other hand, the classical raking ratio algorithm may not converge if some $n_{ij} = 0$. The solution to (16) can be found using the same algorithm of Chen et al. (2002) for solving (6).

Under a general probability sampling design and with known population mean $\overline{\boldsymbol{Z}}$ on a vector of auxiliary variables $\boldsymbol{z}$ in addition to known marginal totals in, for instance, a two-dimensional contingency table, a pseudo EL alternative to raking is as follows. Let π_{ijk} be the inclusion probability and $d_{ijk} = 1/\pi_{ijk}$ be the basic design weight associated with the kth unit in the (i, j)th cell, $k = 1, \ldots, N_{ij}$. Let $\boldsymbol{z}_{ijk}$ be the additional vector-valued auxiliary variable observed only for units in the sample but with known population mean $\overline{\boldsymbol{Z}}$. The population size N is also known from the contingency table. The goal is to estimate the population total $Y = \sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{N_{ij}} y_{ijk}$, or equivalently the mean $\overline{Y} = Y/N$, of a study variable y. Estimation of the cell totals N_{ij} is a special case of Y, where y is the (i, j)th cell indicator variable.

Under the current setting, the nonstratified pseudo EL function is defined as

$$l_1(\boldsymbol{p}) = \sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{n_{ij}} d_{ijk}\log(p_{ijk}),$$

where p_{ijk} is the probability mass assigned to the kth unit in the (i, j)th cell and the p_{ijk} are subject to

$$\sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{n_{ij}} p_{ijk} = 1. \tag{17}$$

The MPEL estimator of $\overline{Y}$ is computed as

$$\hat{\overline{Y}} = \sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{n_{ij}} \hat{p}_{ijk}y_{ijk},$$

where the $\hat{p}_{ijk}$ maximize the pseudo EL function $l_1(\boldsymbol{p})$ subject to (17) and the set of benchmark constraints

$$\sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{n_{ij}} p_{ijk}\boldsymbol{x}_{ijk} = \overline{\boldsymbol{X}}. \tag{18}$$

The vector-valued $\boldsymbol{x}_{ijk}$ consists of the first $r-1$ row indicator variables and the first $c-1$ column indicator variables as well as $\boldsymbol{z}_{ijk}$. It is very important to note that the row and column indicator variables used here are defined at the unit level while those used in (15) are defined at the (i, j) cell level. For instance, the first row indicator variable used in (18) is defined as $x_{(1)ijk} = 1$ if $i = 1$ and zero otherwise. The population mean

$\overline{\boldsymbol{X}}$ used in (18) consists of the first $r-1$ marginal row proportions $N_{i\cdot}/N$ and the first $c-1$ marginal column proportions $N_{\cdot j}/N$, as well as the mean $\overline{\boldsymbol{Z}}$.

It can be shown that the $\hat{p}_{ijk}$ are given by $\hat{p}_{ijk} = \tilde{d}_{ijk}(s)/(1+\boldsymbol{\lambda}'\boldsymbol{u}_{ijk})$, where $\tilde{d}_{ijk}(s) = d_{ijk}/\sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{n_{ij}} d_{ijk}$ are the normalized design weights over the sample s, $\boldsymbol{u}_{ijk} = \boldsymbol{x}_{ijk} - \overline{\boldsymbol{X}}$, and $\boldsymbol{\lambda}$ is the solution to

$$g_1(\boldsymbol{\lambda}) = \sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{n_{ij}} \frac{\tilde{d}_{ijk}(s)\boldsymbol{u}_{ijk}}{1+\boldsymbol{\lambda}'\boldsymbol{u}_{ijk}} = \boldsymbol{0}. \tag{19}$$

A unique solution to (19) exists if none of the observed sample marginal totals $n_{i\cdot}$ or $n_{\cdot j}$ is zero and $\overline{\boldsymbol{Z}}$ is an inner point of the convex hull formed by the $\boldsymbol{z}_{ijk}$ observed in the sample. One of the major features of the pseudo EL approach under the above formulation is that $g_1(\boldsymbol{\lambda}) = \boldsymbol{0}$ is very well structured and can be solved using the same algorithm of Chen et al. (2002).

The cell totals N_{ij} are estimated by $\hat{N}_{ij} = N\sum_{k=1}^{n_{ij}} \hat{p}_{ijk}$. It is apparent that

$$\sum_{j=1}^{c} \hat{N}_{ij} = N_{i\cdot}, i = 1, \ldots, r \quad \text{and} \quad \sum_{i=1}^{r} \hat{N}_{ij} = N_{\cdot j}, j = 1, \ldots, c.$$

Under simple random sampling and without the additional auxiliary variables $\boldsymbol{z}_{ijk}$, we have $d_{ijk} = N/n$ and the probability mass p_{ijk} take the same value for all the units k in the same (i, j)th cell. Apart from a multiplying constant, the pseudo EL function $l_1(\boldsymbol{p})$ reduces to the EL function $l_0(\boldsymbol{p})$ used for the classical simple random sampling setup.

4.3. Pseudo empirical likelihood ratio confidence intervals

One of the major attractive features of the EL approach is the nonparametric confidence intervals constructed through profiling the EL ratio function. With designs other than simple random sampling, the pseudo EL ratio statistic needs to be adjusted for the design effect. The exact definition of the design effect depends not only on the probability sampling design but also on the auxiliary information used. We consider pseudo EL intervals for the population mean $\overline{Y}$ and the distribution function $F_N(t)$ and discuss three scenarios.

For a nonstratified sampling design, the PELL ratio function for $\overline{Y}$ without using any auxiliary information at the estimation stage is given by

$$r_{\mathrm{ns}}(\theta) = -2\{l_{\mathrm{ns}}(\hat{\boldsymbol{p}}(\theta)) - l_{\mathrm{ns}}(\hat{\boldsymbol{p}})\}, \tag{20}$$

where the $\hat{p}_i = \tilde{d}_i(s)$ maximize $l_{\mathrm{ns}}(\boldsymbol{p})$ given by (2) subject to $p_i > 0$ and $\sum_{i\in s} p_i = 1$, and the $\hat{p}_i(\theta)$ are the values of p_i obtained by maximizing $l_{\mathrm{ns}}(\boldsymbol{p})$ subject to

$$\sum_{i\in s} p_i = 1 \quad \text{and} \quad \sum_{i\in s} p_i y_i = \theta \tag{21}$$

for a fixed θ. The design effect (abbreviated deff) in this case is associated with the estimator $\hat{\overline{Y}}_{\mathrm{H}}$ and is defined as

$$\mathrm{deff}_{\mathrm{H}} = V_p(\hat{\overline{Y}}_{\mathrm{H}})/(S_y^2/n), \tag{22}$$

where S_y^2 is the population variance and $V_p(\cdot)$ denotes the variance under the specified design $p(s)$. Under regularity conditions C1–C3 specified below, the pseudo EL ratio function $r_{\rm ns}(\theta)$ converges in distribution to a scaled χ^2 random variable with one degree of freedom when $\theta = \overline{Y}$, where the scale factor is equal to the design effect $\text{deff}_{\rm H}$ (Wu and Rao, 2006). Hence, the adjusted pseudo EL ratio function

$$r_{\rm ns}^{[a]}(\theta) = \{r_{\rm ns}(\theta)\}/\text{deff}_{\rm H} \tag{23}$$

converges in distribution to a χ^2 random variable with one degree of freedom when $\theta = \overline{Y}$. The regularity conditions are as follows:

- C1: the sampling design $p(s)$ and the study variable y satisfy $\max_{i\in s}|y_i| = o_p(n^{1/2})$, where the stochastic order $o_p(\cdot)$ is with respect to the sampling design $p(s)$.
- C2: the sampling design $p(s)$ satisfies $N^{-1}\sum_{i\in s} d_i - 1 = O_p(n^{-1/2})$.
- C3: the HT estimator $\hat{\theta}_{\rm HT} = N^{-1}\sum_{i\in s} d_i y_i$ of $\theta_0 = \overline{Y}$ is asymptotically normally distributed.

A $(1-\alpha)$-level confidence interval on $\overline{Y}$ can be constructed as $\{\theta \mid r_{\rm ns}^{[a]}(\theta) \le \chi_1^2(\alpha)\}$, where $\chi_1^2(\alpha)$ is the upper α quantile of the χ_1^2 distribution. Finding such an interval, however, involves profile analysis described in Section 5.

For nonstratified sampling designs and a vector of auxiliary variables with known population means $\overline{\boldsymbol{X}}$, the PELL ratio function for $\overline{Y}$ is similarly defined as $r_{\rm ns}(\theta)$ given in (20) but with the benchmark constraints (5) included in finding both $\hat{p}_i$ and $\hat{p}_i(\theta)$. The design effect under this scenario is associated with the estimator $\hat{\overline{Y}}_{\rm GR}$ and is defined as

$$\text{deff}_{\rm GR} = V_p(\hat{\overline{Y}}_{\rm GR})/(S_r^2/n), \tag{24}$$

where $V_p(\hat{\overline{Y}}_{\rm GR}) = V_p\{\sum_{i\in s}\tilde{d}_i(s)r_i\}$, $r_i = y_i - \overline{Y} - \boldsymbol{B}'\boldsymbol{u}_i$, $\boldsymbol{B} = (\sum_{i=1}^N \boldsymbol{u}_i\boldsymbol{u}_i')^{-1}\sum_{i=1}^N \boldsymbol{u}_i y_i$, $\boldsymbol{u}_i = \boldsymbol{x}_i - \overline{\boldsymbol{X}}$, and $S_r^2 = (N-1)^{-1}\sum_{i=1}^N r_i^2$. Under conditions C1–C3 and assuming that C1 and C3 also apply to the components of $\boldsymbol{x}$, the adjusted PELL ratio statistic

$$r_{\rm ns}^{(a)}(\theta) = \{r_{\rm ns}(\theta)\}/\text{deff}_{\rm GR} \tag{25}$$

is asymptotically distributed as χ_1^2 when $\theta = \overline{Y}$ (Wu and Rao, 2006).

Under stratified sampling and with known overall population mean $\overline{\boldsymbol{X}} = \sum_{h=1}^L W_h\overline{\boldsymbol{X}}_h$ but unknown strata means $\overline{\boldsymbol{X}}_h$, the pseudo EL ratio function of $\overline{Y}$ is defined as

$$r_{\rm st}(\theta) = -2\{l_{\rm st}(\hat{\boldsymbol{p}}_1(\theta), \ldots, \hat{\boldsymbol{p}}_L(\theta)) - l_{\rm st}(\hat{\boldsymbol{p}}_1, \ldots, \hat{\boldsymbol{p}}_L)\}, \tag{26}$$

where $\hat{p}_{hi}$ maximize $l_{\rm st}(\boldsymbol{p}_1, \ldots, \boldsymbol{p}_L)$ defined by (3) subject to the set of constraints

$$\sum_{i\in s_h} p_{hi} = 1\,,\ h = 1, \ldots, L \quad \text{and} \quad \sum_{h=1}^L W_h \sum_{i\in s_h} p_{hi}\boldsymbol{x}_{hi} = \overline{\boldsymbol{X}}, \tag{27}$$

and $\hat{p}_{hi}(\theta)$ maximize $l_{\rm st}(\boldsymbol{p}_1, \ldots, \boldsymbol{p}_L)$ subject to (27) plus an additional constraint

$$\sum_{h=1}^L W_h \sum_{i\in s_h} p_{hi} y_{hi} = \theta$$

for a fixed θ. Under suitable regularity conditions on the sampling design and variables involved within each stratum, the adjusted pseudo EL ratio statistic

$$r_{\mathrm{st}}^{[a]}(\theta) = \{r_{\mathrm{st}}(\theta)\}/\mathrm{deff}_{\mathrm{GR(st)}}$$

is asymptotically distributed as χ_1^2 when $\theta = \overline{Y}$ (Wu and Rao, 2006). The design effect $\mathrm{deff}_{\mathrm{GR(st)}}$ is defined through an augmented vector of variables and is given in Section 5.

For each of the three scenarios discussed above, the design effect is defined as a population quantity and needs to be replaced by sample-based estimates for the construction of the pseudo EL ratio confidence intervals. The asymptotic coverage level remains valid if the design effect is consistently estimated; see Wu and Rao (2006) for details on the estimation of design effects.

Pseudo empirical likelihood ratio confidence intervals on $F(t)$ for a given t can be obtained by simply changing y_i to $I(y_i \le t)$. Table 2 contains results on 95% confidence intervals on $\theta_0 = F_N(t)$ at the qth population quantile $t = t_q$ for selected values of q. This was part of an extensive simulation study originally reported in Wu and Rao (2006). The well-known Rao–Sampford unequal probability sampling method is used and the sample size is $n = 80$. Three types of confidence intervals were examined: normal approximation (NA) interval in the form of $(\hat{\theta}_0 - 1.96\{\mathrm{var}(\hat{\theta}_0)\}^{1/2}, \hat{\theta}_0 + 1.96\{\mathrm{var}(\hat{\theta}_0)\}^{1/2})$ with truncation of the LB at 0 or upper bound at 1 if necessary; pseudo EL ratio interval (EL1) without using any additional auxiliary information; and pseudo EL ratio interval (EL2) using the constraint (4). The performance of these intervals is measured in terms of simulated coverage probability (CP), lower (L) and upper (U) tail error rates, and average length (AL).

The message conveyed by Table 2 highlights the advantages of the pseudo EL ratio confidence intervals. For $q = 0.50$ and $t = t_{0.5}$, the population median, the sampling distribution of $\hat{\theta}_0 = \hat{F}_{\mathrm{EL}}(t_{0.5})$ is nearly symmetric. In this case, the NA interval usually performs well in terms of coverage probabilities and balanced tail error rates. But for small or large population quantiles ($q = 0.10$ or 0.90) where the underlying sampling distribution of $\hat{\theta}_0$ is skewed, NA intervals perform poorly: coverage probabilities are lower than the nominal level and tail error rates are not balanced. The EL-based intervals EL1 and EL2, on the other hand, perform well in all cases in terms of coverage probabilities and tail error rates, and EL2 gives shorter average length than EL1. In addition, both the LB and the upper bound of the EL-based intervals automatically locate within the

Table 2
95% confidence intervals for the distribution function

q	CI	CP	L	U	AL
0.10	NA	90.7	0.2	9.1	0.134
	EL1	94.1	1.7	4.2	0.134
	EL2	94.5	1.9	3.6	0.127
0.50	NA	95.3	2.4	2.3	0.212
	EL1	95.5	2.4	2.1	0.208
	EL2	95.4	2.8	1.8	0.187
0.90	NA	93.9	5.0	1.1	0.116
	EL1	95.2	2.7	2.1	0.115
	EL2	93.5	4.0	2.5	0.110

range of the parameter space, (0, 1), which is not always the case for the conventional NA intervals.

5. Computational algorithms

There are three major computational tasks for implementing the EL-based methods: (i) to find the Lagrange multiplier $\boldsymbol{\lambda}$ as the solution to (6) with a single nonstratified sample; (ii) to obtain the MPEL solutions for stratified sampling, raking ratio estimation, and other "irregular" cases; and (iii) to construct the pseudo EL ratio confidence intervals through profiling.

We assume that the population mean $\overline{\boldsymbol{X}}$ is an inner point of the convex hull formed by the sample observations $\{\boldsymbol{x}_i, i \in s\}$ so that a unique solution to (6) exists. Chen et al. (2002) proposed a simple algorithm for solving (6) with guaranteed convergence if the solution exists. The uniqueness of the solution and the convergence of the algorithm are proved based on a duality argument: maximizing $l_{\rm ns}(\boldsymbol{p})$ *with respect to* $\boldsymbol{p}$ subject to $p_i > 0, \sum_{i\in s} p_i = 1$, and the benchmark constraints (5) is a dual problem of maximizing $H(\boldsymbol{\lambda}) = \sum_{i\in s} \tilde{d}_i(s) \log(1 + \boldsymbol{\lambda}'\boldsymbol{u}_i)$ *with respect to* $\boldsymbol{\lambda}$ with no restrictions on $\boldsymbol{\lambda}$. In both cases, the solution $\boldsymbol{\lambda}$ solves the equation system (6). Since $H(\boldsymbol{\lambda})$ is a concave function of $\boldsymbol{\lambda}$ with the matrix of second-order derivatives negative definite, a unique maximum point to $H(\boldsymbol{\lambda})$ exists and can be found using the Newton–Raphson search algorithm. Denoting $\boldsymbol{x}_i - \overline{\boldsymbol{X}}$ by $\boldsymbol{u}_i$, the algorithm of Chen et al. (2002) for solving (6) is as follows.

Step 0: Let $\boldsymbol{\lambda}_0 = \boldsymbol{0}$. Set $k = 0$, $\gamma_0 = 1$, and $\epsilon = 10^{-8}$.

Step 1: Calculate $\Delta_1(\boldsymbol{\lambda}_k)$ and $\Delta_2(\boldsymbol{\lambda}_k)$, where

$$\Delta_1(\boldsymbol{\lambda}) = \sum_{i\in s} \tilde{d}_i(s) \frac{\boldsymbol{u}_i}{1 + \boldsymbol{\lambda}'\boldsymbol{u}_i} \quad \text{and} \quad \Delta_2(\boldsymbol{\lambda}) = \left\{ -\sum_{i\in s} \tilde{d}_i(s) \frac{\boldsymbol{u}_i \boldsymbol{u}_i'}{(1 + \boldsymbol{\lambda}'\boldsymbol{u}_i)^2} \right\}^{-1} \Delta_1(\boldsymbol{\lambda}).$$

If $\|\Delta_2(\boldsymbol{\lambda}_k)\| < \epsilon$, stop the algorithm and report $\boldsymbol{\lambda}_k$; otherwise go to Step 2.

Step 2: Calculate $\boldsymbol{\delta}_k = \gamma_k \Delta_2(\boldsymbol{\lambda}_k)$. If $1 + (\boldsymbol{\lambda}_k - \boldsymbol{\delta}_k)'\boldsymbol{u}_i \le 0$ for some i, let $\gamma_k = \gamma_k/2$ and repeat Step 2.

Step 3: Set $\boldsymbol{\lambda}_{k+1} = \boldsymbol{\lambda}_k - \boldsymbol{\delta}_k$, $k = k + 1$, and $\gamma_{k+1} = (k + 1)^{-1/2}$. Go to Step 1.

It turns out that this modified Newton–Raphson algorithm for a single nonstratified sample is also applicable to stratified samples and a variety of other "irregular" cases after suitable reformulation. Under stratified sampling with known $\overline{\boldsymbol{X}}$, the basic problem is to maximize $l_{\rm st}(\boldsymbol{p}_1, \ldots, \boldsymbol{p}_L)$ subject to (27). If we use $\overline{\boldsymbol{X}}^*$ to denote the augmented $\overline{\boldsymbol{X}}$ to include $W_1, \ldots, W_{L-1}$ as its first $L - 1$ components and $\boldsymbol{x}^*_{hi}$ to denote the augmented $\boldsymbol{x}_{hi}$ to include the first $L - 1$ stratum indicator variables, then the set of constraints (27) can equivalently be rewritten as (Wu, 2004b)

$$\sum_{h=1}^{L} W_h \sum_{i\in s_h} p_{hi} = 1 \quad \text{and} \quad \sum_{h=1}^{L} W_h \sum_{i\in s_h} p_{hi} \boldsymbol{x}^*_{hi} = \overline{\boldsymbol{X}}^*. \tag{28}$$

This reformulation makes all the steps for maximization under nonstratified sampling applicable to stratified sampling. The difference between nonstratified and stratified sampling is simply a matter of single or double summation. Maximizing (3) subject to

(27), or equivalently (28), gives $\tilde{p}_{hi} = \tilde{d}_{hi}(s_h)/(1 + \boldsymbol{\lambda}' \boldsymbol{u}^*_{hi})$, where $\boldsymbol{u}^*_{hi} = \boldsymbol{x}^*_{hi} - \overline{\boldsymbol{X}}^*$ and the vector-valued $\boldsymbol{\lambda}$ is the solution to

$$\sum_{h=1}^{L} W_h \sum_{i \in s_h} \frac{\tilde{d}_{hi}(s_h)\boldsymbol{u}^*_{hi}}{1 + \boldsymbol{\lambda}' \boldsymbol{u}^*_{hi}} = \mathbf{0},$$

which can be solved using the same algorithm of Chen et al. (2002). Using the augmented variables $\boldsymbol{x}^*_i$ and $\overline{\boldsymbol{X}}^*$, the design effect $\text{deff}_{\text{GR(st)}}$ required for the pseudo EL ratio confidence intervals discussed in Section 4.3 is defined as

$$\text{deff}_{\text{GR(st)}} = \left\{ \sum_{h=1}^{L} W_h^2 \, V_p \left(\sum_{i \in s_h} \tilde{d}_{hi}(s_h) r_{hi} \right) \right\} \Big/ \left(\frac{S_r^2}{n} \right),$$

where $r_{hi} = (y_{hi} - \overline{Y}) - (\boldsymbol{B}^*)'(\boldsymbol{x}^*_{hi} - \overline{\boldsymbol{X}}^*)$, $\boldsymbol{B}^*$ is the population vector of regression coefficients similarly defined as $\boldsymbol{B}$ for deff_{GR} given by (24) but using $\boldsymbol{x}^*_i$ and $\overline{\boldsymbol{X}}^*$, and $S_r^2 = (N-1)^{-1} \sum_{h=1}^{L} \sum_{i=1}^{N_h} r_{hi}^2$.

The pseudo EL alternative to raking discussed in Section 4.2 involves finding solutions to (16) or (19), which has the same structure as (6) and can be solved using the same algorithm. Other "irregular" EL-based methods include combining information from multiple surveys (Wu, 2004a), where the algorithm of Chen et al. (2002) once again is the fundamental piece for computational procedures.

Construction of the pseudo EL ratio confidence intervals for $\theta_0 = \overline{Y}$ involves two steps: (i) calculate $r^{[a]}(\theta)$ for a given θ; (ii) find the lower and upper bounds for $\{\theta \mid r^{[a]}(\theta) \leq \chi^2_1(\alpha)\}$. The first step invokes no additional complications since, for instance, the constraint $\sum_{i \in s} p_i y_i = \theta$ for a given θ can be treated as an additional component for the benchmark constraints $\sum_{i \in s} p_i \boldsymbol{x}_i = \overline{\boldsymbol{X}}$. For the second step, the lower and upper bounds of the interval can be found through a simple bisection search method since the interval is confined in between $y_{(1)} = \min_{i \in s} y_i$ and $y_{(n)} = \max_{i \in s} y_i$, and the adjusted pseudo EL ratio function $r^{[a]}(\theta)$ is monotone decreasing for $\theta \in (y_{(1)}, \hat{\overline{Y}}_{\text{EL}})$ and monotone increasing for $\theta \in (\hat{\overline{Y}}_{\text{EL}}, y_{(n)})$. Wu (2005) contains the detailed argument and also provides R/SPLUS functions and codes for several key computational procedures.

6. Discussion

The EL and PEL approaches are flexible enough to handle a variety of other problems. In particular, data from two or more independent surveys from the same target population can be combined naturally through the PEL approach and efficient point estimators, and pseudo EL ratio confidence intervals for the population mean can be obtained. For example, suppose $\{(y_i, \boldsymbol{x}_{1i}, \boldsymbol{z}_i), i \in s_1\}$ are the sample data from the first survey and $\{(\boldsymbol{x}_{2j}, \boldsymbol{z}_j), j \in s_2\}$ are the data from the second survey, where the auxiliary variables $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ have known population means $\overline{\boldsymbol{X}}_1$ and $\overline{\boldsymbol{X}}_2$ but the population mean of the "common" auxiliary variable $\boldsymbol{z}$ is unknown. The goal here is to make inference on the population mean $\overline{Y}$ associated with the first survey, taking advantage of the supplementary data from the two sources. The maximum PEL estimator of $\overline{Y}$ is obtained by maximizing a combined PELL function $l(\boldsymbol{p}_1, \boldsymbol{p}_2)$ based on the two independent samples subject

to the following normalization, benchmarking, and internal consistency constraints (Wu, 2004a):

$$\sum_{i\in s_t} p_{ti} = 1, \quad \sum_{i\in s_t} p_{ti}\boldsymbol{x}_{ti} = \overline{\boldsymbol{X}}_t\,, \quad t = 1, 2, \quad \text{and} \quad \sum_{i\in s_1} p_{1i}\boldsymbol{z}_i = \sum_{i\in s_2} p_{2i}\boldsymbol{z}_i\,.$$

The estimator is given by $\hat{\overline{Y}} = \sum_{i\in s_1} \hat{p}_{1i} y_i$, where the $\hat{p}_{ti}$ $(t = 1, 2)$ maximize $l(\boldsymbol{p}_1, \boldsymbol{p}_2)$. Similarly, the PEL ratio confidence intervals for $\overline{Y}$ can be obtained. The above approach can be extended to handle data from independent samples taken from two or more incomplete frames together covering the population of interest. The PEL approach is flexible in combining data from different sources as demonstrated above. Depending on what is available, new constraints can be added to and existing ones can be removed from the system of constraints.

Another problem of interest is to make inference on the population parameters of interest in the presence of imputation for item nonresponse. Again, the EL and pseudo EL approaches can be applied in a systematic manner to handle imputed data and any auxiliary population information. Recent work has focused on EL inference on the mean, distribution function, and quantiles of a variable of interest y, assuming that an *iid* sample $\{(y_i, x_i), i = 1, \ldots, n\}$ subject to missing y_i is available. The missing y-values are imputed using regression imputation, assuming a missing at random response mechanism and a linear regression model (Qin et al., 2006; Wang and Rao, 2002a). The EL inference using kernel regression imputation, assuming only that the conditional expectation of y given x is a smooth function of x, has likewise been studied (Wang and Chen, 2006; Wang and Rao, 2002b). Various extensions have also been analyzed. The pseudo EL approach can be applied to extend the above work to survey data.

In conclusion, the EL and pseudo EL approaches have several advantages over the traditional approaches to inference from survey data. The advantages include (i) likelihood based motivations; (ii) intrinsic positive weights and efficient point and interval estimation taking account of benchmark and other constraints naturally; (iii) orientation of the confidence intervals is determined by the data and the range of the parameter space is fully preserved, unlike the customary normal theory intervals; and (iv) flexibility in handling a variety of problems in a systematic manner. However, the true empirical likelihood is not available for unequal probability sampling without replacement and other complex designs, and the rationale in Section 4 for using a pseudo empirical likelihood to overcome this problem is not entirely appealing from a theoretical point of view.

Acknowledgments

This work was supported by grants from the Natural Sciences and Engineering Research Council of Canada. We are grateful to Phil Kott for many useful comments and suggestions.

Introduction to Part 5

Gad Nathan and Danny Pfeffermann

1. Preface

The rapid development of sample survey theory and practice over the past 50 years is well represented in this volume. The development has been particularly rapid during the last 20 years, as can be seen by comparison of this volume with its predecessor, Krishnaiah and Rao (1988). Although advances have been made in all aspects of sample surveys during these past 20 years, one of the major developments has been the rapid integration of model-based ideas in mainstream theory and practice of sample survey inference. Twenty years ago, the fundamental divide between advocates of classical design-based inference and design, and those who preferred basing both the sample design and inference only on models was still at its zenith and the controversies of the two previous decades, exemplified by Brewer and Mellor (1973), were raging. The early randomization-based approach, developed by the pioneers of classical design-based sampling theory, was challenged by the study of the logical foundations of estimation theory in survey sampling, for example, Godambe (1955), and by early advocates of pure model-based design and inference, for example, Royall (1970). Although these controversies continued to be fiercely discussed well into the 1980s, see, for example, Hansen et al. (1983), the extreme views have mellowed considerably over the past two decades, and sample survey theory and practice are currently based on a variety of combined approaches, such as model-assisted methods, which integrate models with a randomization-based approach.

The "hybrid" approaches, which are based on both randomization ideas and on models, are well represented in many chapters of this volume, and especially with respect to inference from sample surveys, in Parts 4–6. The introductions to Part 4 and to Part 6 provide a brief survey of the various modes of inference in current use, which are then discussed and exemplified in the different chapters contained in these parts.

While the chapters of Part 4 and Part 6 deal with general methods and approaches to inference from survey data, those of Part 5 consider the application and extension of these general methods to special problems of estimation and inference. The problems considered cover estimation for small areas or domains, the design and analysis of repeated and longitudinal surveys, the analysis of categorical data, inference on distribution functions and quantiles, and graphical presentation of data. This eclectic collection of problems is treated, under a variety of approaches. They range from a pure randomization-based approach (but possibly model-assisted), such as in Lehtonen and Vaijanen (Chapter 31), to pure model-based approaches in Datta (Chapter 32) and

Nathan (Chapter 34). The other four chapters provide examples of a mixed mode approach. Thus, Graubard and Korn (Chapter 37) propose modifications of standard scatter plots by use of randomization weights, but also propose model-assisted methods, such as polynomial regression and kernel smoothing, to improve the graphical display of survey data. Dorfman (Chapter 36) considers both design-based and model-based estimation of distribution functions, emphasizing the necessity for good model diagnostics to ensure robustness to model misspecification. Singh (Chapter 35) considers both pure likelihood-based and quasi-likelihood methods to analyze categorical data, but also their modification by randomization weights to obtain weighted quasi-likelihood inference. Similarly, Steel and McLaren (Chapter 33), who discuss primarily the design of repeated surveys under a randomization approach, also consider the use of time series models, linear models, and correlation models, for the estimation of cross-sectional means and totals, and of sampling errors. We mention in this regard that Kalton (Chapter 5) also deals with the design and analysis of panel surveys and their use for cross-sectional estimation and for estimation of changes, under a design-based approach. This chapter addresses some special issues of panel surveys, such as the effects of changing modes of collection, and the use of weight adjustments for attritions and wave nonresponse.

2. Overview of chapters in Part 5

The first two chapters of Part 5 treat the rapidly developing topic of estimation for small domains or areas, for which only small samples or no samples are available.

Lehtonen and Vaijanen (Chapter 31) consider mostly the estimation of domain totals, but also the estimation of ratios and quantiles, from a design-based (randomization-based) perspective. The emphasis is on the use of design-consistent estimators for which the bias and variance under the randomization distribution tend to zero as the domain size increases. Estimators are obtained by use of calibration techniques, and model-assisted methods via generalized regression estimation (GREG). The first approach is largely model free while the second approach uses a "working model" for the construction of the estimators. A good performance of the estimators under either approach requires the availability of sufficient auxiliary information, preferably at the unit or domain level. Most of the working models considered in this chapter are fixed effects models, but generalizations to generalized linear mixed models with random effects are also considered. The authors distinguish between direct estimators, which only use the data observed for the response variable in the domain of interest and at the time period of interest, and indirect estimators, which use data on the response variable from other domains or time periods. The use of a direct, model-assisted estimator requires fitting the working model for each domain separately. Note that direct estimators are generally design-unbiased or nearly design-unbiased, but often with large design variances, due to the small sample sizes, in which case the requirement for design consistency is of little relevance. Indirect estimators have often a much smaller design variance, because they use observations from many domains, but they may be design biased, such as when using a composite estimator, defined as a linear combination of a GREG and a synthetic estimator that is based on the underlying working model.

The chapter reviews a large number of plausible direct and indirect estimators, as derived under the two general approaches mentioned earlier, with appropriate variance

estimators that take the sampling design into account. An interesting compromise between direct and indirect estimators, obtained by the use of indirect estimators within groups of domains is also considered. The pros and cons of the use of these estimators are discussed and illustrated using real and simulated populations, and the readers are directed to computer software, such as SAS and SUDAAN, where these estimators can be computed.

Datta (Chapter 32) discusses the use of model-based methods for small area estimation. The use of a model defines in an optimal or approximately optimal manner the amount of information borrowed from other areas or past surveys for the construction of the indirect estimators. Unlike in Chapter 31, no adjustments are made to make the estimators approximately design-unbiased over repeated sampling, and the measures of error reflect the error under the model, given the selected sample. In fact, model-based small area estimation generally ignores the sampling design, assuming that the model holding for the sample data also holds for the population from which the sample is taken. The use of a model permits predicting the unknown quantities of interest for areas with no samples, which cannot be taken care of under the design-based approach. Model-based small area estimation can generally be implemented using a frequency-based approach, the full Bayesian approach, or some sort of a "compromise" between the two approaches, known as empirical Bayes. The Bayesian approach is more flexible computationally in terms of the complexity of models that it can accommodate, via the use of Markov Chain Monte Carlo (MCMC) simulations. On the other hand, the reliance on models (and possibly prior distributions) raises the question of the robustness of the estimators to possible model misspecification.

The chapter distinguishes between area-level models that only use area-level values of the auxiliary variables, (e.g., the true area means of these variables), and unit-level models that use unit-level values for at least some of the auxiliary variables (e.g., individual demographic and socio-economic information). Models considered include linear mixed models (including univariate and multivariate cross-sectional time-series models), for which the response variables are continuous, and the more general family of generalized linear mixed models that are used for modeling categorical or count data. Both classes of models incorporate area random effects that account for the variation of the area quantities of interest, not explained by the auxiliary variables. The optimal predictors under the various models in the case of known model parameters are presented, along with the corresponding prediction mean square errors (MSE). When the model parameters are unknown, the corresponding empirical predictors are obtained by replacing the unknown parameters by suitable sample estimates. (Replacing the parameters of the prior distributions for the unknown parameters appearing in the sample distribution by sample estimates in the Bayes predictor, yields the corresponding empirical Bayes predictor.) The use of empirical predictors requires special techniques for estimating the corresponding prediction MSE, and several model-based and resampling procedures are considered. Alternatively, one can use the full Hierarchical Bayes approach that requires in addition fully specified prior distributions for the unknown hyperparameters governing the prior distributions of the parameters appearing in the sample distribution. Application of this approach produces the whole posterior distribution of the small area quantities of interest, so that the prediction MSE and credibility intervals are obtained straightforwardly. The chapter concludes with some brief remarks on other important topics in model-based small area estimation, not covered in this chapter.

Steel and McLaren (Chapter 33) overview the main issues involved in the design and analysis of surveys that are repeated over time. The authors concentrate on the effects of the inter-period correlation structure on the design and analysis of repeated surveys with sample overlap, such as rotating panel surveys and split panel surveys. They consider primarily the extent of the overlap for the estimation of changes in the population level over time and of current means and totals, rather than for the estimation of individual microlevel changes, such as measures of gross flows. In deciding on the extent of the sample overlap and the rotation pattern, conflicting considerations of efficiency and practical considerations have to be balanced. For example, for estimating change, a fixed panel without rotation is the most efficient design from a variance point of view. However, in practice, it is usually necessary to limit the number of times that a person or business is surveyed, to spread the respondent burden and to maintain response rates and the quality of the reported data. Other considerations that need to be taken into account are the additional costs of sampling new units and the necessity to deal with population changes over time. The relative importance of measuring changes over time, versus estimating cross-sectional means for given time points must also be taken into account when designing the overlap and rotation patterns.

As mentioned above, the basic approach in this chapter is design-based. Thus, the randomization variance of linear combinations of estimators for different survey periods, covering measures of change and long-term moving or simple averages, is the basic criterion proposed for evaluating the impact of different rotation patterns. The initial methods of estimation proposed for repeated surveys, BLUE, GREG, and composite estimates, are model-assisted, but are considered in a design-based context. However, the difficulty in obtaining stable randomization-based estimates of the correlation structure leads to the use of ANOVA and time-series models to obtain smooth estimates, which are required for the efficient design of the rotation pattern. For estimation from repeated surveys, a state-space Basic Structural Model, with the associated Kalman filter is proposed. Finally, the case of direct interest in the time series structure of the data is discussed. Methods for the estimation of the trend or the seasonally adjusted series, as well as estimates of the variances of these components, are proposed.

Nathan (Chapter 34) discusses various aspects of the analysis of longitudinal surveys under a model-based approach. In longitudinal surveys, the same units are investigated on several occasions over extensive periods of time. The different modes of collection of longitudinal data, such as prospective measurement, retrospective measurement, observational studies and intervention studies, and the specific problems involved in their implementation and subsequent analysis are discussed. The predominant method of analysis for longitudinal data is based on the fitting of generalized linear models and the use of generalized estimating equations, for estimating the model parameters. Multilevel modeling, often used to describe the hierarchical structure of a population, can be extended to describe the time series relationships between repeated measurements, with the random effects accounting for the variability of higher-level groups, such as households. A model is proposed, which combines standard multilevel models operating at given points in time with a state-space model that represents the time series relationships of the random effects and of the individual measurements. Other general methods of analysis applicable to longitudinal studies and considered in the chapter are as follows: extensions of path analysis, such as graphical chain modeling and structural equation modeling, which provide pictorial representations of the association between

variables, to identify direct and indirect effects of one variable on another; structured and unstructured ante-dependence models, which deal with the inherent nonstationarity of longitudinal data; event history analysis to model the movement of individuals between states; and multivariate counting processes, a flexible framework for modeling event histories.

Chapter 34 considers also special aspects of nonresponse in longitudinal data. Although repeated requests for information from the same individuals or households may lead to attrition and wave nonresponse, the existence of observations for other points in time for the same unit suggests that this information can assist in dealing efficiently with the effects of nonresponse, by considering plausible relationships over time between individual measurements. In this chapter, the focus is on the treatment of missing data resulting from wave nonresponse, where data are available for some points in time and missing for others, rather than on complete nonresponse, which can be dealt with similarly to the ways used for dealing with nonresponse in cross-sectional surveys. In the sample survey context, the effects of nonresponse in a longitudinal survey are accounted for by modifications of the hierarchical linear model, described previously, and by the combined use of time series structures with hierarchical modeling, based on an augmented regression method or on a state-space model. Finally, the chapter examines the effects of informative sampling designs and proposes an extension to longitudinal surveys of a general method of inference on the population distribution under informative sampling, developed for cross-sectional samples.

Singh (Chapter 35) considers methods by which standard quasi-likelihood and quasi-score methods, used for the analysis of categorical data under simple random sampling can be modified, primarily by weighting, to take into account complex sample designs. The methods are evaluated with respect to four basic aspects of data analysis, which are considered throughout the chapter: model selection, model diagnostics, inferential estimation, and inferential testing. The chapter focuses on the quadratic score statistic and its relationship to Pearson's chi-square statistic, rather than on the maximum-likelihood ratio test, in view of the usefulness of the former statistic for the quasi-likelihood approach. Neyman's version of the score function – the nuisance parameter adjusted score – is likewise emphasized to deal with nested hypotheses. Also reviewed are the use of Cholesky residuals to obtain Pearson-type residuals at the cell level, and R^2-type measures of model goodness-of-fit. Next, the results are generalized to the quasi-likelihood framework. This is carried out primarily by the construction of appropriate quasi nuisance parameter adjusted score functions, again considering the four basic aspects of analysis, mentioned above. The semiparametric framework of quasi-likelihood estimation is especially helpful for the analysis of categorical data under a complex sampling design, because it does not require the specification of moments beyond the first two.

To extend the results to the analysis of categorical data obtained under complex sampling, an underlying two-phase randomization scheme is assumed. The first phase generates a finite population from a conceptual super-population model. The second phase selects the sample from the finite population. The weighted quasi-score function is then derived by applying sample weights to the sample estimating function, to obtain an estimate of the finite population estimating function. This leads to the problem of possible instability of the covariance matrix of the weighted quasi-score function. Use of a suitable working covariance matrix followed by Rao–Scott corrections for testing

purposes, and the use of generalized design effects to smooth the covariance matrix for estimation purposes, are reviewed. When covariates are available at the unit level, the use of unit level models is better than the use of aggregate level models from the point of view of estimation efficiency. Likelihood-based methods, quasi-likelihood methods, and weighted quasi-likelihood methods are developed for unit-level models, in a similar way to the methods proposed for aggregate-level models. In particular, for model diagnostics, a Rao–Scott type chi-square approximation to the distribution of the commonly used Hosmer–Lemeshow goodness-of-fit statistic for unit-level models is obtained.

Dorfman (Chapter 36) emphasizes the importance of estimating the finite population distribution function. The distribution function underlies many important statistics, such as quantiles, and is very useful in assessing and comparing finite populations. The predominant approach is design-based, but model-based and model-assisted methods are also considered, with emphasis on protecting the inference against model misspecification. Many desirable properties of estimators of the distribution function are listed: estimation by a proper distribution function satisfying boundary conditions; simplicity and invertibility; calibration to correlated auxiliary variables; efficiency, consistency, and robustness; and simplicity of variance estimation. Although these properties are not entirely compatible, attempts are made to consider them with appropriate priorities. The design-based estimators considered are the well-known weighted Hájek estimator and several alternatives, based on weighted averages of several estimators and on linearly interpolated estimators. When unit level auxiliary information is available, a model-dependent method for estimating the population distribution function is proposed by Chambers and Dunstan (1986). This method is based on estimating the expectation of the distribution function under a hetroscedastic regression model. An alternative model-dependent, design-consistent estimator, designed to protect against possible model misspecification, is the difference estimator of Rao et al. (1990), also constructed with reference to a specific linear regression model. Other alternatives proposed, based on these two approaches, attempt to further robustify the estimation against model misspecification and to enhance diagnostic capabilities. These include weighted averages of the aforementioned estimators, nonparametric regression-based estimators, calibration and GREG estimators, and pseudolikelihood estimators. The performance of these estimators are compared with respect to their variance, consistency, robustness, and diagnostic capabilities. The possibility of using the estimators when only partial auxiliary information, such as means or totals is available, is also discussed.

The primary way of estimating quantiles is by inverting any of the proposed estimates of the distribution function. However, a number of direct estimators, which do not require explicit expressions for the estimates of the distribution function are also proposed. The chapter ends with remarks on the possibilities for variance estimation and the construction of confidence intervals for distribution functions and for quantiles, based on the previously proposed estimators. Variance estimators considered are based on three basic approaches: (1) plug-in model based estimators, (2) design-based estimators, and (3) estimators obtained by replication methods such as the bootstrap and jackknife. Additional remarks relate to the effect of independence assumptions on quantile estimation, and to the effect of measurement error in the variable of interest.

Graubard and Korn (Chapter 37) discuss modifications of standard scatterplots for representing complex survey data. The usefulness of simple scatterplots for data collected in a survey is limited by several factors: individuals in the sample represent

different numbers of individuals in the population; the use of imputations for item nonresponse; the large sample sizes; and intraclass correlations due to cluster sampling. Scatterplots that ignore these features can be misleading. Several modifications to deal with these issues are discussed. In "bubble plots," single dots are replaced by circles with areas that are proportional to the sample weights. Bubble plots are better than a simple scatterplot in describing the population distribution and in identifying influential points. However, for moderate-to-large sample sizes, a bubble plot can be difficult to interpret because of overlapping bubbles. A "sampled scatterplot" is proposed, where the subsample units are selected from the original sample with probabilities that are proportional to the weights, yielding a dataset that represents the population distribution without weighting. Overlapping bubbles can pose a problem in large samples when the data are rounded. A possible solution is "jittering" the data by adding random noise to each data point before plotting. Nonresponse in survey data can affect its graphical presentation. For unit nonresponse, a nonresponse adjustment to the base sample weights can be accounted for in the scatterplot. When there is item nonresponse, which is handled by imputing values for the missing values, a proper variability of the imputed values can be achieved by adding a random error to each imputed value. It is important to distinguish visibly between imputed and nonimputed values in graphical displays.

Displaying smooth curves through a scatterplot that reveal the underlying X–Y structural relationship are useful to the analyst. The large number of observations in surveys allows the use of less model-dependent approaches for fitting these curves. For example, "strip box plots" can be used to show conditional percentiles for data grouped along the x-axis where each box plot displays the sample-weighted percentiles. A more pleasing plot removes the boxes and generates smooth curves through the percentiles using a piecewise cubic spline to fit third degree polynomials. Smooth conditional percentiles curves can be estimated more directly using the original ungrouped data. The primary method is the kernel method, which is first introduced for estimating the conditional mean of the variable of interest, y_i, given the value of an auxiliary variable x. The conditional mean is estimated by a weighted mean of the y_i whose corresponding x_i values are near x. The weights used for this estimator incorporate the sample weights. The choice of the bandwidth, which is critical in determining the smoothness of the resulting conditional mean curve, is discussed. Extensions of the kernel method for estimating the conditional percentiles of y given x are proposed. In addition, a median correction is used to reduce the bias involved in estimating conditional percentiles. Finally, the authors propose several approaches for obtaining standard errors of kernel estimators based on replication methods for variance estimation, such as balanced half-samples and the jackknife.

Sample Surveys: Inference and Analysis, Vol. 29B
ISSN: 0169-7161

© 2009 Elsevier B.V. All rights reserved
DOI: 10.1016/S0169-7161(09)00231-4

Design-based Methods of Estimation for Domains and Small Areas

Risto Lehtonen and Ari Veijanen

1. Introduction

This chapter is devoted to the estimation for population subgroups or domains. Regional areas constructed by administrative criteria, such as county or municipality, are typical *domains of study* (Yates, 1949), also called *domains of interest*. *Estimation for domains*, or *domain estimation* for short, refers to the estimation of population quantities, such as totals or means, for the desired population subgroups. Domain estimation will be examined in the context of *design-based estimation*. Design-based methods for domain estimation are frequently used in many areas of empirical research and official statistics production.

Design-based estimation for a finite population quantity refers to an estimation approach where the randomness is introduced by the sampling design. Thus, the approach also is called *randomization approach*. In design-based estimation, it is emphasized that estimators should be design consistent and, preferably, essentially (or nearly) design unbiased at least in medium-sized samples.

Some early milestones of design-based estimation for domains are Yates (1953, 1960) and Durbin (1958). Hartley (1959) introduced the so-called domain-specific variables for domain estimation with standard design-based estimators of population quantities. This technique has appeared fruitful for example in software development for domain estimation.

We focus on the estimation of *population totals* for domains. Totals are chosen because of their fundamental role in survey sampling and because more complex parameters can often be expressed as functions of totals. The estimation of *ratios* and *quantiles*, such as median, is also discussed. The availability of high-quality *auxiliary information* is crucial for reliable estimation for domains. The reason for incorporating auxiliary data in a domain estimation procedure is obvious: improved accuracy is attained if strong auxiliary data are available for domain estimation.

Different types of auxiliary data can be used in design-based estimation for domains. The available auxiliary data can be aggregated at the population level, at the domain level, or at an intermediate level. Aggregates are often taken from reliable auxiliary sources such as population census or other official statistics; this case is common, for

example in North America. If the auxiliary data are included in a sampling frame, as is the case in many European countries, notably in Scandinavia, the necessary auxiliary totals can be aggregated at the desired level from unit-level data sources.

Calibration techniques and *model-assisted methods* using aggregated auxiliary data offer efficient tools for design-based domain estimation. Calibration is discussed, for example, in Deville and Särndal (1992) and Kott (2003). Särndal (2007) provides a comprehensive treatment of the calibration approach in survey theory and practice. An overview on calibration weighting is given in Chapter 25. Calibration methods were developed for domain estimation in Estevao and Särndal (1999, 2006). The proposed approach to calibration is sometimes called linear or *model-free calibration*. Model-assisted methods using *generalized regression* (GREG) *estimators* were extensively discussed in Särndal et al. (1992). GREG estimation was introduced for domain estimation in Särndal (1981, 1984), Hidiroglou and Särndal (1985), and Särndal and Hidiroglou (1989) and were developed further (including computational tools) in Estevao et al. (1995). We elaborate to some extent these developments; it will appear that the level at which the auxiliary data are used is crucial: efficiency tends to improve when the aggregation level comes close to the domain level when compared to the use of higher-level aggregates.

A statistician also can be in a favorable position to use unit-level auxiliary data for domain estimation. These data are incorporated in the estimation procedure by unit-level statistical models. We illustrate various members of the family of GREG estimators for these cases. For this purpose, we assume that register data (such as population census register, business register, different administrative registers) are available as frame populations and sources of auxiliary data, and the registers contain unique identification keys that can be used in merging at microlevel data from registers and sample surveys. Known domain membership for all population elements is often assumed. Many countries, both in Europe and elsewhere, are progressing in the development of reliable population and business registers that can be accessed for statistical purposes. Obviously, access to micromerged register and survey data provides great flexibility for domain estimation. In GREG estimation, this view has been adopted, for example, in Lehtonen and Veijanen (1998), Särndal (2001), Lehtonen et al. (2003, 2005), and Hidiroglou and Patak (2004). Wu and Sitter (2001a) use unit-level auxiliary information in their *model calibration* method.

Design-consistent estimation for domains contrasts with *model-dependent* estimators, which can have desirable properties under the model but whose design bias does not necessarily tend to zero with increasing sample size (Hansen et al., 1978, 1983; Lehtonen et al., 2003; Särndal, 1984). Design-consistent domain estimators also have been proposed in the context of *model-based* estimation. Model-based and model-dependent methods falling under the headline of *small-area estimation* may be required for the smallest domains (with a small sample size in a domain), where design-based estimators often fail. The methods include a variety of model-based techniques such as synthetic and composite estimators, empirical best linear unbiased predictor (EBLUP) type estimators and various Bayesian techniques. The monograph by Rao (2003a) provides a comprehensive treatment of model-based small-area estimation. Model-based small-area estimation is discussed in Chapter 32.

In design-based estimation, the existence of a model is not necessarily recognized. For example in model-free calibration, an explicit model is not present but exists in

the model calibration method. An assisting or "working" model is postulated in model-assisted estimation. In GREG estimation, the main goal is to obtain favorable design-based properties, such as small design bias. These design-based properties should hold even when the model is misspecified. If our model fits well, decreased design variance is expected for a GREG estimator. Thus, a model is used as an assisting tool in constructing the estimator, which is then modified to meet the desired design-based properties. For example, a GREG estimator for a domain total is often constructed by adding a bias correction term to the sum of fitted values calculated over the population domain. The bias correction term is obtained as a weighted sum of the sample residuals over the domain.

In this chapter, we do not address design-based techniques for nonresponse adjustment (see Chapter 8). Calibration approach to nonresponse treatment is discussed in Särndal and Lundström (2005). Additional topics that are not covered include informative sampling in the context of domain estimation (e.g., Pfeffermann and Sverchkov, 2007) and estimation for domains in the presence of outliers (see Chapter 11).

This chapter is organized as follows. Theoretical framework, terminology, and notation are introduced in Section 2. Section 3 discusses direct estimation for domains by the Horvitz–Thompson (HT) estimator, calibration and GREG estimators. In these cases, domains are often considered as strata in the sampling design. We extend in Section 4 our discussion to more general estimator types and domain structures that are often encountered in practice. GREG estimators for domains are discussed extensively; we also address composite estimation from a design-based perspective. In all these cases, auxiliary information is needed at an aggregated level. Extensions are discussed in Section 5, where a number of empirical examples based on simulation experiments are presented. In these cases, access to unit-level auxiliary data is assumed. Section 6 summarizes some properties of selected software products that can be used for design-based domain estimation.

2. Theoretical framework, terminology, and notation

2.1. Design-based inference at the population level

Let us consider a collection of random variables $(Y_1, Y_2, \ldots, Y_k, \ldots, Y_N)$ with unknown values $(y_1, y_2, \ldots, y_k, \ldots, y_N)$ of a variable of interest y in a *fixed* and *finite population* $U = \{1, 2, \ldots, k, \ldots, N\}$, where k refers to the label of population element. The fixed population is said to be generated from a *superpopulation*. For practical purposes, we are interested in one particular realized population U with $(y_1, y_2, \ldots, y_N)$, not in the more general properties of the model explaining how the population evolved. This is important especially in national statistical agencies, which attempt to describe the current state of the population of a country.

In the design-based approach, the values of the variable of interest are regarded as fixed but unknown quantities. The only source of randomness is the sampling design, and our conclusions should apply to hypothetical repeated sampling from the fixed population.

In estimation for the whole population, we are mainly interested in the total $t = \sum_{k \in U} y_k$ or mean $\bar{y} = \sum_{k \in U} y_k / N$ of the variable y. Notation $\sum_{k \in U}$ refers to summation over all population units $k \in U$. In practice, the values y_k of y are observed

in an n element sample $s \subset U$, which is drawn at random by a sampling design giving probability $p(s)$ to each sample s. The sampling design can be *complex* involving stratification and clustering and several sampling stages.

The design expectation of an estimator $\hat{t}$ of population total t is determined by the probabilities $p(s)$: let $\hat{t}(s)$ denote the value of estimator that depends on y observed in s. Then the expectation is $E(\hat{t}) = \sum_s p(s)\hat{t}(s)$. A *design unbiased* estimator has $E(\hat{t}) = t$. *Design variance* is defined as $\mathrm{Var}(\hat{t}) = \sum_s p(s) \left(\hat{t}(s) - E(\hat{t})\right)^2$. An estimator of design variance is denoted by $\hat{V}(\hat{t})$.

An estimator is *design consistent* if its design bias and variance tend to zero as the sample size increases. An estimator is *nearly design unbiased* if its bias ratio (bias divided by standard deviation) approaches zero with order $O(n^{-1/2})$ when the total sample size n tends to infinity (Estevao and Särndal, 2004). For a nearly design unbiased estimator, the design bias is, under mild conditions, an asymptotically insignificant contribution to the estimator's mean squared error (MSE) (Särndal, 2007, p. 99).

Variance estimators are derived in two steps. First, the theoretical design-based variance $\mathrm{Var}(\hat{t})$ (or its approximation if the theoretical design variance is intractable) is derived. Second, the derived quantity is estimated by a design unbiased or design-consistent estimator $\hat{V}(\hat{t})$.

When the estimator is a weighted sum of observations over sample, it is practical to derive expectation and variance using *inclusion probabilities*. An observation k is included in the sample with probability $\pi_k = P\{k \in s\}$. The inverse probabilities are called *design weights* $a_k = 1/\pi_k$. A useful tool is a sample membership indicator $I_k = I\{k \in s\}$ with value 1 if k is in the sample and 0 otherwise, $E(I_k) = \pi_k$. In variances, we have to consider inclusion of pairs of observations: the probability of including both k and $l (k \neq l)$ is $\pi_{kl} = E(I_k I_l)$ with inverse $a_{kl} = 1/\pi_{kl}$, and $a_{kl} = a_k$ when $k = l$. The covariance of I_k and I_l is $\mathrm{Cov}(I_k, I_l) = \pi_{kl} - \pi_k \pi_l$; this quantity is needed in constructing design variances and their estimators, especially for without-replacement type designs.

2.2. Basic features of design-based inference for domains

2.2.1. Planned and unplanned domain structures

In domain estimation, we are mainly interested in totals or averages of a variable of interest y over D nonoverlapping domains $U_d \subset U$, $d = 1, 2, \ldots, D$, with possibly known domain sizes N_d. As an example, consider the population of a country divided into D domains by regional classification, with N_d households in domain U_d, and the aim is to estimate statistics on household poverty for the regional areas. A domain total is $t_d = t_{dy} = \sum_{k \in U_d} y_k$, where y_k refers to measurement for household k, and domain mean is $\bar{y}_d = t_d / N_d$, $d = 1, \ldots, D$.

Corresponding to population domains, the sample s is divided into subsamples s_d, $d = 1, \ldots, D$. Sampling design may be based on knowledge of domain membership of units in population. If the sampling design is stratified, domains being the strata, the domains are called *planned* (Singh et al., 1994) or *primary domains* (Hidiroglou and Patak, 2004); sometimes also *design domains* (Kish, 1980) or *identified domains* (Särndal, 2007). For planned domain structures, the population domains U_d can be regarded as separate subpopulations. Therefore, standard population estimators are applicable as such. The domain size N_d in every domain U_d is often assumed known and the sample

size n_d in domain sample $s_d \subset U_d$ is fixed in advance. Stratified sampling in connection to a suitable allocation scheme such as optimal (Neyman) or power (Bankier) allocation is advisable in practical applications to obtain control over domain sample sizes (e.g., Lehtonen and Pahkinen, 2004). Singh et al. (1994) describe allocation strategies to attain reasonable accuracy for small domains, still retaining good accuracy for large domains. Falorsi et al. (2006) propose sample balancing and coordination techniques for cases with a large number of different stratification structures to be addressed in domain estimation. If the domain membership is not incorporated into the sampling design, the sizes n_{s_d} of domain samples $s_d = s \cap U_d$ will be random. The domains are then called *unplanned* or *secondary domains*. Unplanned domain structures typically cut across design strata. The property of random domain sample sizes introduces an increase in the variance of domain estimators. In addition, extremely small number (even zero) of sample elements in a domain can be realized if the domain size in the population is small. Unplanned domain structures are commonly encountered in practice because it is impossible to include all relevant domain structures into the sampling design of a given survey.

2.2.2. Extended domain variables of interest

A general tool for domain estimation is the *extended domain variable of interest* y_d defined as $y_{dk} = y_k$ for $k \in U_d$ and $y_{dk} = 0$ for $k \notin U_d$ (Hartley, 1959). In other words, $y_{dk} = I\{k \in U_d\}y_k$. Because $t_d = \sum_{k \in U_d} y_k = \sum_{k \in U} y_{dk}$, we can estimate the domain total of y by estimating the population total of y_{dk} (e.g., Estevao et al., 1995; Estevao and Särndal, 1999; Hidiroglou and Patak, 2004). Consequently, any population total or mean estimator applied to y_{dk} is usable as a corresponding domain estimator. Extended domain variables are useful for estimation for unplanned domains because the contribution of extra variance caused by random domain sample sizes can be easily incorporated in variance expressions. The technique of extended domain variables allows building of generally applicable software for domain estimation and is implemented, for example, in survey sampling oriented SAS procedures and the GES software of Statistics Canada (Estevao et al., 1995).

Extended domain variables can be incorporated in a model-assisted estimation procedure. However, a model fitted to the whole sample is not always going to fit well because most of the y_{dk} are zeroes. But when using extended domain variables, the main interest is not necessarily in the goodness of fit; the primary objective is to attain a single set of weights for all domains. Moreover, the estimates are additive: their sum over the domains equals the estimate for the whole population (Estevao et al., 1995). This can be considered as a benefit of practical importance, especially for routine official statistics production. On the other hand, possible efficiency gains might not be attained and therefore, we usually attempt to derive estimators using the original y_k values.

2.2.3. Direct and indirect estimators

It is advisable to separate direct and indirect estimators for domains. A *direct* estimator uses values of the variable of interest only from the time period of interest and only from units in the domain of interest (Federal Committee on Statistical Methodology, 1993). A HT type estimator $\hat{t}_d = \sum_{k \in s_d} y_k/\pi_k$ provides a simple example of direct estimator. In model-assisted estimation, direct estimators are constructed by using models fitted separately in each domain; an example is a model $Y_k = \mathbf{x}_k'\boldsymbol{\beta}_d + \varepsilon_k$, $k \in U_d$, with domain-specific auxiliary x-data and a vector of regression coefficients

$\boldsymbol{\beta}_d, d = 1, \ldots, D$. A direct domain estimator can still incorporate auxiliary data outside the domain of interest. This is relevant if accurate population data about the auxiliary x-variables are only available at a higher aggregate level.

An *indirect* domain estimator uses values of the variable of interest from a domain and/or time period other than the domain and time period of interest (Federal Committee on Statistical Methodology, 1993). For example, if a linear model $Y_k = \mathbf{x}'_k\boldsymbol{\beta}+\varepsilon_k, k \in U$, with a common vector $\boldsymbol{\beta}$ is used as an assisting model, the resulting domain estimator will be indirect. In general, indirect estimators are attempting to "borrow strength" from other domains and/or in a temporal dimension. The concept of "borrowing strength" is often used in model-based small-area estimation (e.g., Rao, 2003a). Indirect model-assisted estimators for domains are discussed in the literature (e.g., Estevao and Särndal, 1999; Hidiroglou and Patak, 2004; Lehtonen et al., 2003, 2005). Estevao and Särndal (2004) have argued in favor of direct estimators in the context of design-based estimation for domains.

2.2.4. Conditional design-based inference for domains

For unplanned domain structures, observed domain sample sizes can be taken into account in estimation and in theory. We are interested in the average properties of estimators in samples with observed domain sample sizes $\mathbf{n} = (n_1, n_2, \ldots, n_d, \ldots n_D)'$. In conditional design-based inference for domains (Falorsi et al., 2000; Hidiroglou and Patak, 2004; Särndal and Hidiroglou, 1989) given $\mathbf{n}$, the hypothetical repeated sampling yields only samples s with $\mathbf{n}(s) = \mathbf{n}$. This subset of samples, $S_\mathbf{n} = \{s^* \subset U : \mathbf{n}(s^*) = \mathbf{n}\}$, is based on observed information, so it has been considered more relevant than the set of all possible samples. By using conditional probabilities $p_c(s) = p(s)/P\{\mathbf{n}(s^*) = \mathbf{n}\}$, if $\mathbf{n}(s) = \mathbf{n}$, and $p_c(s) = 0$ otherwise, the conditional expectation of an estimator is defined as $E(\hat{t}_d|\mathbf{n}(s) = \mathbf{n}) = \sum_{s:\mathbf{n}(s)=\mathbf{n}} p_c(s)\hat{t}_d(s)$. The conditional MSE and variance are defined in the same way.

We prefer conditionally unbiased estimators to conditionally biased ones. We do not encounter estimators that are conditionally unbiased but unconditionally biased because the unconditional expectation is an average over conditional expectations. The conditional approach may also result in changes in a domain estimation procedure. For example, Falorsi et al. (2000) introduced a HT type estimator and a ratio estimator incorporating conditional inclusion probabilities. Park and Fuller (2005) used conditional inclusion probabilities for a calibrated GREG estimator.

The estimator of the conditional variance is, in general, different from the estimator of the unconditional variance. Conditional variance estimate yields a conditional confidence interval. In repeated sampling from the subset $S_\mathbf{n}$, the conventional t-based conditional confidence interval covers the true value approximately at a given rate if the estimator is approximately normally distributed. Because this holds for all values of $\mathbf{n}$, the conditional confidence interval is also an unconditional confidence interval with the same coverage rate. If the model is only approximately correct, a model-assisted method does not always yield conditionally valid inference. It can be argued (Rao, 1997) that model-assisted approach should be restricted to methods with good conditional properties. Conditional inference has been based on other properties besides domain sizes; there are examples of conditioning on strata sample sizes (Holt and Smith, 1979) and on HT estimates of the auxiliary variables (Montanari and Ranalli, 2002; Rao, 1985).

2.2.5. Design-based properties of domain estimators

Known design-based properties related to bias and accuracy of model-assisted estimators are summarized in Table 1. For comparison, design-based properties of corresponding model-dependent estimators are also included in the table. Model-assisted estimators such as GREG are design consistent or nearly design unbiased by definition, but their variance can become large in domains where the sample size is small. Model-dependent estimators such as synthetic and EBLUP estimators are design biased: the bias can be large for domains where the model does not fit well. The variance of a model-dependent estimator can be small even for small domains, but the accuracy can be poor if the squared bias dominates the MSE, as shown, for example, by Lehtonen et al. (2003, 2005). For a model-dependent estimator, the dominance of the bias component together with a small variance can cause poor coverage rates and invalid design-based confidence intervals. For design-based model-assisted estimators, on the other hand, valid confidence intervals can be constructed. Typically, model-assisted estimators are used for major or not-so-small domains, and model-dependent estimators are used for small domains where model-assisted estimators can fail.

Table 1 indicates that small domains present problems in the design-based approach. Purcell and Kish (1980) call domain a minidomain when $N_d/N < 1\%$. In such small domains, especially, direct estimators can have large variance. Small domains are the main reason to prefer indirect model-based estimators to design-based estimators (Rao, 2005). By proper planning of the sampling strategy, it is possible to decrease the variance of a design-based estimator in the small domains. Singh et al. (1994) and Marker (2001) give examples of such strategies.

In practice, there are two main approaches to design-based estimation for domains: direct estimators that are usually applied for planned domain structures and indirect estimators whose natural applications are for unplanned domains. The two main approaches are discussed in Sections 3 and 4, respectively.

Table 1
Design-based properties of model-assisted and model-dependent estimators for domains and small areas

	Design-based model-assisted methods	Model-dependent methods
	GREG and calibration estimators	Synthetic and EBLUP estimators
Bias	Design unbiased (approximately) by the construction principle	Design biased Bias does not necessarily approach zero with increasing domain sample size
Precision (Variance)	Variance may be large for small domains Variance tends to decrease with increasing domain sample size	Variance can be small even for small domains Variance tends to decrease with increasing domain sample size
Accuracy (MSE)	MSE = Variance (or nearly so)	MSE = Variance + squared bias Accuracy can be poor if the bias is substantial
Confidence intervals	Valid design-based intervals can be constructed	Valid design-based intervals not necessarily obtained

3. Direct estimators for domain estimation

The HT type estimator does not incorporate auxiliary information. GREG estimation is assisted by a model fitted at the domain level and uses auxiliary data from the domain. Calibration incorporates auxiliary data from the domain of interest or from a higher-level aggregate. All these estimators are direct because the y-values are taken from the domain of interest. When domain membership is known for all population elements, domain sizes N_d are also known.

3.1. Horvitz–Thompson estimator

The basic design-based direct estimator of the domain total t_d is the HT estimator, also known as the Narain-Horvitz-Thompson (NHT) and the *expansion estimator*:

$$\hat{t}_{d\mathrm{HT}} = \sum_{k\in U_d} I_k y_k/\pi_k = \sum_{k\in s_d} y_k/\pi_k = \sum_{k\in s_d} a_k y_k \tag{1}$$

(Horvitz and Thompson, 1952; Narain, 1951; notation as in Section 2.1). HT estimates of domain totals are additive: they sum up to the HT estimator $\hat{t}_{\mathrm{HT}} = \sum_{k\in s} a_k y_k$ of the population total. As $E(I_k) = \pi_k$, the HT estimator is design unbiased for t_d. Under mild conditions on the π_k, the corresponding mean estimator $\hat{t}_{d\mathrm{HT}}/N_d$ is also design consistent (Isaki and Fuller, 1982). The estimator $\hat{t}_{d\mathrm{HT}}$ has design variance

$$\begin{aligned}\mathrm{Var}\big(\hat{t}_{d\mathrm{HT}}\big) &= E\left(\sum_{k\in U_d} \frac{I_k-\pi_k}{\pi_k} y_k\right)^2 = \sum_{k\in U_d}\sum_{l\in U_d} E(I_k-\pi_k)(I_l-\pi_l)\,\frac{y_k}{\pi_k}\,\frac{y_l}{\pi_l} \\ &= \sum_{k\in U_d}\sum_{l\in U_d} (\pi_{kl}-\pi_k\pi_l)\frac{y_k}{\pi_k}\,\frac{y_l}{\pi_l} = \sum_{k\in U_d}\sum_{l\in U_d} (a_k a_l/a_{kl}-1)y_k y_l. \end{aligned} \tag{2}$$

From $a_{kl}E(I_k I_l) = 1$, we see that an unbiased estimator for the design variance is

$$\hat{V}\big(\hat{t}_{d\mathrm{HT}}\big) = \sum_{k\in U_d}\sum_{l\in U_d} a_{kl} I_k I_l (a_k a_l/a_{kl}-1) y_k y_l = \sum_{k\in s_d}\sum_{l\in s_d} (a_k a_l - a_{kl}) y_k y_l. \tag{3}$$

An alternative Sen–Yates–Grundy formula for fixed sample size designs is (Sen, 1953; Yates, 1953):

$$\begin{aligned}\hat{V}\big(\hat{t}_{d\mathrm{HT}}\big) &= -\sum_{k\in s_d}\sum_{l<k;\,l\in s_d} a_{kl}(\pi_{kl}-\pi_k\pi_l)(a_k y_k - a_l y_l)^2 \\ &= \sum_{k\in s_d}\sum_{l<k;\,l\in s_d} (a_{kl}/a_k a_l - 1)(a_k y_k - a_l y_l)^2. \end{aligned}$$

These variance estimators are impractical because they contain second-order inclusion probabilities π_{kl} whose computation is often laborious for practical purposes. Hájek (1964) and Berger (2004, 2005b) proposed approximations to π_{kl}. Särndal (1996) developed efficient strategies with simple variance estimators under fixed sample size probability proportional-to-size (πPS) schemes, including a combination of Poisson sampling or stratified simple random sampling without replacement (SRSWOR) with

GREG estimation. Berger and Skinner (2005) proposed a jackknife variance estimator and Kott (2006a) introduced a delete-a-group jackknife variance estimator for πPS designs. The SAS procedure SURVEYSELECT is able to compute π_{kl} under certain unequal probability without-replacement sampling designs. Some software products can incorporate the π_{kl} into variance estimation procedures; an example is the SUDAAN software. The SAS macro CLAN includes the Sen–Yates–Grundy formula. Such estimators are discussed in Chapter 2.

Many πPS designs allow using of Hájek approximation (Berger, 2004, 2005b; Hájek, 1964) of second-order inclusion probabilities by $\pi_{kl} \approx \pi_k \pi_l \left[1 - (1 - \pi_k)(1 - \pi_l) m_d^{-1}\right]$ for $k \neq l$, where $m_d = \sum_{i \in U_d} \pi_i (1 - \pi_i)$. The approximation is used in a simple variance estimator $\hat{V}\left(\hat{t}_{d\mathrm{HT}}\right) = \sum_{k \in s_d} c_k e_k^2$, where $c_i = n_d (n_d - 1)^{-1} (1 - \pi_i)$ and $e_k = a_k y_k - \left(\sum_{i \in s_d} c_i\right)^{-1} \sum_{i \in s_d} c_i a_i y_i$.

For unequal probability sampling designs, the variance of the ordinary HT estimator has been approximated under a with-replacement (WR) assumption, leading to Hansen–Hurwitz (1943) type variance estimator (Lehtonen and Pahkinen, 2004, p. 228, and SAS procedure SURVEYMEANS) given by

$$\hat{V}_A(\hat{t}_{d\mathrm{HT}}) = \frac{1}{n_d(n_d - 1)} \sum_{k \in s_d} (n_d a_k y_k - \hat{t}_{d\mathrm{HT}})^2. \tag{4}$$

For unplanned domains, the variance estimator for HT should account for random domain sizes. An approximate variance estimator applied, for example, in SAS procedure SURVEYMEANS contains extended domain variables y_{dk}:

$$\hat{V}_U(\hat{t}_{d\mathrm{HT}}) = \frac{n}{n - 1} \sum_{k \in s} (a_k y_{dk} - \hat{t}_d / n)^2, \tag{5}$$

where n is the total sample size. Under SRSWOR, an alternative to (5) is

$$\hat{V}_{\mathrm{srswor}}(\hat{t}_{d\mathrm{HT}}) = N^2 \left(1 - \frac{n}{N}\right) \left(\frac{1}{n}\right) p_d \hat{s}_{dy}^2 \left(1 + \frac{q_d}{\mathrm{c.v}_{dy}^2}\right),$$

where $p_d = n_{s_d}/n$, $q_d = 1 - p_d$, variance estimator is, $\hat{s}_{dy}^2 = \sum_{k \in s_d} (y_k - \bar{y}_d)^2 / (n_{s_d} - 1)$, and estimated coefficient of variation is $\mathrm{c.v}_{dy} = \hat{s}_{dy}/\bar{y}_d$ for $\bar{y}_d = \sum_{k \in s_d} y_k / n_{s_d}$.

The HT estimator can be regarded as a model-dependent estimator under a model $Y_k = \beta \pi_k + \pi_k \varepsilon_k$ (Zheng and Little, 2003). HT is nearly optimal estimator among weighted sums of Y values when Y depends on scalar x as $E(Y_k) = \beta x_k$, the variance of errors is proportional to x_k^2, and the sampling design assigns π_k proportional to x_k. On the other hand, HT is very inefficient when the intercept of the model is far from zero. Disastrous results are possible in HT estimation, as the famous example of Basu (1971) shows (e.g., citation in Little, 2004).

If the domain size N_d is known, we expect better results with a "Hájek" type direct estimator $\hat{t}_{dH(N)} = N_d \hat{\bar{y}}_d$ (e.g., Hidiroglou and Patak, 2004; Särndal et al., 1992, p. 391) derived from the domain mean $\hat{\bar{y}}_d = \sum_{k \in s_d} a_k y_k / \hat{N}_d$ with $\hat{N}_d = \sum_{k \in s_d} a_k$. This is a special case of ratio estimation (Section 4.3.1). The variance of $\hat{t}_{dH(N)}$ is estimated by

$$\hat{V}(\hat{t}_{dH(N)}) = \left(\frac{N_d}{\hat{N}_d}\right)^2 \sum_{k \in s_d} \sum_{l \in s_d} (a_k a_l - a_{kl})(y_k - \hat{\bar{y}}_d)(y_l - \hat{\bar{y}}_d). \tag{6}$$

3.2. Population fit regression estimator

The population fit regression estimator is a theoretical tool used in approximating real-world estimators. We first consider *difference estimators* (Särndal, 1980; Särndal et al., 1992, p. 221). If known values y_k^0 are close to y_k, we write the estimable population total as

$$t = \sum_{k\in U} y_k = \sum_{k\in U} y_k^0 + \sum_{k\in U} (y_k - y_k^0).$$

A difference estimator is defined by estimating the second sum using HT:

$$\hat{t}_{\mathrm{DIFF}} = \sum_{k\in U} y_k^0 + \sum_{k\in s} a_k (y_k - y_k^0).$$

As the y_k^0 are constants, $\hat{t}_{\mathrm{DIFF}}$ is unbiased for t.

Consider a regression superpopulation model $Y_k = \mathbf{x}'_k\boldsymbol{\beta} + \varepsilon_k$, where $\mathbf{x}_k = (1, x_{1k}, \ldots, x_{Jk})'$ is the vector of auxiliary x-variables, $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_J)'$ is the vector of regression coefficients, and ε_k are the residuals with variances $\sigma_k^2 = \mathrm{Var}(\varepsilon_k)$. Hypothetically, we can fit the model to the population by calculating generalized least squares (GLS) estimator $\mathbf{B} = \hat{\boldsymbol{\beta}}$ as

$$\mathbf{B} = \left(\sum_{k\in U} \frac{\mathbf{x}_k \mathbf{x}'_k}{\sigma_k^2}\right)^{-1} \left(\sum_{k\in U} \frac{\mathbf{x}_k y_k}{\sigma_k^2}\right).$$

In practice, the error variance $\mathrm{Var}(\varepsilon_k) = \sigma_k^2$ can often be assumed constant, $\sigma_k^2 = \sigma^2$, and then it cancels out. When the variance varies between observations, the σ_k^2 should be included in the estimators. Straightforward cases are known σ_k^2 or an assumption that the variances differ by known constants c_k such that $\sigma_k^2 = c_k\sigma^2$. A special case is when $c_k = 1$ for all $k \in U$. For more details on the treatment of σ_k^2, see, for example, Särndal et al. (1992, p. 229 and Chapter 7).

A difference estimator with fitted values $\hat{y}_k^0 = \mathbf{x}'_k\mathbf{B}$ defines the *population fit regression estimator*,

$$\hat{t}_{\mathrm{REG}} = \sum_{k\in U} \hat{y}_k^0 + \sum_{k\in s} a_k (y_k - \hat{y}_k^0).$$

If an estimator $\hat{t}$ can be well approximated by $\hat{t}_{\mathrm{REG}}$, then $\mathrm{Var}(\hat{t})$ can be estimated by a sample-based estimator of

$$\mathrm{Var}(\hat{t}_{\mathrm{REG}}) = \mathrm{Var}\left(\sum_{k\in s} a_k E_k\right) = \sum_{k\in U}\sum_{l\in U} (a_k a_l / a_{kl} - 1) E_k E_l,$$

where $E_k = y_k - \hat{y}_k^0$ are the population fit residuals. To estimate $\mathrm{Var}(\hat{t}_{\mathrm{REG}})$ from sample, we replace the E_k by corresponding sample residuals $e_k = y_k - \mathbf{x}'_k\hat{\mathbf{B}}$. If $\hat{\mathbf{B}}$ is nearly unbiased for $\mathbf{B}$, we can verify using $E(a_{kl} I_k I_l) = 1$ that a nearly unbiased estimator for $\mathrm{Var}(\hat{t}_{\mathrm{REG}})$ is

$$\hat{V}(\hat{t}_{\mathrm{REG}}) = \sum_{k\in s}\sum_{l\in s} (a_k a_l - a_{kl}) e_k e_l. \tag{7}$$

One approach to estimate $\mathbf{B}$ is to plug in HT estimators of both of its sum components. When σ_k^2 is constant, we use a weighted least squares (WLS) estimator

$$\hat{\mathbf{B}} = \left(\sum_{k\in s} a_k \mathbf{x}_k \mathbf{x}'_k\right)^{-1} \left(\sum_{k\in s} a_k \mathbf{x}_k y_k\right).$$

This estimator is only approximately unbiased due to its nonlinearity. Another approach is to consider the population maximum likelihood (ML) estimator maximizing $f(\boldsymbol{\beta}) = -\sum_{k\in U} \left(y_k - \mathbf{x}'_k\boldsymbol{\beta}\right)^2/\sigma^2$. As only the sample is available, we use an estimated log-likelihood, the so-called *pseudolikelihood*, instead (Binder, 1983; Godambe and Thompson, 1986a; Nordberg, 1989). The function $f(\boldsymbol{\beta})$ is estimated by an unbiased HT type estimator $\hat{f}(\boldsymbol{\beta}) = -\sum_{k\in s} a_k \left(y_k - \mathbf{x}'_k\boldsymbol{\beta}\right)^2/\sigma^2$. This function is maximized by $\hat{\mathbf{B}}$. Robust alternatives are presented in Beaumont and Alavi (2004).

Särndal et al. (1992) and Estevao and Särndal (2006) have approximated GREG and calibration estimators (Sections 3.3 and 3.4) by Taylor linearization yielding a population fit regression estimator. Because many approximations are involved, the resulting variance estimators are at least slightly biased.

3.3. GREG estimators

The GREG estimator is a sample-based substitute for the population fit regression estimator (Section 3.2). A direct type GREG estimator of domain total t_d is assisted by a regression model $Y_k = \mathbf{x}'_k\boldsymbol{\beta}_d + \varepsilon_k$, $\mathrm{Var}(\varepsilon_k) = \sigma_k^2$. Assuming constant error variance σ_k^2, the domain-specific parameter $\mathbf{B}_d$ of the population fit defined for U_d is estimated as in Section 3.2 by

$$\hat{\mathbf{B}}_d = \left(\sum_{k\in s_d} a_k \mathbf{x}_k \mathbf{x}'_k\right)^{-1} \left(\sum_{k\in s_d} a_k \mathbf{x}_k y_k\right),$$

and the fitted values $\hat{y}_k = \mathbf{x}'_k\hat{\mathbf{B}}_d$ and residuals $e_k = y_k - \hat{y}_k$ are incorporated into the GREG estimator

$$\hat{t}_{d\mathrm{GREG}} = \sum_{k\in U_d} \hat{y}_k + \sum_{k\in s_d} a_k(y_k - \hat{y}_k) = \sum_{k\in U_d} \hat{y}_k + \sum_{k\in s_d} a_k e_k \tag{8}$$

(Särndal, 1980; Särndal et al., 1992). The first part in $\hat{t}_{d\mathrm{GREG}}$, the population sum of fitted values over the domain, is sometimes called a synthetic estimator (Särndal, 1984). When compared with direct GREG, it may have smaller variance but possibly large design bias. The weighted sum of residuals tends to correct for the design bias. In some cases, however, the weighted sum of the residual terms is zero. This happens when the model contains an intercept.

Rearranging the terms of GREG we obtain the traditional regression estimator

$$\hat{t}_{d\mathrm{GREG}} = \hat{t}_{d\mathrm{HT}} + (\mathbf{t}_{dx} - \hat{\mathbf{t}}_{dx})'\hat{\mathbf{B}}_d,$$

where $\mathbf{t}_{dx} = \sum_{k\in U_d} \mathbf{x}_k = \left(N_d, \sum_{k\in U_d} \mathbf{x}_{1k}, \ldots, \sum_{k\in U_d} \mathbf{x}_{Jk}\right)'$ and $\hat{\mathbf{t}}_{dx} = \sum_{k\in s_d} a_k \mathbf{x}_k$. By Taylor linearization, $\hat{t}_{d\mathrm{GREG}}$ is approximated by a population fit regression estimator

$\hat{t}_{d\mathrm{REG}} = \hat{t}_{d\mathrm{HT}} + (\mathbf{t}_{dx} - \hat{\mathbf{t}}_{dx})'\mathbf{B}_d$ applied in U_d. The estimator $\hat{t}_{d\mathrm{REG}}$ is unbiased for t_d, and so the GREG estimator is nearly unbiased. Although GREG incorporates a model, it is model-assisted, not model-dependent, because the model only yields a fixed population quantity $\mathbf{B}_d$, and GREG is nearly design unbiased even when the model is not valid. By (7), the variance of $\hat{t}_{d\mathrm{GREG}}$ can be estimated using sample residuals $e_k = y_k - \mathbf{x}'_k\hat{\mathbf{B}}_d$:

$$\hat{V}_1(\hat{t}_{d\mathrm{GREG}}) = \sum_{k\in s_d}\sum_{l\in s_d}(a_k a_l - a_{kl})e_k e_l. \tag{9}$$

The GREG estimator can be written as a weighted sum of observations incorporating so-called g-weights:

$$\hat{t}_{d\mathrm{GREG}} = \sum_{k\in s_d} a_k g_{dk} y_k;\ g_{dk} = I_{dk} + I_{dk}(\mathbf{t}_{dx} - \hat{\mathbf{t}}_{dx})'\hat{\mathbf{M}}_d^{-1}\mathbf{x}_k,$$

where $\hat{\mathbf{M}}_d = \sum_{i\in s_d} a_i\mathbf{x}_i\mathbf{x}'_i$ and $I_{dk} = I\{k \in U_d\}$ is the domain membership indicator. The g-weights are used in a variance estimator

$$\hat{V}_2(\hat{t}_{d\mathrm{GREG}}) = \sum_{k\in s_d}\sum_{l\in s_d}(a_k a_l - a_{kl})g_{dk}e_k g_{dl}e_l \tag{10}$$

(Hidiroglou and Patak, 2004; Särndal et al., 1989 and 1992, p. 235). In practice, $\hat{V}_1$ and $\hat{V}_2$ often yield similar results but $\hat{V}_2$ in (10) is preferable (Fuller, 2002; Särndal et al., 1989).

3.4. Calibration estimators

Calibration is based on information about known totals of auxiliary variables $\mathbf{x}_k$, also called *benchmark variables*, at an aggregate level. In model-free calibration (Särndal, 2007) discussed here, it is not necessary to impose a model on the data. Suppose the population is divided into *calibration groups* U_c ($c = 1, 2, \ldots, C$) so that every domain U_d is contained within one of the groups and the population totals $\mathbf{t}_{cx} = \sum_{k\in U_c}\mathbf{x}_k$ of auxiliary variables are known. The domain totals $\mathbf{t}_{dx}$ are not required. Direct *calibration estimator* of the domain total t_d is a weighted sum of observations:

$$\hat{t}_{d\mathrm{CAL}} = \sum_{k\in s_d} w_k y_k,$$

where the *calibration weights* w_k have to satisfy the *calibration equation*s

$$\sum_{k\in s_c} w_k\mathbf{x}_k = \sum_{k\in U_c}\mathbf{x}_k = \mathbf{t}_{cx}$$

for every calibration group. It follows immediately that calibration estimator applied to the auxiliary data yields the known totals. We therefore expect that the weighted sum of y over s_d is close to t_d.

There are two main approaches to calibration, one based on a *distance measure* and the other based on *instrument vectors* (Chapter 25). In the distance measure approach, the weights w_k minimize a distance to the design weights a_k, subject to the calibration equations (Deville and Särndal, 1992; Singh and Mohl, 1996). An example of a

calibration estimator incorporating an instrument vector $\mathbf{z}_k$ is

$$\hat{t}_{d\mathrm{CAL}} = \sum_{k \in s_d} a_k (1 + \boldsymbol{\lambda}' \mathbf{z}_k) y_k,$$

where $\boldsymbol{\lambda}' = (\mathbf{t}_{cx} - \hat{\mathbf{t}}_{cx})' \left(\sum_{k \in s_c} a_k \mathbf{z}_k \mathbf{x}'_k \right)^{-1}$. It should be noted that the values of instrument z-variables need to be known only for the sample (or need to be estimated); they are not necessarily treated as proper auxiliary information in the same manner as the auxiliary x-variables. For practical purposes, a natural choice is $\mathbf{z}_k = \mathbf{x}_k$; an optimal choice is discussed in Estevao and Särndal (2004).

As in (7), the variance of $\hat{t}_{d\mathrm{CAL}}$ is estimated by

$$\hat{V}(\hat{t}_{d\mathrm{CAL}}) = \sum_{k \in s_c} \sum_{l \in s_c} (a_k a_l - a_{kl})(y_{dk} - \mathbf{x}'_{ck} \hat{\mathbf{B}}_{cd})(y_{dl} - \mathbf{x}'_{cl} \hat{\mathbf{B}}_{cd}),$$

where $\mathbf{x}_{ck} = I\{k \in U_c\} \mathbf{x}_k$ (Estevao and Särndal, 2006), and

$$\hat{\mathbf{B}}_{cd} = \left(\sum_{k \in s_c} a_k \mathbf{z}_k \mathbf{x}'_{ck} \right)^{-1} \left(\sum_{k \in s_c} a_k \mathbf{z}_k y_{dk} \right).$$

When U_c is much larger than U_d, the variance can become large. Therefore, we should attempt to find a calibration group that agrees closely with the domain of interest.

Our GREG estimator of Section 3.3 is actually a special case of calibration, sometimes called linear calibration estimator, as the weights $a_k g_{dk}$ minimize a certain chi-square distance to design weights a_k, subject to domain-level calibration equations $\sum_{k \in s_d} a_k g_{dk} \mathbf{x}_k = \mathbf{t}_{dx}$.

Calibration is contrasted with GREG estimation in Särndal (2007). Särndal and Lundström (2005) discuss calibration in the context of adjustment for unit nonresponse in sample surveys.

3.5. Computational example with direct estimation under a planned domain structure

In this section, we demonstrate with real data the direct Horvitz–Thompson, Hájek, and GREG estimation of totals for domains. The data set contains disposable income of households in $D = 12$ regions of Western Finland. The population consists of $N = 431{,}000$ households. In addition to the income data, the record of a household shows the number of household members who had higher education (variable EDUC) and the number of months in total the household members were employed (EMP) during last year. All three variables were determined using administrative registers. For this computational exercise, we had access to population level information on all variables. This gives a possibility to compare sample estimates to the known population values.

We were interested in the yearly total disposable income $t_d = \sum_{k \in U_d} y_k$ in the regions $U_d (d = 1, \ldots, D)$. A sample of 1000 households was drawn from the population by using stratified πPS (without-replacement type probability proportional to size sampling) with household size as the size variable. To demonstrate estimation for planned domains, we interpret here the sample as a stratified sample where the regions constitute the strata. Thus, the domain structure is of planned type, where the regional sample sizes are considered fixed by the sampling design. In Section 4.2, we use the same sample

in estimation for unplanned domains, where the regional sample sizes are considered random.

In Table 2, we grouped the domains by sample size into minor ($8 \leq n_d \leq 33$), medium-sized ($34 \leq n_d \leq 45$) and major ($46 \leq n_d \leq 277$) domains, where n_d is the observed domain sample size in domain U_d. There were four domains in each domain size class.

Results are shown in Table 2. The absolute relative error of an estimator in domain d is calculated as $|\hat{t}_d - t_d|/t_d$ and domain group's MARE is the mean of absolute relative errors over domains in the group. Correspondingly, MCV is the mean coefficient of variation of the estimate over domain group. The coefficient of variation is calculated as $\text{s.e}(\hat{t}_d)/\hat{t}_d$, where s.e refers to the estimated standard error of an estimator. For variance estimation, we approximated the design by with-replacement type probability-proportional-to-size sampling (PPS). The variance estimators for ordinary HT (column 1) and the Hájek type estimator (column 2) were defined by (4) and (6), respectively. The Hájek estimator, which contains the known domain sizes N_d, yielded better results than ordinary HT.

A calibration estimator, the direct GREG estimator with linear assisting model,

$$Y_k = \beta_{0d} + \beta_{1d}\text{EMP}_k + \varepsilon_k (\text{column 3}) \text{ or}$$
$$Y_k = \beta_{0d} + \beta_{1d}\text{EMP}_k + \beta_{2d}\text{EDUC}_k + \varepsilon_k (\text{column 4}),$$

and variance estimator (10) incorporated the known domain sizes and domain totals of EMP (column 3) and EDUC (column 4). The model parameters were estimated by WLS with weights $a_k = 1/\pi_k$. By GREG, we obtained clearly smaller MARE and MCV figures than by HT.

Adding information in the estimation procedure improved the results until the assisting model contained both EMP and EDUC: inclusion of EDUC in GREG decreased MCV but average errors did not always decrease. In large domains, the average error and MCV were usually smaller than in small domains.

Table 2
Mean absolute relative error (MARE) and mean coefficient of variation (MCV) of direct HT, Hájek, and calibration (GREG) estimators of totals for minor, medium-sized, and major domains by using various amounts of auxiliary information in a planned domains case

	HT		Hájek		Calibration (GREG)			
	1		2		3		4	
Auxiliary Information	None		Domain Sizes		Domain Sizes and Domain Totals of EMP		Domain Sizes and Domain Totals of EMP and EDUC	
Domain sample size class	MARE (%)	MCV (%)	MARE (%)	MCV (%)	MARE (%)	MCV (%)	MARE (%)	MCV (%)
Minor $8 \leq n_d \leq 33$	11.5	11.9	5.3	10.9	5.8	7.7	6.4	6.8
Medium $34 \leq n_d \leq 45$	7.6	9.0	6.4	9.0	3.7	8.0	3.6	8.1
Major $46 \leq n_d \leq 277$	12.5	5.2	4.7	5.6	4.3	4.7	5.2	3.7

4. Indirect estimators in domain estimation

4.1. Generalized regression estimators

4.1.1. Linear GREG

Indirect estimators use y-values also from other domains than the domain of interest. While direct estimators can be derived from corresponding estimators for population, indirect estimators require new results. This holds for unplanned domain structures in particular, but the methodology below applies also to planned domains when indirect estimators are used, for example, when the GREG estimator is assisted by a model fitted to the whole sample. Thus, direct estimators can be treated as a special case of indirect estimators. If the auxiliary information is not available at the domain level but at a higher aggregate level, or if the population frame does not include domain membership data, the calibration approach might be preferred to GREG.

We first assume a common linear fixed-effects regression model $Y_k = \mathbf{x}'_k\boldsymbol{\beta} + \varepsilon_k$ for all domains. The corresponding population fit parameter $\mathbf{B}$ (Section 3.2) is estimated as in Section 3.2. The linear GREG estimator of domain total t_d incorporates fitted values $\hat{y}_k$ of the common model:

$$\hat{t}_{d\mathrm{GREG}} = \sum_{k\in U_d} \hat{y}_k + \sum_{k\in s_d} a_k(y_k - \hat{y}_k), \tag{11}$$

where $\hat{y}_k = \mathbf{x}'_k\hat{\mathbf{B}}$. In general, this is an indirect estimator, since all y-values in the sample contribute.

There is a whole spectrum of model types describing various assumptions about differences between domains (e.g., Lehtonen et al., 2003). If the domains are assumed similar enough, the model may contain only intercept and slopes common to all domains. At the other end of the spectrum, the model is equivalent to a set of separate models for each domain, and all estimators are of direct type. A more parsimonious model might have separate parameters for the largest domains and common parameters for the small domains. It is also possible to use a model formulation with domain-specific intercepts and common slopes or nonlinear model formulations (e.g., Lehtonen et al., 2005). These extensions are discussed in Section 5.

In (11), unit-level auxiliary information about $\mathbf{x}_k$, also including known domain membership, for all population units is assumed. Actually, since the assisting model for (11) is linear, GREG estimation does not require unit-level information on $\mathbf{x}_k$. It is enough to have access to the vector $\mathbf{t}_{dx}$ of domain totals of auxiliary variables in the population and the corresponding HT estimates $\hat{\mathbf{t}}_{dx}$ in the sample. This can be seen by writing the GREG estimator in the standard textbook form,

$$\hat{t}_{d\mathrm{GREG}} = \hat{t}_{d\mathrm{HT}} + (\mathbf{t}_{dx} - \hat{\mathbf{t}}_{dx})'\hat{\mathbf{B}}.$$

An alternative calibration form incorporates g-weights:

$$\hat{t}_{d\mathrm{GREG}} = \sum_{k\in s} a_k g_{dk} y_k,$$

where $g_{dk} = I_{dk} + (\mathbf{t}_{dx} - \hat{\mathbf{t}}_{dx})'\hat{\mathbf{M}}^{-1}\mathbf{x}_k$, $I_{dk} = I\{k \in U_d\}$, and $\hat{\mathbf{M}} = \sum_{i\in s} a_i\mathbf{x}_i\mathbf{x}'_i$. The g-weights are often small outside domain sample s_d.

The variance of $\hat{t}_{d\mathrm{GREG}}$ is estimated by a double sum over the whole sample s:

$$\hat{V}(\hat{t}_{d\mathrm{GREG}}) = \sum_{k\in s}\sum_{l\in s}(a_k a_l - a_{kl}) g_{dk} e_k g_{dl} e_l \tag{12}$$

(Särndal et al., 1992, p. 401). Alternatively, the sum extends only over the domain sample s_d (Hidiroglou and Patak, 2004). For the direct estimator, these two forms are identical. These variance estimators do not take into account that the sample size n_{s_d} for an unplanned domain is random. To account for the randomness, we might apply GREG assisted by a model fitted to the extended domain variables $y_{dk} = I\{k \in U_d\} y_k$ (Estevao et al., 1995). It has also been proposed to fit the model to the original y_k and replace the residuals e_k in the variance estimator by "extended residuals" $e_{dk} = I\{k \in U_d\} y_k - \hat{y}_k$ (Lehtonen and Pahkinen, 2004, p. 202; Särndal, 2001, p. 39).

The basic direct and indirect GREG estimators and their variance estimators for the case of planned domains, discussed this far, are presented in Table 3 below. In both GREG estimators, access to domain-level auxiliary totals of x-variables is assumed. A key difference is in the model formulation: the direct GREG estimator employs domain-specific assisting models, whereas a model common for all domains is postulated for the indirect GREG estimator. Direct GREG estimation uses domain sample data in variance estimation; the data use extends to the whole sample in indirect GREG.

The GREG estimator (11) has been modified to take into account the domain size N_d when known:

$$\hat{t}_{d\mathrm{GREG}(N)} = \sum_{k\in U_d}\hat{y}_k + (N_d/\hat{N}_d)\sum_{k\in s_d} a_k(y_k - \hat{y}_k) = \sum_{k\in s} a_k g_{dk(N)} y_k, \tag{13}$$

where $g_{dk(N)} = (N_d/\hat{N}_d) I_{dk} + (\mathbf{t}_{dx} - (N_d/\hat{N}_d)\hat{\mathbf{t}}_{dx})' \hat{\mathbf{M}}^{-1} \mathbf{x}_k$ and $\hat{N}_d = \sum_{k\in s_d} a_k$. This estimator has smaller variance than the estimator (11) because the weighted mean of the residuals is more stable. The variance estimator of $\hat{t}_{d\mathrm{GREG}(N)}$ contains the weights $g_{dk(N)}$ instead of g_{dk}. If inference is conditional on observed sample domain sizes, $\hat{t}_{d\mathrm{GREG}(N)}$ is conditionally nearly unbiased, whereas the ordinary GREG is conditionally biased (Hidiroglou and Patak, 2004; Särndal and Hidiroglou, 1989). Therefore, $\hat{t}_{d\mathrm{GREG}(N)}$ yields better conditional confidence intervals. On the other hand, domain estimators (13) are not additive; their sum is not usually equal to the GREG estimator of the population total.

Table 3
The basic direct and indirect GREG estimators and their variance estimators for the planned domains case

	GREG Estimator Type	
	Direct	Indirect
Model formulation	$Y_k = \mathbf{x}'_k \boldsymbol{\beta}_d + \varepsilon_k$	$Y_k = \mathbf{x}'_k \boldsymbol{\beta} + \varepsilon_k$
GREG estimator	$\hat{t}_{d\mathrm{GREG}} = \hat{t}_{d\mathrm{HT}} + (\mathbf{t}_{dx} - \hat{\mathbf{t}}_{dx})' \hat{\mathbf{B}}_d$	$\hat{t}_{d\mathrm{GREG}} = \hat{t}_{d\mathrm{HT}} + (\mathbf{t}_{dx} - \hat{\mathbf{t}}_{dx})' \hat{\mathbf{B}}$
Variance estimator	$\hat{V}(\hat{t}_{d\mathrm{GREG}}) = \sum_{k\in s_d}\sum_{l\in s_d}(a_k a_l - a_{kl}) g_{dk} e_k g_{dl} e_l$	$\hat{V}(\hat{t}_{d\mathrm{GREG}}) = \sum_{k\in s}\sum_{l\in s}(a_k a_l - a_{kl}) g_{dk} e_k g_{dl} e_l$

4.1.2. Composite estimation for domains

As noted in Section 2.2.5, domains with small sample size can present problems in design-based estimation. This also holds for the GREG estimator (11). For example, the GREG estimate for a small domain is not necessarily bounded within an acceptable range. Even when only positive y-values are valid, the GREG estimate may be negative for a small domain when a negative residual is associated with a large weight a_k. In addition, although the GREG estimator (11) is nearly design unbiased, its design variance becomes large for a small domain. *Composite* or *combined* estimators have been proposed to overcome these kinds of problems.

Consider a composite estimator $\hat{t}_{d\mathrm{COMB}} = \lambda_d \hat{t}_{d\mathrm{GREG}} + (1 - \lambda_d)\hat{t}_{d\mathrm{SYN}}$, which is constructed as a weighted sum of the design-based GREG estimator (11) and a model-based synthetic estimator $\hat{t}_{d\mathrm{SYN}} = \sum_{k \in U_d} \hat{y}_k = \sum_{k \in U_d} \mathbf{x}'_k \hat{\mathbf{B}}$. The domain-specific weight $\lambda_d (0 \leq \lambda_d \leq 1)$ is chosen such that λ_d is close to one for large domains and approaches zero with decreasing domain sample size n_{s_d}. Thus, for small domains, the estimator $\hat{t}_{d\mathrm{COMB}}$ will be close to the synthetic estimator $\hat{t}_{d\mathrm{SYN}}$; the GREG estimator (11) will be obtained when the domain sample size is large. Different strategies in choosing λ_d are possible, leading to composite estimators of optimal type or sample size dependent type (see Rao, 2003a, Section 4.3).

The rationale behind composite estimation is obvious. The composite estimator can be written as $\hat{t}_{d\mathrm{COMB}} = \hat{t}_{d\mathrm{SYN}} + \lambda_d \sum_{k \in s_d} a_k (y_k - \hat{y}_k)$, reproducing the GREG estimator (11) with $\lambda_d = 1$. The design variance is of order $O(n^{-1})$ for the synthetic term $\hat{t}_{d\mathrm{SYN}}$ and of order $O\left(n_{s_d}^{-1}\right)$ for the bias correction term $\sum_{k \in s_d} a_k (y_k - \hat{y}_k)$. If the domain sample size n_{s_d} is large, the weight λ_d should be close to one and a sufficiently small variance will be obtained for $\hat{t}_{d\mathrm{COMB}}$. For a small domain, the variance of the correction factor of the GREG will be large and it is beneficial to decrease the value of λ_d because the variance of the component $\hat{t}_{d\mathrm{SYN}}$ tends to be small. This is an example of "trading bias against variance": by suitable choice of λ_d, a balance between the potential design bias of the synthetic estimator and the instability of the GREG estimator is achieved. The price to be paid for the variance reduction is increased design bias because the synthetic estimator $\hat{t}_{d\mathrm{SYN}}$ is generally design biased. The MSE of the composite estimator will be smaller than the MSE of the GREG estimator if the underlying model is not too bad for the given domain. However, with a poor-fitting model, the bias component of the MSE can dominate, leading to increased MSE.

An example of a *sample size dependent composite estimator* is provided by the GREG estimator (13), with $\lambda_d = N_d / \hat{N}_d$. We noted in Section 4.1.1 that the GREG estimate (13) is not necessarily bounded within an acceptable range. The likelihood of this occurrence is reduced when $N_d / \hat{N}_d$ is replaced by $\hat{N}_d / N_d$ in a domain where $n_{s_d} < \sum_{k \in U_d} \pi_k$ (Hidiroglou and Särndal, 1985). Further, Särndal and Hidiroglou (1989) proposed an estimator called *dampened regression estimator* given by

$$\hat{t}_{d\mathrm{DRE}} = \sum_{k \in U_d} \hat{y}_k + (\hat{N}_d / N_d)^{c-1} \sum_{k \in s_d} a_k (y_k - \hat{y}_k),$$

where $c = 0$ if $\hat{N}_d \geq N_d$ and $c = 2$ if $\hat{N}_d < N_d$.

Variants of composite estimators have often been used in practice. Examples of early references are Schaible et al. (1977), and Kumar and Lee (1983). A method called regression composite estimation is discussed in the context of repeated surveys, such

as a Labour Force Survey, in Singh et al. (1994), Bell (2001), Fuller and Rao (2001), Gambino et al. (2001), and Singh et al. (2001). Design-based composite estimation, including MSE estimation, is discussed more extensively in Rao (2003a). Model-based composite estimation is treated in Chapter 32.

4.1.3. Model groups approach

Instead of using a common model fitted to the whole sample, it is sometimes more convenient to consider a set of regression models defined for nonoverlapping subsets $U_p (p = 1, 2, \ldots, P)$ of the population called *model groups* (Estevao et al., 1995). In regional classification, there is often a hierarchy of regions, and model groups are larger regions composed of domains. More generally, the boundaries of the sets U_p do not have to agree with domain boundaries, and interleaving is allowed. In model group U_p, we define a model $Y_k = \mathbf{x}'_{kp}\boldsymbol{\beta}_p + \varepsilon_k;\ k \in U_p$. Here, the vectors $\mathbf{x}_{kp}$ may contain different variables in different groups U_p. Naturally, this ensemble of models is equivalent with a single regression model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \varepsilon$, where $\mathbf{X} = \text{diag}(\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_P)$ and $\boldsymbol{\beta} = (\boldsymbol{\beta}'_1, \boldsymbol{\beta}'_2, \ldots, \boldsymbol{\beta}'_P)'$. The general theory of the GREG estimator applies, but the $\mathbf{X}$ matrix is perhaps impractically large. It is easier to consider the separate models. For that purpose, the sample is divided into subsets $s_p = U_p \cap s$ and further into sets $s_{pd} = s_p \cap s_d$. If we know the auxiliary totals $\mathbf{t}_{dpx} = \sum_{k \in U_p \cap U_d} \mathbf{x}_{kp}$, estimated by $\hat{\mathbf{t}}_{dpx} = \sum_{k \in s_{pd}} a_k \mathbf{x}_{kp}$, the domain total GREG estimator (11) can be written as

$$\hat{t}_{d\text{GREG}} = \sum_p \sum_{k \in s_{pd}} \hat{y}_k + \sum_p \sum_{k \in s_{pd}} a_k (y_k - \hat{y}_k) = \hat{t}_{d\text{HT}} + \sum_p (\mathbf{t}_{dpx} - \hat{\mathbf{t}}_{dpx})' \hat{\mathbf{B}}_p,$$

where $\hat{\mathbf{B}}_p$ is obtained by fitting the regression model in model group U_p:

$$\hat{\mathbf{B}}_p = \hat{\mathbf{M}}_p^{-1} \sum_{i \in s_p} a_i \mathbf{x}_{ip} y_i \tag{14}$$

and $\hat{\mathbf{M}}_p = \sum_{i \in s_p} a_i \mathbf{x}_{ip} \mathbf{x}'_{ip}$.

The model groups approach can be generalized by the use of *overlapping* sets $U_{p(d)}$ that are defined for each domain U_d so that $U_d \subset U_{p(d)}$. In regional statistics, an example of $U_{p(d)}$ is the neighborhood of a region U_d, the union of U_d and all neighboring regions sharing a common border with the region. This makes sense if the neighboring regions are similar due to spatial correlations (e.g., D'Alo et al., 2006; Petrucci et al., 2005). Since there is no single regression model that is equivalent to the ensemble of separate regression models, the estimators are not necessarily additive.

When the models are defined separately for each domain ($U_p = U_d$), the resulting estimator is direct. In small domains, the direct estimator typically has large variance. Therefore, it has been common to use indirect estimator assisted by a model fitted in a larger subset of the sample. Design-based estimation with an indirect estimator is challenged by Estevao and Särndal (2004) but indirect estimation might be useful at least for the small domains. Hidiroglou and Patak (2004) note that an indirect estimator (13) incorporating $\hat{N}_d$ may be preferred to a corresponding direct estimator when the domain sample size is very small.

The auxiliary totals are not always known in every domain but only in the model groups U_p. This situation can be addressed by calibration. An alternative is the calibration-type GREG estimator discussed in Estevao et al. (1995). It is necessary to fit the regression models to the extended domain variables $y_{dk} = I\{k \in U_d\}y_k$:

$$\hat{t}_{d\mathrm{GREG}(G)} = \sum_{k\in s} a_k y_{dk} + \sum_p (\mathbf{t}_{px} - \hat{\mathbf{t}}_{px})'\hat{\mathbf{B}}_{p(d)}, \tag{15}$$

where the auxiliary total over U_p is denoted by $\mathbf{t}_{px}$, its HT estimate by $\hat{\mathbf{t}}_{px}$, and $\hat{\mathbf{B}}_{p(d)} = \hat{\mathbf{M}}_p^{-1}\sum_{i\in s_p} a_i\mathbf{x}_{ip}y_{di}$. Only model groups U_p that intersect the domain are included in (15). An alternative expression for unit-level auxiliary data is

$$\hat{t}_{d\mathrm{GREG}(G)} = \sum_{k\in U} \hat{y}_{dk} + \sum_{k\in s} a_k\big(y_{dk} - \hat{y}_{dk}\big),$$

where $\hat{y}_{dk} = \mathbf{x}'_{kp}\hat{\mathbf{B}}_{p(d)}$ for $k \in U_p$. The calibration equations hold at the model group level, that is, the total estimates of auxiliary variables agree with the known totals over U_p. This approach is adopted, for example, in GES and CLAN software packages.

The variance estimator for (15) is calculated using all residuals $e_{dk} = y_{dk} - \mathbf{x}'_{kp}\hat{\mathbf{B}}_{p(d)}$:

$$\hat{V}(\hat{t}_{d\mathrm{GREG}(G)}) = \sum_{k\in s}\sum_{l\in s} (a_k a_l - a_{kl}) g_{p(k)k} e_{dk} g_{p(l)l} e_{dl}, \tag{16}$$

with $g_{pk} = 1 + (\mathbf{t}_{px} - \hat{\mathbf{t}}_{px})'\hat{\mathbf{M}}_p^{-1}\mathbf{x}_{kp}$ and $k \in U_{p(k)}$ (Estevao et al., 1995; Hidiroglou and Patak, 2004). Obviously, the regression models fitted to y_{dk} will not fit the data well in a large model group and the residuals are often large. This inflates the variance; the problem is often met if a model group contains several domains.

4.1.4. A general class of domain estimators

Estevao and Särndal (2004) define a general class of estimators including both GREG estimators and calibration estimators based on an instrument vector: suppose the auxiliary totals are known over sets U_p, called calibration groups or model groups. For practical purposes, we again assume that the error variance σ_k^2 is constant. The regression parameter is estimated using subpopulations U_m and U_l:

$$\hat{\mathbf{B}}_{ml} = \left(\sum_{k\in s} a_k \mathbf{z}_k I_{mk} \mathbf{x}'_k\right)^{-1} \left(\sum_{k\in s} a_k \mathbf{z}_k I_{lk} y_k\right),$$

where $I_{mk} = I\{k \in U_m\}$ and $I_{lk} = I\{k \in U_l\}$, and $\mathbf{z}_k$ is an instrument vector, in GREG chosen as $\mathbf{z}_k = \mathbf{x}_k$. The domain estimator for $U_d \subset U_p$ is $\hat{t}_d = \hat{t}_{d\mathrm{HT}} + (\mathbf{t}_{px} - \hat{\mathbf{t}}_{px})'\hat{\mathbf{B}}_{ml}$, where the estimators $\hat{t}_{d\mathrm{HT}}$ and $\hat{\mathbf{t}}_{px}$ are HT estimators of the population totals of y_{dk} and $\mathbf{x}_{kp} = I\{k \in U_p\}\mathbf{x}_k$, respectively. As special cases, the calibration estimator based on the instrument vectors $\mathbf{z}_k$ has $U_m = U_p$ and $U_l = U_d$, as well as the GREG estimator incorporating model groups U_p. The ordinary GREG (11) has $U_m = U_l$. In GREG, the regression model is fitted to the whole sample (when $U_l = U$), to each domain (when $U_l = U_d$) or to calibration groups (when $U_l = U_p$). All these estimators are design consistent, and their relative bias tends to zero as $O(n^{-1/2})$.

Estevao and Särndal (2004) show that the design variance of the estimator is minimized by choosing $U_m = U_p$, $U_l = U_d$, and $\mathbf{z}_k = \sum_{l\in U}(a_k a_l/a_{kl} - 1) I_{pl}\mathbf{x}_l$. These instrument variables are estimated by $\mathbf{z}_k = a_k^{-1}\sum_{l\in s}(a_k a_l - a_{kl})\, I_{pl}\mathbf{x}_l$. The resulting estimator is then essentially identical with the so-called *optimal estimator* (Montanari, 1987; Montanari and Ranalli, 2002; Rao, 1994), which minimizes the design variance (Estevao and Särndal, 2004, p. 656)

$$\begin{aligned}\mathrm{Var}(\hat{t}_d) &= \mathrm{Var}(\hat{t}_{d\mathrm{HT}} + (\mathbf{t}_{px} - \hat{\mathbf{t}}_{px})'\mathbf{B}) \\ &= \mathrm{Var}(\hat{t}_{d\mathrm{HT}}) + \mathbf{B}'\mathrm{Var}(\hat{\mathbf{t}}_{px})\mathbf{B} - 2\mathbf{B}'\mathrm{Cov}(\hat{t}_{d\mathrm{HT}}, \hat{\mathbf{t}}_{px})\end{aligned}$$

with respect to $\mathbf{B}$. Unfortunately, the optimal estimator is often unstable, especially for designs more complex than SRS (Estevao and Särndal, 2004, p. 657). In practice, we should probably use $\mathbf{z}_k = I_{pk}\mathbf{x}_k$ instead. Then the estimator is the GREG estimator based on model groups. Note that the optimal estimator is a direct estimator using the y-values only from the given domain ($U_l = U_d$). The ordinary GREG estimator has approximately the same asymptotic variance as the optimal calibration estimator only if $U_p = U_d$. Andersson and Thorburn (2005) discuss optimality of a calibration estimator in relation to GREG estimation.

4.1.5. One-stage and two-stage designs

In addition to element-level sampling designs discussed so far, we can define GREG estimators for clusters (Estevao et al., 1995). In single-stage cluster sampling, a sample s_C of clusters is first drawn with design weights a_i^C and all elements in each sample cluster are surveyed. Clusters are grouped into model groups $C_p (p = 1, 2, \ldots, P)$. Consider a cluster $i \in C_p$ with elements s_i and auxiliary data $\mathbf{x}_i$. A regression model is defined for the sum y_{di}^C of y-variables $y_{dk} = I_{dk}y_k$ over the cluster:

$$y_{di}^C = \sum_{k\in s_i} y_{dk} = \mathbf{x}_i'\boldsymbol{\beta}_p + \varepsilon_i,$$

where the error variance is $\mathrm{Var}(\varepsilon_i) = \sigma_i^2$. The regression parameter is estimated for group C_p by

$$\hat{\mathbf{B}}_p = \hat{\mathbf{M}}_p^{-1} \sum_{i\in s_C\cap C_p} a_i^C \mathbf{x}_i y_{di}^C/\sigma_i^2,$$

where $y_{di}^C = \sum_{k\in s_i} y_{dk}$ and $\hat{\mathbf{M}}_p = \sum_{i\in s_C\cap C_p} a_i^C \mathbf{x}_i\mathbf{x}_i'/\sigma_i^2$.

The error variance $\mathrm{Var}(\varepsilon_i)$ can hardly be assumed constant, but, for example, it can often be assumed to be proportional to the size n_i of the cluster: $\sigma_i^2 = n_i\sigma^2$. Then the unknown σ^2 cancels out from $\hat{\mathbf{B}}_p$.

Using known auxiliary totals $\mathbf{t}_{px}^C = \sum_{i\in C_p}\mathbf{x}_i$ and their estimates $\hat{\mathbf{t}}_{px}^C = \sum_{i\in s_C\cap C_p} a_i^C\mathbf{x}_i$, we estimate t_d by

$$\hat{t}_{d\mathrm{GREG}(C)} = \sum_p \sum_{i\in s_c\cap C_p} a_i^C g_{pi}^C y_{di}^C,$$

where $g_{pi}^C = 1 + \left(\mathbf{t}_{px}^C - \hat{\mathbf{t}}_{px}^C\right)'\hat{\mathbf{M}}_p^{-1}\mathbf{x}_i/\sigma_i^2$.

The variance of $\hat{t}_{d\mathrm{GREG}(C)}$ is estimated using residuals $e_{di} = y_{di}^C - \mathbf{x}_i'\hat{\mathbf{B}}_p$ and the inclusion probabilities of clusters:

$$\hat{V}(\hat{t}_{d\mathrm{GREG}(C)}) = \sum_{i\in s_C}\sum_{j\in s_C}(a_i^C a_j^C - a_{ij}^C)g_{p(i)i}^C e_{di} g_{p(j)j}^C e_{dj}$$

with $i \in C_{p(i)}$ and $j \in C_{p(j)}$.

In two-stage sampling, the first-stage sample consists of primary sampling units (PSU), such as clusters. Then in each sample PSU, a sample of elements is drawn. The design weight of element k is a product $a_k = a_i^C a_{k|i}$ of the weight a_i^C of PSU i and the conditional design weight $a_{k|i}$ of element k within PSU i. This generalizes to more stages. If the model groups are defined at the PSU level, the regression models define how the PSU totals depend on auxiliary variables. However, the PSU totals are not known, and we use their HT estimates $\hat{t}_{di} = \sum_{k\in s_i} a_{k|i} y_{dk}$ instead. The GREG estimator of the domain total is

$$\hat{t}_{d\mathrm{GREG}(2)} = \sum_p \sum_{i\in s_c\cap C_p} a_i^C g_{pi}^C \hat{t}_{di}$$

but variance estimation requires more complex derivations (e.g., Estevao et al., 1995). Falorsi et al. (2000) discuss some simple estimation methods under two-stage sampling and Estevao and Särndal (2006) discuss calibration under two-stage and two-phase sampling.

4.2. Computational example with direct and indirect estimation under an unplanned domain structure

Domain totals are estimated here by direct Horvitz–Thompson and indirect GREG estimators. We use the same sample as in Section 3.5. This allows a comparison of results with the case of direct estimation for planned domains. There were $D = 12$ regions (domains) in our population. To demonstrate domain estimation for unplanned domains, we recognize that the regional sample sizes n_{s_d} are not fixed in the sampling design but are random (in Section 3.5, we assumed a case of planned domains with domain sample sizes fixed by stratification).

In addition to the income data for households, the sample data set includes the variables EDUC (number of household members who had higher education) and EMP (the number of months in total the household members were employed during last year). We again estimate the domain totals of disposable income of households in the 12 regions. We use the same auxiliary data as in Section 3.5. In addition to direct HT, we computed two indirect GREG estimates. Results are shown in Table 4. MARE is the mean absolute relative error and MCV is the mean coefficient of variation of the estimate over domain group.

The variance of ordinary HT (column 1 in Table 4) was estimated by $\hat{V}_U(\hat{t}_{d\mathrm{HT}})$ (5). As expected, in the present case of unplanned domains, the HT estimator had larger MCV than in the case of planned domains (column 1 in Table 2). The random domain sample size increased the variance of domain estimators.

In GREG, we first illustrate the model groups approach. We assumed that the population size N and the population total of EMP only were known. We thus had a single model

Table 4
Mean absolute relative error (MARE) and mean coefficient of variation (MCV) of HT and indirect GREG estimators of totals for minor, medium-sized, and major domains by using various amounts of auxiliary information in an unplanned domains case

	HT		GREG			
Auxiliary Information	1 None		2 Population Size and EMP Total		3 Domain Sizes and Domain Totals of EMP	
Domain sample size class	MARE (%)	MCV (%)	MARE (%)	MCV (%)	MARE (%)	MCV (%)
Minor $8 \le n_{s_d} \le 33$	11.5	28.3	11.5	28.3	7.6	9.0
Medium $34 \le n_{s_d} \le 45$	7.6	20.3	7.4	20.3	3.8	8.1
Major $46 \le n_{s_d} \le 277$	12.5	9.6	12.5	9.4	4.1	5.0

group, that is, the whole population. The indirect GREG estimator (15) was assisted by model

$$Y_{dk} = \beta_0 + \beta_1 \mathrm{EMP}_k + \varepsilon_k.$$

We thus did not use domain-level auxiliary information. For each domain, we fitted the model to the extended domain variables $y_{dk} = I\{k \in U_d\} y_k$. The variables y_{dk} were also included in the variance estimator (16). This GREG estimator (column 2) did not yield smaller errors or MCV than the HT estimator. The population level information was not powerful for domain estimation in this case, confirming the argument of favoring the use of lower level aggregates of auxiliary variables if available (Estevao and Särndal, 2004).

The second indirect GREG estimator (column 3) was assisted by a common model

$$Y_k = \beta_0 + \beta_1 \mathrm{EMP}_k + \varepsilon_k$$

fitted to the whole sample, and domain sizes and domain totals of EMP were assumed known. The variance was estimated using (12). This estimator outperformed the other three estimators. The MCV was larger than in the comparable direct GREG estimator for planned domains (column 3 in Table 2), as expected. The use of extended domain residuals $e_{dk} = y_{dk} - \hat{y}_k$ in the variance estimator would have affected the MCV only slightly. Increasing the number of auxiliary variables in GREG did not yield further improvement. The size correction with known domain size (13) resulted in small decrease in average errors, but MCV increased slightly.

We did have access to several cross-sectional yearly data sets of the survey and the corresponding auxiliary data. With two last year's data, the domains were defined by cross classification of year and region, yielding altogether 24 domains. We fitted models containing the year and interactions of year with EMP and EDUC, but the results did not

improve. A model fitting the whole sample better does not necessarily fit better for the data in domains of interest, and even if it did, a better fitting model does not guarantee better GREG estimates in one particular sample although improvement is expected on an average.

The problem of model choice is discussed in Lehtonen and Veijanen (1998), Estevao and Särndal (1999), Hedlin et al. (2001), Lehtonen et al. (2003, 2005), and Hidiroglou and Patak (2004). We address model choice in GREG estimation further in Sections 5.1 and 5.2.

4.3. Ratios and percentiles for domains

4.3.1. Ratios and means

Consider estimating the *ratio* $R_d = t_{dy}/t_{dz}$ of two unknown totals $t_{dy} = \sum_{k\in U_d} y_k$ and $t_{dz} = \sum_{k\in U_d} z_k$. An example is the unemployment rate, which is the ratio of the number of unemployed and the size of the labor force in the domain. Another example is the proportional area of fields allocated to, say, wheat in a region, estimated using data obtained from each farm k; we only need to know (y_k, z_k) for units in the sample from area d. A simple, nearly unbiased estimator of R_d is $\hat{R}_d = \hat{t}_{dy}/\hat{t}_{dz}$. We denote the ratio of two HT estimators by $\hat{R}_{d\mathrm{HT}}$ and the ratio of two GREG estimators by $\hat{R}_{d\mathrm{GREG}}$.

In a case of planned domains, the variance estimators for the ratios of direct HT and GREG estimators are defined as follows (Särndal et al., 1992, p. 178, 296):

$$\hat{V}(\hat{R}_{d\mathrm{HT}}) = \frac{1}{\hat{t}^2_{dz\mathrm{HT}}} \sum_{k\in s_d}\sum_{l\in s_d} (a_k a_l - a_{kl})(y_k - \hat{R}_{d\mathrm{HT}} z_k)(y_l - \hat{R}_{d\mathrm{HT}} z_l),$$

$$\hat{V}(\hat{R}_{d\mathrm{GREG}}) = \frac{1}{\hat{t}^2_{dz\mathrm{GREG}}} \sum_{k\in s_d}\sum_{l\in s_d} (a_k a_l - a_{kl}) g_{dk}(e_{yk} - \hat{R}_{d\mathrm{GREG}} e_{zk}) \times g_{dl}(e_{yl} - \hat{R}_{d\mathrm{GREG}} e_{zl}),$$

where the residuals $e_{yk} = y_k - \mathbf{x}'_k \hat{\mathbf{B}}_{dy}$ and $e_{zk} = z_k - \mathbf{x}'_k \hat{\mathbf{B}}_{dz}$ are obtained from regression models fitted in the domain to y_k and z_k, respectively, and the g-weights are common to both models. In the case of indirect GREG,

$$\hat{V}(\hat{R}_{d\mathrm{GREG}}) = \frac{1}{\hat{t}^2_{dz\mathrm{GREG}}} \sum_{k\in s}\sum_{l\in s} (a_k a_l - a_{kl}) g_{dk}(e_{yk} - \hat{R}_{d\mathrm{GREG}} e_{zk}) \times g_{dl}(e_{yl} - \hat{R}_{d\mathrm{GREG}} e_{zl}),$$

where $e_{yk} = y_k - \mathbf{x}'_k \hat{\mathbf{B}}_y$ and $e_{zk} = z_k - \mathbf{x}'_k \hat{\mathbf{B}}_z$ are residuals of models fitted in the whole sample.

With unplanned domains, we can estimate the domain ratio by the ratio of two population level estimators using extended domain variables $y_{dk} = I\{k \in U_d\} y_k$ and $z_{dk} = I\{k \in U_d\} z_k$. In the case of HT, this ratio is actually identical with $\hat{R}_{d\mathrm{HT}}$ defined above:

$$\hat{R}_{d(e)} = \frac{\sum_{k\in s} a_k y_{dk}}{\sum_{k\in s} a_k z_{dk}}.$$

Moreover, $\hat{V}(\hat{R}_{d(e)}) = \hat{V}(\hat{R}_{d\mathrm{HT}})$. In contrast, the variance estimator of a ratio of two GREG estimators incorporating the extended domain variables is

$$\hat{V}(\hat{R}_{d\mathrm{GREG}}) = \frac{1}{\hat{t}^2_{dz\mathrm{GREG}}} \sum_{k\in s}\sum_{l\in s}(a_k a_l - a_{kl}) g_{dk}(e_{ydk} - \hat{R}_{d\mathrm{GREG}} e_{zdk}) \times g_{dl}(e_{ydl} - \hat{R}_{d\mathrm{GREG}} e_{zdl}),$$

where $e_{ydk} = y_{dk} - \hat{y}_{dk}$ and $e_{zdk} = z_{dk} - \hat{z}_{dk}$ are from models fitted to the extended domain variables.

Domain mean $\bar{y}_d = t_d/N_d$ can be estimated by $\hat{\bar{y}}_d = \hat{t}_d/N_d$ when the domain size N_d is known. The variance estimator is correspondingly $\hat{V}(\hat{t}_d)/N_d^2$. An alternative is to interpret the domain mean as a ratio $R_d = t_d/t_{dz}$, where $t_{dz} = \sum_{k\in U_d} z_k = N_d$ is defined for $z_k = I_{dk} = I\{k \in U_d\}$. The estimator $\hat{t}_{dz}$ is an estimator of the domain size: $\hat{t}_{dz} = \hat{N}_d = \sum_{k\in s_d} a_k$. This is applicable also when N_d is unknown. The mean estimator is then $\hat{R}_d = \hat{t}_d/\hat{t}_{dz}$, and the variance is estimated by the formula for $\hat{V}(\hat{R}_d)$ with $z_k = I_{dk}$. Comparison of estimators of domain means is studied in Särndal et al. (1992), p. 412.

The *ratio estimator* is an estimator of t_d based on $\hat{R}_d$ and a known total t_{dz}: $\hat{t}_{dR} = t_{dz}\hat{R}_d$. It is nearly unbiased for t_d and its variance is estimated by $t^2_{dz}\hat{V}(\hat{R}_d)$. If the domain size N_d is known, a ratio estimator of t_d derived from an estimator of the domain mean is $\hat{t}_{d(N)} = N_d\hat{\bar{y}}_d$. If $\hat{\bar{y}}_d$ is estimated by $\hat{\bar{y}}_d = \hat{t}_{d\mathrm{HT}}/\hat{N}_d$, then $\hat{t}_{d(N)}$ is a special case of the Hájek type estimator. The estimates $\hat{t}_{d(N)}$ do not, in general, add up to the estimate of the population total.

4.3.2. Percentile estimation for domains

Percentiles, such as median and quartiles, are important in certain surveys, notably surveys of income statistics including median household income, income deciles, and derived poverty measures. The percentiles can be estimated using an estimated distribution function (Chambers and Dunstan, 1986; Chambers and Tzavidis, 2006; Rao et al., 1990; Tzavidis et al., 2007); recently, calibration has been used (Rueda et al., 2007a; Särndal, 2007; Wu and Sitter, 2001a). Harms and Duschene (2006) use known percentiles of auxiliary variables. These studies have not considered estimation of domain percentiles, but most population estimators can be apparently generalized for domain estimation. We also suggest straightforward application of the estimation equation approach of Binder and Patak (1994). Percentile estimation is discussed also in Chapter 36.

The distribution function is defined for a finite population domain U_d of size N_d as

$$F_d(t) = \sum_{k\in U_d} I(y_k \le t)/N_d,$$

where the indicator function $I(y_k \le t)$ equals 1 when $y_k \le t$ and 0 otherwise. The pth percentile is $\theta = \theta(p) = \inf\{t : F_d(t) \ge p\}$, that is, we find the smallest value θ for which proportion p of the y_k are smaller than or equal to θ. In the finite population, we choose percentiles among the values y_k. Then the percentile is $\theta = \min\{y_k : F_d(y_k) > p; k \in U_d\}$. It is useful to restate the problem as follows: the solution of $F_d(\theta) = p$ satisfies an estimating equation defined for $u(y,\theta) = I(y \le \theta) - p$:

$W_d(\theta) = \int_{-\infty}^{\infty} u(y, \theta) dF_d(y) = 0$. As F_d is a step function, the equation is

$$W_d(\theta) = \sum_{k \in U_d} u(y_k, \theta)/N_d = \sum_{k \in U_d} (I(y_k \leq \theta) - p)/N_d = 0.$$

When using a sample without auxiliary information, we estimate $W_d(\theta)$ by HT and use equation

$$\hat{W}_d(\theta) = \sum_{k \in s_d} a_k (I(y_k \leq \theta) - p)/N_d = 0.$$

This has the same form as the optimal estimating function of Godambe and Thompson (1986a), although their theory seems to require differentiable $u(y, \theta)$. The solution satisfies

$$\hat{F}_{d\mathrm{HT}}(\theta) = \sum_{k \in s_d} a_k I\,(y_k \leq \theta)/\hat{N}_d = p.$$

The function $\hat{F}_{d\mathrm{HT}}(\theta)$ is interpreted as an HT estimator of the distribution function. It is monotone, nondecreasing, and bounded in [0, 1]. This simplifies finding the percentile. The smallest value y_k for which $\hat{F}_{d\mathrm{HT}}(y) > p$ is found from sorted data in the same way as in the binary search algorithm. Percentile searching can be more complicated if the estimated distribution function is not monotone. Rao et al. (1990) have suggested that a monotone distribution function estimate is derived by tracking maxima. Rueda et al. (2007a) have presented a calibration-based monotone and nondecreasing estimator of the distribution function.

Särndal et al. (1992, p. 203) give an approximate variance estimator of $\hat{F}_{d\mathrm{HT}}(\theta)$:

$$\hat{V}_{\hat{F}}(\theta) = \hat{V}(\hat{F}_{d\mathrm{HT}}(\theta)) = \frac{1}{\hat{N}_d^2} \sum_{k \in s_d} \sum_{l \in s_d} (a_k a_l - a_{kl})(I(y_k \leq \hat{\theta}) - p)(I(y_l \leq \hat{\theta}) - p).$$

Under the assumption of normality of $\hat{F}_{d\mathrm{HT}}(\theta)$ close to p, a 95% confidence interval for $F_d(\theta)$ is $[p - 1.96\hat{V}_{\hat{F}}(\theta)^{1/2},\ p + 1.96\hat{V}_{\hat{F}}(\theta)^{1/2}]$. A confidence interval for θ is obtained from the equivalent equality $P\left\{\hat{F}_{d\mathrm{HT}}^{-1}(p - 1.96\hat{V}_{\hat{F}}(\theta)) \leq \theta \leq \hat{F}_{d\mathrm{HT}}^{-1}(p + 1.96\hat{V}_{\hat{F}}(\theta))\right\} = 0.95$.

When auxiliary data are available, Binder and Patak (1994) have proposed a generalization of the estimating equation containing $\alpha(\mathbf{x}, \boldsymbol{\beta}, \theta) = E\,(u(y, \theta)|\mathbf{x})$ under a model with parameter $\boldsymbol{\beta}$:

$$\int_{-\infty}^{\infty} \alpha\,(\mathbf{x}, \boldsymbol{\beta}, \theta)\, d[F_{X;d}(\mathbf{x}) - \hat{F}_{X;d}(\mathbf{x})] + \int_{-\infty}^{\infty} u(y, \theta) d\hat{F}_d(y) = 0,$$

where $F_{X;d}$ is the distribution function of $\mathbf{x}$ in domain d. In the case of percentile estimation, $\alpha(\mathbf{x}, \boldsymbol{\beta}, \theta) = P\{Y \leq \theta|\mathbf{x}; \boldsymbol{\beta}\} - p$. Let us denote the probability $P\{Y \leq \theta|\mathbf{x} = \mathbf{x}_k; \boldsymbol{\beta}\}$ by p_k. The estimating equation is

$$\sum_{k \in U_d} \frac{1}{N_d}\,(p_k - p) - \sum_{k \in s_d} \frac{1}{\hat{N}_d} a_k\,(p_k - p) + \sum_{k \in s_d} \frac{1}{\hat{N}_d} a_k\,(I(y_k \leq \theta) - p) = 0.$$

When we substitute estimated probabilities $\hat{p}_k$ for p_k (see below), we obtain an equation

$$\hat{F}_{d\mathrm{GREG}}(\theta) = p \quad \text{with} \quad \hat{F}_{d\mathrm{GREG}}(\theta) = \frac{1}{N_d}\sum_{k\in U_d}\hat{p}_k + \frac{1}{\hat{N}_d}\sum_{k\in s_d} a_k(I(y_k \le \theta) - \hat{p}_k).$$

This is interpreted as a GREG estimator (13) of the distribution function; the indicators $I(y_k \le \theta)$ are the observations and $\hat{p}_k$ are the fitted values. This estimator is similar to a difference estimator defined in Rao et al. (1990). It is indirect if the probabilities $\hat{p}_k$ are estimated from the whole sample, and then the percentile should be searched using all observations of the sample, but variance is still probably large in a small domain. We can estimate the variance of $\hat{F}_{d\mathrm{GREG}}(\theta)$ using the ordinary variance estimator $\hat{V}$ of GREG (13). This would yield a confidence interval with end points $\hat{F}^{-1}_{d\mathrm{GREG}}(p - 1.96\hat{V}^{1/2})$ and $\hat{F}^{-1}_{d\mathrm{GREG}}(p + 1.96\hat{V}^{1/2})$, but its properties are not known yet.

The estimates $\hat{p}_k$ are obtained from a logistic regression model fitted to the indicators $I(y_k \le \theta)$ in the sample, preferably by maximizing a pseudolikelihood that contains design weights. Alternatively, one can obtain $\hat{p}_k$ using the empirical distribution function $\hat{F}_{\hat{e}}$ of the standardized residuals $\hat{e}_k = (y_k - \mathbf{x}'_k\hat{\boldsymbol{\beta}})/\hat{\sigma}$ in the sample; $\hat{p}_k = \hat{F}_{\hat{e}}((\theta - \mathbf{x}'_k\hat{\boldsymbol{\beta}})/\hat{\sigma})$, or by using the fitted values $\hat{y}_k$ in the population; $\hat{p}_k = I\big(\hat{y}_k \le \theta\big)$ (Rao et al., 1990; Wu and Sitter, 2001a). In domain estimation, it is an open question whether to use only the data in the domain or a larger data set to obtain possibly better estimates $\hat{p}_k$.

5. Extended GREG family for domain estimation

5.1. Assisting models

A fixed-effects linear model is often chosen as an assisting model for a GREG estimator of direct or indirect type; this was the case in Sections 3 and 4. When the model does not fit well in a domain, the population fit residuals $E_k = y_k - \hat{y}_k$ in that domain can be large, inflating the estimator's variance. Nonlinear models may fit better, especially if the variable of interest is binary or multinomial. Mixed models can offer an interesting alternative for direct and indirect GREG estimators with fixed-effects type assisting models. By introducing suitable random components in the model, flexible accounting for the domain differences is allowed. The extended GREG family of domain estimators refers to GREG type estimators where the assisting model is a member of the family of generalized linear mixed models (GLMM; e.g., Breslow and Clayton, 1993; McCulloch and Searle, 2001). Lehtonen and Veijanen (1998) and Lehtonen et al. (2003, 2005) have introduced GREG estimators of the form (11) assisted by logistic, multinomial logistic, and mixed models. This approach might be attractive at least from a modeller's point of view. Torabi and Rao (2008) compare the MSE behavior of an EBLUP estimator with a GREG estimator assisted by a mixed model, introduced in Lehtonen and Veijanen (1999).

Access to reliable auxiliary information is essential for accurate domain estimation. In Sections 3 and 4, we worked with aggregate-level auxiliary data. Now, we assume access at unit-level auxiliary data. Let us assume that the auxiliary vector value $\mathbf{x}_k = (1, x_{1k}, \ldots, x_{jk}, \ldots, x_{Jk})'$ and domain membership is known and specified in the frame for every unit $k \in U$. Consider first a generalized linear fixed-effects model,

$E_m(Y_k) = f(\mathbf{x}_k; \boldsymbol{\beta})$ for a given function $f(\cdot\,; \boldsymbol{\beta})$, where $\boldsymbol{\beta}$ requires estimation, and E_m refers to the expectation under the model. Examples of $f(\cdot\,; \boldsymbol{\beta})$ are a linear functional form or a logistic function. The model fit to the sample data $\{(y_k, \mathbf{x}_k); k \in s\}$ yields the estimate $\hat{\mathbf{B}}$ of $\mathbf{B}$, a finite population counterpart of $\boldsymbol{\beta}$. Using the estimated parameter values, the vector value $\mathbf{x}_k$, and the domain membership of k, we compute the predicted value $\hat{y}_k = f(\mathbf{x}_k; \hat{\mathbf{B}})$ for every $k \in U$, which is possible under our assumptions.

A similar reasoning applies for a generalized linear mixed model involving random effects in addition to the fixed effects. The model specification is $E_m(Y_k|\mathbf{u}_d) = f(\mathbf{x}'_k(\boldsymbol{\beta} + \mathbf{u}_d))$, where $\mathbf{u}_d$ is a vector of random effects defined at the domain level. Using the estimated parameters, predicted values $\hat{y}_k = f(\mathbf{x}'_k(\hat{\mathbf{B}} + \hat{\mathbf{u}}_d))$ are computed for all $k \in U$.

An example of a mixed model formulation is a multinomial logistic mixed model for a binary or polytomous y-variable. In addition to domains U_d, a second subdivision of U arises: for an m-class polytomous variable, the population is also subdivided into classes denoted U_i, $i = 1, \ldots, m$. For class U_i, denote the response variable as y_i with value $y_{ik} = 1$ if $k \in U_i$ and $y_{ik} = 0$ otherwise. We want to estimate the class frequencies or totals $t_{id} = \sum_{k \in U_d} y_{ik}$, $i = 1, \ldots, m$, for all domains U_d. For a binary y-variable ($m = 2$), the domain totals are $t_d = \sum_{k \in U_d} y_k$. The multinomial logistic mixed model is of the form

$$E_m(y_{ik}|\mathbf{u}_d) = P\{y_{ik} = 1|\mathbf{u}_d\} = \frac{\exp(\mathbf{x}'_k(\boldsymbol{\beta}_i + \mathbf{u}_{id}))}{1 + \sum_{r=2}^{m} \exp(\mathbf{x}'_k(\boldsymbol{\beta}_r + \mathbf{u}_{rd}))}$$

for $k \in U_d$, $i = 1, \ldots, m$, $d = 1, \ldots, D$, where $\mathbf{x}_k$ is a known vector value for every $k \in U$, $\boldsymbol{\beta}_i$ is a vector of fixed effects common for all domains, $\mathbf{u}_d = (\mathbf{u}'_{1d}, \ldots, \mathbf{u}'_{id}, \ldots, \mathbf{u}'_{md})'$, and $\mathbf{u}_{id}$ is a vector of domain-specific random effects, defined for the classes of the y-variable. To avoid identifiability problems, we set $\boldsymbol{\beta}_1 = 0$. Lehtonen et al. (2005) give special cases of the model.

Obviously, the possible nonlinearity of the model complicates the method. For example, we cannot express the sum of fitted values using the sum of auxiliary variables; in general, $\sum_{k \in U_d} \hat{y}_k \neq \left(\sum_{k \in U_d} \mathbf{x}_k\right)' \hat{\mathbf{B}}$. As a consequence, the GREG estimator cannot be written using the totals of auxiliary variables. The representation incorporating g-weights is also invalid, and the variance estimator with g-weights is not appropriate. For a given model specification, the GREG estimator of domain total $t_d = \sum_{k \in U_d} y_k$ remains the one given by (11), that is, the form $\hat{t}_{d\mathrm{GREG}} = \sum_{k \in U_d} \hat{y}_k + \sum_{k \in s_d} a_k(y_k - \hat{y}_k)$, $d = 1, \ldots, D$. The latter component in GREG, an HT estimator of the residual total, aims at correcting for the bias of the synthetic part.

We could use a simpler variance estimator (9), but it is probably negatively biased. A resampling-based variance estimator might be preferred. Stukel et al. (1996) discuss jackknife type variance estimation for calibration estimators.

For simplicity, we concentrate now on linear models (Lehtonen et al., 2003). The model specification of a linear mixed model is $E_m(Y_k|\mathbf{u}_d) = \mathbf{x}'_k(\boldsymbol{\beta} + \mathbf{u}_d) = (\beta_0 + u_{0d}) + (\beta_1 + u_{1d})\, x_{1k} + \cdots + (\beta_J + u_{Jd})\, x_{Jk}$, where $\mathbf{u}_d = (u_{0d}, u_{1d}, \ldots, u_{Jd})'$ is a vector of random effects defined at the domain level. The random effects are assumed to have common distribution. In estimation, they are often shrunken towards zero. The random components of $\mathbf{u}_d$ represent deviations from the corresponding coefficients of the fixed-effects part of the model. In practice, not all components are treated as random; for some j, $u_{jd} = 0$ for every d. A simple example is a model that includes domain-specific random intercepts u_{0d} as the only random term. If all components of $\mathbf{u}_d$ are set

to zero, a fixed-effects model is attained. The mixed model is usually estimated by using ML and restricted or residual maximum likelihood (REML) methods (e.g., Goldstein, 2002; McCulloch and Searle, 2001). Using the estimated parameters, predicted values $\hat{y}_k = \mathbf{x}'_k(\hat{\mathbf{B}} + \hat{\mathbf{u}}_d)$ are computed for all $k \in U$. The predictions $\{\hat{y}_k; k \in U\}$ differ from one model specification to another.

An additional possible direction for extension of the GREG concept is explored in Breidt and Opsomer (2000). These authors use nonparametric regression techniques to obtain the fitted values necessary for a GREG type estimator. Zheng and Little (2003, 2004) use penalized spline nonparametric mixed models for a similar purpose. Nonparametric and semiparametric estimation is discussed in Chapter 27. By using suitable mixed models, Jiang and Lahiri (2006) introduce a model-assisted empirical best prediction approach for domain means.

5.2. *Computational example for extended GREG family estimators*

We compare empirically the design bias and accuracy of model-assisted GREG type estimators of domain totals of a continuous y-variable for different linear assisting models (fixed-effects, mixed). Results are based on Monte Carlo simulation experiments, where repeated systematic probability proportional-to-size samples (πPS design) were drawn from an artificially generated fixed and finite population. The inclusion probabilities were $\pi_k = nx_{1k}\big/\sum_{k\in U} x_{1k}$. The weights $a_k = 1/\pi_k$ varied between 54.5 and 599.8. We used unit-level auxiliary data.

In the Monte Carlo experiment, for an estimate $\hat{t}_d(s_v)$ obtained for sample s_v; $v = 1, 2, \ldots, K$, we computed for each domain U_d the absolute relative bias (ARB; defined as the ratio of the absolute value of bias to the true value), given by $\big|(1/K)\sum_{v=1}^{K} \hat{t}_d(s_v) - t_d\big|/t_d$, and relative root mean squared error (RRMSE), defined as the ratio of the root MSE to the true value, given by $\sqrt{(1/K)\sum_{v=1}^{K} (\hat{t}_d(s_v) - t_d)^2}\big/t_d$.

There were $D = 100$ domains in the population. The size of domain U_d was proportional to $\exp(q_d)$, where q_d was simulated from U(0,2.9). We had 47 domains with minor sample sizes, 19 domains with medium sample sizes, and 34 domains with major sample sizes. These three size classes were defined on the basis of expected sample size $n(t_{dx_1}/t_{x_1})$ in domain U_d, where x_1 is the size variable used in πPS sampling. The domain size classes were less than 70, 70–119, and 120 or more units. The smallest domain of the generated population had 1721 units and the largest had 28,614.

The auxiliary variable x_1 was simulated from uniform distribution U(1,11). Another auxiliary variable x_2, unrelated to the sampling design, was simulated from $U(-5, 5)$. The random effects u_d and random slopes v_{id}, $i = 1, 2$, were simulated for each domain from multinormal distribution with variances $\mathrm{Var}(u_d) = 1$, $\mathrm{Var}(v_{id}) = 0.125$ and correlations $\mathrm{Corr}(u_d, v_{id}) = -0.5$; $\mathrm{Corr}(v_{1d}, v_{2d}) = 0$. The error term ε was generated from N(0,100). Values of the y-variable were simulated as $y_k = 1 + (1 + v_{1d})x_{1k} + (1 + v_{2d})x_{2k} + u_d + \varepsilon_k$. Correlations of the variables in the population were as follows: $\mathrm{corr}(y, x_1) = 0.44$, $\mathrm{corr}(y, x_2) = 0.45$, and $\mathrm{corr}(x_1, x_2) \approx 0$. Domain means of the y-variable were approximately equal but the totals differed considerably: The means of domain totals were 50,977 for minor domains, 131,776 for medium domains, and 263,979 for major domains.

Our population size was $N = 1{,}000{,}000$ and sample size $n = 10{,}000$. $K = 1000$ independent samples were selected. The following assisting models (groups A, B, C, and D) were considered:

Model A1, $Y_k = \beta_{0d} + \varepsilon_k$, $k \in U_d$, producing a direct estimator GREG-A1.

Model A2, $Y_k = \beta_0 + u_d + \varepsilon_k$, $k \in U$, producing an indirect estimator MGREG-A2.

Model B1, $Y_k = \beta_{0d} + \beta_2 x_{2k} + \varepsilon_k$, $k \in U$, producing an indirect estimator GREG-B1.

Model B2, $Y_k = \beta_0 + u_d + \beta_2 x_{2k} + \varepsilon_k$, $k \in U$, producing an indirect estimator MGREG-B2.

Model C1, $Y_k = \beta_{0d} + \beta_1 x_{1k} + \varepsilon_k$, $k \in U$, producing an indirect estimator GREG-C1.

Model C2, $Y_k = \beta_0 + u_d + \beta_1 x_{1k} + \varepsilon_k$, $k \in U$, producing an indirect estimator MGREG-C2.

Model D1, $Y_k = \beta_{0d} + \beta_1 x_{1k} + \beta_2 x_{2k} + \varepsilon_k$, $k \in U$, producing an indirect estimator GREG-D1.

Model D2, $Y_k = \beta_0 + u_d + \beta_1 x_{1k} + \beta_2 x_{2k} + \varepsilon_k$, $k \in U$, producing an indirect estimator MGREG-D2.

A-models did not contain auxiliary information. In B-models, the auxiliary variable x_2 was used, whereas the πPS size variable x_1 was included in C-models. Both auxiliary variables were included in D-models. Note that for the models A1, A2, B1, and B2, the sampling is *informative* (see Chapter 39), because the values of the y-variable depend on x_1 but the predictor is not included in the model. In models A1, B1, C1, and D1, the domain differences were accounted for by domain-specific fixed effects β_{0d}, and in A2, B2, C2, and D2 by domain-specific random intercepts $\beta_0 + u_d$. We incorporated the design weights a_k in the estimation procedures of model parameters, including the mixed models. This facilitates the condition of "internal bias calibration" (a proper combination of model formulation and estimation procedure under a given sampling design) proposed, for example, by Firth and Bennett (1998). The design weights were included in a REML method introduced in Saei and Chambers (2004) by modifying matrix products of $\mathbf{X}$, $\mathbf{y}$, the $\mathbf{Z}$ matrix whose columns are domain indicators, and $\mathbf{e}$, the vector of residuals: for example, the sample-based $\mathbf{X}'_s\mathbf{X}_s$ in the original algorithm was replaced by $\mathbf{X}'_s\mathbf{W}\mathbf{X}_s$, where $\mathbf{W}$ is the diagonal matrix of design weights. $\mathbf{X}'_s\mathbf{W}\mathbf{X}_s$ is an estimate of the corresponding product $\mathbf{X}'_U\mathbf{X}_U$ defined in the population.

The design bias of GREG estimators remained negligible for all model formulations considered (Table 5). In model groups A, B, C, and D, a mixed model formulation yielded slightly better results than fixed model formulation. Accuracy improved when incorporating in B-type assisting models the auxiliary variable x_2 (which was unrelated to the sampling design). GREG-C1 and GREG-C2 outperformed the A-type and B-type estimators. Best accuracy was obtained for the D-models. Thus, the inclusion of the πPS size variable x_1 in C-type and D-type assisting models appears powerful in this case. This strategy facilitates "double use" (Särndal, 1996) of the auxiliary information (i.e., to use it both in the sampling design and in the estimation phase).

Table 5
Average absolute relative bias (ARB) and average relative root mean squared error (RRMSE) of GREG estimators of domain totals for minor, medium-sized, and major domains of the generated population

Model and Estimator	Average ARB (%)			Average RRMSE (%)		
	Domain Size Class			Domain Size Class		
	Minor (20 − 69)	Medium (70 − 119)	Major (120+)	Minor (20 − 69)	Medium (70 − 119)	Major (120+)
Model A1 $Y_k = \beta_{0d} + \varepsilon_k$						
GREG-A1	1.2	0.7	0.3	20.2	11.9	8.5
Model A2 $Y_k = \beta_0 + u_d + \varepsilon_k$						
MGREG-A2	0.5	0.5	0.3	19.9	11.8	8.5
Model B1 $Y_k = \beta_{0d} + \beta_2 x_{2k} + \varepsilon_k$						
GREG-B1	1.2	0.6	0.3	18.3	10.7	7.7
Model B2 $Y_k = \beta_0 + u_d + \beta_2 x_{2k} + \varepsilon_k$						
MGREG-B2	0.5	0.4	0.2	18.0	10.6	7.7
Model C1 $Y_k = \beta_{0d} + \beta_1 x_{1k} + \varepsilon_k$						
GREG-C1	0.4	0.3	0.2	17.5	10.3	7.5
Model C2 $Y_k = \beta_0 + u_d + \beta_1 x_{1k} + \varepsilon_k$						
MGREG-C2	0.3	0.3	0.2	17.3	10.2	7.5
Model D1 $Y_k = \beta_{0d} + \beta_1 x_{1k} + \beta_2 x_{2k} + \varepsilon_k$						
GREG-D1	0.4	0.3	0.2	15.3	8.8	6.5
Model D2 $Y_k = \beta_0 + u_d + \beta_1 x_{1k} + \beta_2 x_{2k} + \varepsilon_k$						
MGREG-D2	0.3	0.3	0.2	15.1	8.7	6.5

5.3. Other extensions

A class of extended generalized regression estimators (EGRE) has been introduced by Montanari and Ranalli (2002) but it has not been applied in domain estimation yet. Calibration has been generalized in various ways. Wu and Sitter (2001a) discuss model calibration approach by defining the calibration equations for the fitted values: $\sum_{k\in s} w_{ks}\hat{y}_k = \sum_{k\in U} \hat{y}_k$. This approach works well with nonlinear models but auxiliary information is needed at unit level. Nonparametric model calibration by neural networks is studied in Montanari and Ranalli (2005), who assumed access to unit-level auxiliary information. Lehtonen et al. (2008) compared model calibration and GREG in the context of domain estimation.

6. Software

6.1. SAS applications and macros

SAS procedure SURVEYMEANS can be used in HT estimation for domains (STRATA and DOMAIN statements) under unequal probability sampling. SAS procedure SURVEYFREQ is available for domain analysis of frequency tables. With some additional programming, SAS procedure SURVEYREG yields GREG estimates for domains. Extended domain variables y_d with $y_{dk} = I_{dk}y_k = I\{k \in U_d\}y_k$ can be used for unplanned domain structures. Variance estimation is based on Taylor linearization.

CALMAR (CALibration on MARgins) and CALMAR 2 are calibration-oriented SAS macro programs of INSEE (Caron and Sautory, 2004; Le Guennec and Sautory, 2003). Methods of Deville and Särndal (1992) and Deville et al. (1993), for example, are implemented.

CLAN is a freely available SAS macro developed at Statistics Sweden (Andersson and Nordberg, 1998). CLAN contains GREG and different calibration methods. Variance estimation is based on Taylor linearization.

GES, Statistics Canada's Generalized Estimation System is a domain estimation package including GREG and calibration estimation (Estevao et al., 1995). The same g-weights can be applied to different y-variables and different domains. Variance is estimated by Taylor linearization or jackknife.

Computer software for sample surveys is discussed further in Chapter 13.

6.2. Application Domest

Domest is an interactive Java application developed for the estimation of totals or means for domains and small areas. It uses methods described in Lehtonen et al. (2003) and Saei and Chambers (2004). Domest provides both model-based and design-based domain estimators. Mixed models are incorporated into EBLUP, synthetic estimator, and pseudo EBLUP (Rao, 2003a). Design-based methods include HT and most GREG methods presented in this chapter. GREG estimation is assisted by fixed-effects regression models or mixed models, fitted with or without design weights. Currently, GREG variance estimation allows SRSWOR, Poisson sampling, and πPS with approximated second-order inclusion probabilities (Berger, 2004, 2005b; Hájek, 1964).

A linear regression model is fitted by OLS or WLS, and a mixed model is fitted by ML or REML (Saei and Chambers, 2004). When the fitting of a mixed model incorporates design weights in the same way as in pseudolikelihood estimation, the design bias of EBLUP seems to decrease.

The mixed model can include both area and time effects. The area effects are then assumed independent and time effects have AR(1) correlations. In a mixed model with spatially correlated random effects, the correlation of the random effects associated with regions a and b distance d_{ab} apart is $\mathrm{Corr}(u_a, u_b) \propto \exp(-d_{ab})$. Spatial correlations may improve the predictive power of a synthetic domain estimator. In a domain missing from the sample, the correlation structure yields a nonzero estimate of the associated random effect.

SAS data or text files can be imported into Domest and output tables are saved as text files or added incrementally to an HTML file.

Domest is developed at Statistics Finland by Ari Veijanen with Risto Lehtonen. It is freely available from the authors.

Acknowledgments

Discussions with Carl-Erik Särndal over past years have inspired us in writing the chapter. He also contributed to Section 5.2, as did Mikko Myrskylä of University of Pennsylvania. Comments of two referees and the Editor were of great help in finalizing the materials. We are thankful to Kari Djerf of Statistics Finland for providing us with the real-life survey data for our empirical examples in Sections 3 and 4. Thanks are also due to Statistics Finland and University of Helsinki for kindly supporting this work.

Sample Surveys: Inference and Analysis, Vol. 29B
ISSN: 0169-7161
© 2009 Elsevier B.V. All rights reserved
DOI: 10.1016/S0169-7161(09)00232-6

Model-Based Approach to Small Area Estimation

Gauri S. Datta

1. Introduction

Sample surveys have long been used as cost-effective means for data collection. Such data have been effectively used to provide suitable statistics not only for the population targeted by the survey but also for a variety of subpopulations, often called domains or areas. Domains may be geographical areas such as states, or socio-demographic groups (for example, white male) or other subpopulations. A domain or an area is considered a large or a major domain if the domain sample is sufficiently large so that the domain sample can provide a "direct" estimate of the domain parameter (for example, the mean) with adequate precision. On the other hand, a domain or an area is regarded as "small" if the domain-specific sample is not large enough to produce a direct estimate with reliable precision. In the survey sampling literature, areas or domains with small sample are referred to as "small areas." Small areas are also often referred to as "small domains," "local areas," "subdomains," "substates," etc. (cf. Rao, 2003a). Following the title of the book by Rao (2003a), in this chapter we will stick to the popular usage "small area."

Research on small area estimation has experienced a rapid growth in the last 25 years. Small area estimation methods are enjoying increasing popularity in survey sampling because of the growing demand for reliable small area estimates both from public and private sectors. In the United States, Canada, and other countries there is an "increasing government concern with the issues of distribution, equity, and disparity" (Brackstone, 1987). For example, there may exist underprivileged geographical subgroups within a given population that are far below the average in many respects, and need an uplift. To implement remedial program, it is necessary to identify such regions, and accordingly, one must have suitable statistical data for these regions. Government agencies, both at the national level and local level, use small area statistics for distribution of government funds and planning for welfare and service. In the private sector, businesses make decisions based on local income, population, and environmental data to evaluate markets for new products and to determine areas for location, expansion, and contraction of their activities.

To make better policy decision and to address emerging or existing social issues, many governments have passed laws requiring regular production of reliable and up-to-date small area estimates. For example, the U.S. Congress has passed a law requiring the Secretary of Commerce to produce and publish at least every 2 years, beginning in 1996, current data related to poverty. Specifically, the law requires that "to the extent feasible," the Secretary shall produce estimates of poverty for states, counties, and local jurisdictions of government and school districts. For school districts, estimates are to be made of the number of poor children in the 5–17 years age interval. It also specifies production of state and county estimates of the number of poor persons aged 65 and over. These estimates will be used by a broad range of customers including policy makers at the state and local levels as well as the private sector. This includes allocation of federal and state funds, federal funds annually being nearly $100 billion in the recent years.

There are many more important applications of small area estimation encountered by various government agencies. For many examples and case studies in small area estimation one may refer to Ghosh and Rao (1994), Rao (2003a), and Longford (2005). Schaible (1996) has an excellent account describing the use of indirect estimation, in particular, small area estimation in many U.S. Federal programs. In our treatment of model-based small area estimation we have mostly ignored the sampling design. Sverchkov and Pfeffermann (2008) in Chapter 39 present a treatment on small area estimation under informative sampling of areas and within the areas. For a design-based treatment of domain and small area estimation, we refer to Chapter 31 by Lehtonen and Veijanen (2008).

Indirect estimates of small area means that borrow strength from other areas are referred to as cross-sectional estimates. On the other hand for a survey which is repeated regularly, one can obtain indirect estimates of small area means by borrowing strength both from other areas and the time series. The latter estimates are referred to as cross-sectional time series estimates. The remainder of the chapter is structured as follows. Section 2 deals with a systematic development of frequentist model-based small area estimation. It begins with the development of synthetic and composite estimators in Section 2.1. This is followed by small area estimators based on mixed linear models in Section 2.2. Such estimators are viewed as empirical best linear unbiased predictors (EBLUPS). Second-order accurate mean squared error (MSE) approximation of the EBLUPs and second-order unbiased estimators of the accurate MSE approximations are also given. Small area estimators based on multivariate data are also derived in this section. Section 2.3 introduces cross-sectional time series small area estimators based on time series data, while Section 2.4 discusses empirical Bayes (EB) small area estimators for generalized linear mixed models (GLMM). Section 3 discusses hierarchical Bayesian (HB) methods for small area estimation. It begins in Section 3.1 with Bayesian normal theory small area estimation methods for unit-level models. Multivariate HB methods suitable for area-level and unit-level data are discussed in Section 3.2. Section 3.3 discusses both EB and HB methods for area-level data that exhibit correlated sampling errors. The cross-sectional time series approach to small area estimation is discussed in Section 3.4. HB small area estimation in GLMM is discussed in Section 3.5. Some final remarks including some open problems are made in Section 4.

We conclude this section by briefly describing and comparing an EB estimator and an HB estimator. In the Bayesian approach one needs to specify a prior distribution on the parameter appearing in the sampling distribution, namely the distribution of the

data conditional on the parameter. If this prior distribution is completely specified then one can use the conditional distribution of the parameter given the data, which is also known as the posterior distribution of the parameter. One can obtain the Bayes estimator from this posterior distribution. Often, however, the prior distribution itself involves some unknown parameter known as hyperparameter. In such situation, one cannot use the Bayes estimator because it involves the unknown hyperparameter. One can either estimate the hyperparameter from the marginal (by integrating out the parameter) distribution of the data or can assign another prior distribution (which is completely known), known as hyperprior, on the hyperparameter. In the first case, one simply replaces the hyperparameter by its estimator (obtained from the marginal distribution) in the Bayes estimator to get the EB estimator. In the second case, one integrates out the unknown hyperparameter in the Bayes estimator with respect to posterior distribution of the hyperparameter. The resulting estimator is known as the HB estimator. An EB estimator is essentially treated as a frequentist estimator. Indeed for a normal mixed linear model, the EBLUP and EB predictors of small area means are identical. However, the EB has wider applicability than the EBLUP approach because the former can be applied also to nonlinear or generalized linear models. We present EB procedures under the frequentist approach.

2. Model-based frequentist small area estimation

2.1. Synthetic and composite estimation

Much of the popularity and usefulness of small area estimation methods can be attributed to the model-based approach developed in the past three decades. A direct estimate of a small area mean is based on the sample from that area alone. It is often unreliable due to small sample from that area. Model-based methods in small area estimation are now in extensive use to compute indirect estimates of small area means. These methods facilitate "borrowing strength" from neighboring areas using suitable linking models that explicitly connect the direct estimators from the small areas.

Prior to developments of model-based small area estimates, synthetic estimates were used in many government agencies (see Gonzalez, 1973). According to Gonzalez (1973), "an unbiased estimate is obtained from a sample survey for a large area; when this estimate is used to derive estimates for subareas under the assumption that the small areas have the same characteristics as the large area, we identify these estimates as synthetic estimates." The NCHS (1968) in the U.S. proposed synthetic estimates to calculate at the state level the estimates of long- and short-term physical disabilities from the National Health Interview Survey data. According to Ghosh and Rao (1994), the popularity of synthetic estimates results from their simplicity, wide applicability to general sampling designs, and potential of producing more accurate estimates by implicitly borrowing strength from similar small areas.

Synthetic estimates are quite popular among practitioners because they do not use explicit models. However, such estimates can be justified through models. At unit-level let y_{ij} denote the value for the jth unit in the ith small area, with $j = 1, \ldots, N_i$, $i = 1, \ldots, m$, where N_i is the size of the finite population corresponding to the ith small area, and m is the number of small areas. Let $\gamma_i = N_i^{-1} \sum_{j=1}^{N_i} y_{ij}$ denote the

finite population mean for the ith small area. We use μ_i to denote this quantity when the population is infinite, that is, when N_i goes to ∞. For notational simplicity let $j = 1, \ldots, n_i$ be the sample from the ith small area. Let the vector $\boldsymbol{y}(s)$ denote the sampled values from all the areas. In our treatment of the model-based approach we mostly ignore the sampling design. For exception, see the Fay–Herriot model discussed later, and the pseudo EBLUP, discussed, for example, in Rao (2003a, Chapter 7, and also Chapter 10). A direct estimate for γ_i is the corresponding sample mean $\bar{y}_{is} = n_i^{-1} \sum_{j=1}^{n_i} y_{ij}$, provided $n_i > 0$.

To develop reliable estimates of the small area means it is important that we have useful auxiliary information. An auxiliary variable for which information is available at the population level and correlated with a response variable may be used to develop reliable indirect small area estimates. If the correlation is high, the auxiliary variable will explain a significant portion of the variability of the response variable. We now discuss synthetic and composite estimation in the presence of such a scalar auxiliary variable X. Let x_{ij} denote the value of X associated with the jth unit of the ith small area. Let $\bar{X}_i = N_i^{-1} \sum_{j=1}^{N_i} x_{ij}$, $\bar{x}_{is} = n_i^{-1} \sum_{j=1}^{n_i} x_{ij}$, and $\bar{x}_s = \sum_{i=1}^{m} n_i \bar{x}_{is} / \sum_{i=1}^{m} n_i$. A direct estimator of the finite population mean γ_i is the ratio estimator $(\bar{y}_{is}/\bar{x}_{is})\bar{X}_i$. The "ratio synthetic estimator" of γ_i is given by $\hat{\gamma}_i^{\mathrm{RS}} = (\bar{y}_s/\bar{x}_s)\bar{X}_i$. A composite estimator of γ_i is given by $\hat{\gamma}_i^{*} = (n_i/N_i)\bar{y}_{is} + (1 - n_i/N_i)\hat{\gamma}_i^{\mathrm{RS}}$. If $\bar{X}_i^{*} = (N_i - n_i)^{-1} \sum_{j=n_i+1}^{N_i} x_{ij}$, an alternative composite estimator of γ_i is given by $\hat{\gamma}_i^{C} = (n_i/N_i)\bar{y}_{is} + (1 - n_i/N_i)(\bar{y}_s/\bar{x}_s)\bar{X}_i^{*}$. For a model-based justification of $\hat{\gamma}_i^{C}$ suppose y_{ij} are independent with mean bx_{ij}, and variance $\sigma^2 x_{ij}$ (assuming $x_{ij} > 0$). Since under this model

$$E[\gamma_i | \boldsymbol{y}(s)] = (n_i/N_i)\bar{y}_{is} + (1 - n_i/N_i)b\bar{X}_i^{*}, \tag{1}$$

and the best linear unbiased estimator (BLUE) of b is $\hat{b} = \bar{y}_s/\bar{x}_s$, the estimator $\hat{\gamma}_i^{C}$ is obtained by substituting $\hat{b}$ for b in (1). This estimator is proposed in Holt et al. (1979). A Bayesian interpretation of this estimator is presented in Section 3 of this chatper. To conclude this subsection, we refer the reader to Rao (2003a, Chapter 4) for an authoritative discussion on synthetic and composite estimation of small area means.

2.2. *Linear mixed models in small area estimation*

Synthetic and composite estimates of small area means presented in the last subsection do not employ explicit models. These estimates are proposed based on some implicit models that link a number of small areas using auxiliary information. Although it is relatively easy to get point estimates of small area means using synthetic or composite estimates, a lack of a model makes it difficult often to realize the rationale behind these estimates and to derive suitable measure of uncertainty associated with them. The approach to small area estimation based on explicit models, particularly, linear mixed models has played a key role in the tremendous growth of small area estimation research. In this chapter, we consider both univariate and multivariate small area estimation problems. In the univariate case the dependent variable is a scalar measuring a single characteristic of a unit, and in the multivariate case the dependent variable is a vector consisting of measurements of multiple charcteristics.

2.2.1. Frequentist model-based small area estimation: Univariate case

In this subsection, we consider a number of popular linear mixed models in small area estimation when the characteristic variable or dependendent variable is univariate. Depending on the extent of information for auxiliary variables available for the population units, there are two basic types of small area models: area-level models and unit-level models (cf. Rao, 2003a, Chapter 5). Area-level models are useful when only area-level summary of auxiliary variable is available and unit-specific auxiliary data is unavailable. Using auxiliary data at the area-level, a linear mixed model is proposed for the traditional direct estimators. A basic area-level model, popularly known as the Fay–Herriot model, was first proposed by Fay and Herriot (1979) to produce estimates of per capita income for small places in the United States. On the other hand, unit-level models use unit-specific auxiliary data to build a model for the values of the response variable for all the units in the population. The nested error regression model is a popular unit-level model that was proposed by Battese et al. (1988) to estimate the crop areas under corn and soybeans for certain counties of Iowa.

Prasad and Rao (1990) and Datta and Lahiri (2000) considered the following general normal linear mixed model in small area estimation:

$$\boldsymbol{Y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{v}_i + \boldsymbol{e}_i, \quad i = 1, \ldots, m, \tag{2}$$

where $\boldsymbol{X}_i(n_i \times p)$ and $\boldsymbol{Z}_i(n_i \times b_i)$ are known matrices, $\boldsymbol{v}_i$ and $\boldsymbol{e}_i$ are independently distributed with $\boldsymbol{v}_i \overset{ind}{\sim} N(\boldsymbol{0}, \boldsymbol{G}_i)$ and $\boldsymbol{e}_i \overset{ind}{\sim} N(\boldsymbol{0}, \boldsymbol{R}_i)$. The vector $\boldsymbol{Y}_i$ is $n_i \times 1$ corresponding to the sampled units in the unit-level model, and it is a scalar denoting a direct estimator corresponding to an area-level model. Though $\boldsymbol{e}_i$ models the sampling error, $\boldsymbol{v}_i$ explicitly models the area specific random effects. One important difference between synthetic estimation and model-based small area estimation is in the inclusion of an area specific random effect term that accounts for between area variation not explained by the auxiliary variables included in $\boldsymbol{X}_i$. We assume that $\boldsymbol{G}_i = \boldsymbol{G}_i(\boldsymbol{\psi})$ and $\boldsymbol{R}_i = \boldsymbol{R}_i(\boldsymbol{\psi})$ possibly depend on $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_q)^T$, a vector of variance components. Using the notation of Prasad and Rao (1990), write $\boldsymbol{Y} = \mathrm{col}_{1\le i\le m}\boldsymbol{Y}_i$, $\boldsymbol{e} = \mathrm{col}_{1\le i\le m}\boldsymbol{e}_i$, $\boldsymbol{X} = \mathrm{col}_{1\le i\le m}\boldsymbol{X}_i$, $\boldsymbol{Z} = \mathrm{diag}_{1\le i\le m}\boldsymbol{Z}_i$, $\boldsymbol{G}(\boldsymbol{\psi}) = \mathrm{diag}_{1\le i\le m}\boldsymbol{G}_i$, $\boldsymbol{v} = \mathrm{col}_{1\le i\le m}\boldsymbol{v}_i$, and $\boldsymbol{R}(\boldsymbol{\psi}) = \mathrm{diag}_{1\le i\le m}\boldsymbol{R}_i$. We assume that $\boldsymbol{X}$ has full column rank p. Let $n = \sum_{i=1}^m n_i$ and $b = \sum_{i=1}^m b_i$, and $\boldsymbol{\Sigma}(\boldsymbol{\psi}) = \boldsymbol{R}(\boldsymbol{\psi}) + \boldsymbol{Z}\boldsymbol{G}(\boldsymbol{\psi})\boldsymbol{Z}^T$, the variance–covariance matrix of $\boldsymbol{Y}$. With this notation we can rewrite (2) as

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{v} + \boldsymbol{e}, \tag{3}$$

where $\boldsymbol{v}$ and $\boldsymbol{e}$ are independently distributed with $\boldsymbol{v} \sim N_b(\boldsymbol{0}, \boldsymbol{G})$ and $\boldsymbol{e} \sim N_n(\boldsymbol{0}, \boldsymbol{R})$. This model covers the following two important small area models.

Fay–Herriot Model: In their first application of this area-level model in (4) below, to improve the direct estimator Y_i for estimating the per capita income of small places, denoted by μ_i, Fay and Herriot (1979) assumed that a p-vector of auxiliary variables $\boldsymbol{x}_i$ was available for each area i. They assumed

$$Y_i = \mu_i + e_i, \quad \mu_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + v_i, \quad i = 1, \ldots, m, \tag{4}$$

where v_i and e_i are independent with $e_i \overset{ind}{\sim} N(0, D_i)$ and $v_i \overset{iid}{\sim} N(0, \sigma_v^2)$. Sampling variances D_i are assumed to be known, but the variance component σ_v^2 for the random

effect v_i is unknown. We are interested in estimating μ_i. Here, $n_i = b_i = 1$, $\boldsymbol{Z}_i = 1$, $\boldsymbol{\psi} = \sigma_v^2$, $\boldsymbol{R}_i = D_i$, and $\boldsymbol{G}_i = \sigma_v^2$.

In reality the sampling variances D_i are unknown, and they are estimated from the sample. Because an estimate of D_i based on the sample from the i-th small area may be unreliable (due to small sample), D_i are obtained by smoothing the direct estimates of sampling variances. Smoothing is based on certain model assumptions and provides stability to the D_i. Though in our treatment we will consider D_i is known, Rivest and Vandal (2003) and Wang and Fuller (2003) considered the case of estimating μ_i when D_i is also unknown. Although in the Fay–Herriot model we typically have a diagonal error covariance matrix, this is not necessarily the case since in applications small areas may not be strata (see Subsection 3.3 where correlated sampling errors are considered).

Nested Error Regression Model: This is a unit-level model proposed by Battese et al. (1988) to estimate areas under corn and soybeans for each of 12 counties in North-Central Iowa. The nested error regression model is given by

$$Y_{ij} = \boldsymbol{x}_{ij}^T \boldsymbol{\beta} + v_i + e_{ij}, \quad j = 1, \ldots, n_i, \quad i = 1, \ldots, m, \tag{5}$$

where $\boldsymbol{x}_{ij}$ is a $p \times 1$ vector of auxiliary variables, v_i and e_{ij} are independently distributed with $v_i \overset{iid}{\sim} N(0, \sigma_v^2)$ and $e_{ij} \overset{ind}{\sim} N(0, \sigma_e^2)$, $j = 1, \ldots, n_i$, $i = 1, \ldots, m$. Here, $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{in_i})^T$, $\boldsymbol{X}_i = \mathrm{col}_{1 \le j \le n_i} \boldsymbol{x}_{ij}^T$, $\boldsymbol{Z}_i = \boldsymbol{1}_{n_i}$, $\boldsymbol{e}_i = (e_{i1}, \ldots, e_{in_i})^T$, $\boldsymbol{\psi} = (\sigma_e^2, \sigma_v^2)^T$, $\boldsymbol{R}_i(\boldsymbol{\psi}) = \sigma_e^2 \boldsymbol{I}_{n_i}$, $\boldsymbol{G}_i(\boldsymbol{\psi}) = \sigma_v^2$, where $\boldsymbol{I}_d$ is a $d \times d$ identity matrix and $\boldsymbol{1}_d$ is a $d \times 1$ vector of ones. Let $\bar{\boldsymbol{X}}_i = N_i^{-1} \sum_{j=1}^{N_i} \boldsymbol{x}_{ij}$. If N_i is large so that $N_i^{-1} \sum_{j=1}^{N_i} e_{ij} \approx 0$, a predictor of γ_i may be approximated by a predictor of

$$\mu_i = \bar{\boldsymbol{X}}_i^T \boldsymbol{\beta} + v_i. \tag{6}$$

Random Regression Coefficients Model: The random regression coefficients model of Dempster et al. (1981) has been adapted in small area estimation for unit-level data. Although Datta and Ghosh (1991) in the Bayesian formulation of the small area estimation problem discussed the more general random regression coefficients model, Prasad and Rao (1990) in the frequentist formulation used a simplified version with a single auxiliary variable. The model used by Prasad and Rao (1990) is given by

$$Y_{ij} = \beta_i x_{ij} + e_{ij} = \beta x_{ij} + v_i x_{ij} + e_{ij}, \quad j = 1, \ldots, N_i, \quad i = 1, \ldots, m, \tag{7}$$

where $\beta_i = \beta + v_i$ and v_i and e_{ij} are the same as in the nested error model in (6). Note that $p = 1$. As in the nested error regression model, here $\boldsymbol{X}_i^{(1)} = (x_{i1}, \ldots, x_{in_i})^T$, $\boldsymbol{Z}_i^{(1)} = \boldsymbol{X}_i^{(1)}$. The other entities are the same as in the nested error model. Noting that $\bar{X}_i$ is a scalar, the finite population mean γ_i for the ith area, assuming that N_i is large, may be approximated by

$$\mu_i = \bar{X}_i \beta + \bar{X}_i v_i. \tag{8}$$

2.2.2. *Best linear unbiased predictor (BLUP) and estimated BLUP*

Prasad and Rao (1990) and Datta and Lahiri (2000) under the linear mixed model (3) discussed prediction of a general mixed effect $\eta = \boldsymbol{h}^T \boldsymbol{\beta} + \boldsymbol{\lambda}^T \boldsymbol{v}$, where $\boldsymbol{h}$ and $\boldsymbol{\lambda}$ are known vectors. For example, if we take the ith component of $\boldsymbol{\lambda}$ equal to 1 and all other

components zero, and if we take $\boldsymbol{h} = \bar{\boldsymbol{X}}_i$, then η reduces to μ_i in (6). With the same $\boldsymbol{\lambda}$, and $\boldsymbol{h} = \boldsymbol{x}_i$, the η reduces to μ_i in (4). For the known variance components case, Henderson (1975) derived the BLUP of η given by

$$\tilde{\eta}(\boldsymbol{\psi}, \boldsymbol{Y}) = \boldsymbol{h}^T\tilde{\boldsymbol{\beta}} + \boldsymbol{\lambda}^T\boldsymbol{G}\boldsymbol{Z}^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{Y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}}), \tag{9}$$

where $\tilde{\boldsymbol{\beta}} = \tilde{\boldsymbol{\beta}}(\boldsymbol{\psi}) = (\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{Y}$ is the generalized least squares (GLS) estimator of $\boldsymbol{\beta}$. The normality assumption is not needed to derive the BLUP. Indeed, $\tilde{\boldsymbol{\beta}}$ is the BLUE of $\boldsymbol{\beta}$ and $\tilde{\boldsymbol{v}} = \boldsymbol{G}\boldsymbol{Z}^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{Y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}})$ is the BLUP of $\boldsymbol{v}$. Under the normality assumption mentioned earlier, Henderson et al. (1959) showed that $\tilde{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{v}}$ can be obtained by maximizing $\{-(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{v})^T\boldsymbol{R}^{-1}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{v}) - \boldsymbol{v}^T\boldsymbol{G}^{-1}\boldsymbol{v}\}$ (see also, Rao, 2003a, p. 97). This can be viewed as the posterior mode of $\boldsymbol{\beta}$ and $\boldsymbol{v}$ where the Bayesian model is completed by putting independent prior distribution on $\boldsymbol{v}$ and $\boldsymbol{\beta}$ with $\boldsymbol{v} \sim N_b(\boldsymbol{0}, \boldsymbol{G})$ which is deemed as a random effect and a uniform improper prior on $\boldsymbol{\beta}$, which is deemed as a fixed effect in the standard linear model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{v} + \boldsymbol{e}$. Because the joint posterior distribution of $\boldsymbol{\beta}$ and $\boldsymbol{v}$ is multivariate normal, the BLUE of $\boldsymbol{\beta}$ and the BLUP of $\boldsymbol{v}$ are also the posterior expectations. Thus the BLUP possesses a Bayesian interpretation.

From (9), for the Fay–Herriot model, with $\delta_i = \sigma_v^2(\sigma_v^2 + D_i)^{-1}$, the BLUP of μ_i is given by $\tilde{\mu}_i(\boldsymbol{\psi}, \boldsymbol{Y}) = \boldsymbol{x}_i^T\tilde{\boldsymbol{\beta}} + \delta_i(Y_i - \boldsymbol{x}_i^T\tilde{\boldsymbol{\beta}})$. The BLUP of μ_i for the nested error regression model simplifies to

$$\tilde{\mu}_i(\boldsymbol{\psi}, \boldsymbol{Y}) = \bar{\boldsymbol{X}}_i^T\tilde{\boldsymbol{\beta}} + \delta_i(\bar{Y}_{is} - \bar{\boldsymbol{x}}_{is}^T\tilde{\boldsymbol{\beta}}), \tag{10}$$

where $\bar{\boldsymbol{x}}_{is}$ is the sample mean of $\boldsymbol{x}_{ij}$ for the ith area and $\delta_i = \sigma_v^2(\sigma_v^2 + \sigma_e^2 n_i^{-1})^{-1}$. Under the superpopulation approach if the model given by (5) holds for all the N_i units, following Prasad and Rao (1990) one can show that the BLUP of the finite population mean γ_i under the nested error regression model is given by $\tilde{\gamma}_i(\boldsymbol{\psi}, \boldsymbol{Y}) = f_i\bar{Y}_{is} + (1 - f_i)\tilde{\mu}_{i(u)}(\boldsymbol{\psi}, \boldsymbol{Y})$, where $f_i = n_i/N_i$, $\tilde{\mu}_{i(u)}(\boldsymbol{\psi}, \boldsymbol{Y})$ is given by (10), with $\bar{\boldsymbol{X}}_i$ replaced by $\bar{\boldsymbol{x}}_{i(u)}$, the mean of $\boldsymbol{x}_{ij}$'s for the $N_i - n_i$ unsampled units from the ith area.

The BLUP of η, or in particular, the BLUP of the small area mean γ_i usually depends on the ratios of the variance components, in which practice will be unknown. Replacing the unknown $\boldsymbol{\psi}$ in $\tilde{\eta}(\boldsymbol{\psi}, \boldsymbol{Y})$ by an estimator of $\boldsymbol{\psi}$ leads to a two-stage estimator of η, which is more popularly known as an empirical or estimated BLUP, or EBLUP. We denote an EBLUP of η by $\hat{\eta}(\boldsymbol{Y})$, which is the same as $\tilde{\eta}(\hat{\boldsymbol{\psi}}, \boldsymbol{Y})$. For simplicity of notation, we will often denote the EBLUP by $\hat{\eta}$. In particular, for the nested error model, the EBLUP of μ_i is given by $\tilde{\mu}_i(\hat{\boldsymbol{\psi}}, \boldsymbol{Y}) = \bar{\boldsymbol{X}}_i^T\tilde{\boldsymbol{\beta}}(\hat{\boldsymbol{\psi}}) + \delta_i\hat{\boldsymbol{\psi}}(\bar{Y}_{is} - \bar{\boldsymbol{x}}_{is}^T\tilde{\boldsymbol{\beta}}(\hat{\boldsymbol{\psi}})$. In practice, some of the small areas may have no sample (that it, $n_i = 0$ for some i). For such an area with no sample, the EBLUP of μ_i (or, equivalently, of γ_i) is given by the model-based estimator $\bar{\boldsymbol{X}}_i^T\tilde{\boldsymbol{\beta}}(\hat{\boldsymbol{\psi}})$. In contrast with the design-based approach where there is no estimator of a small area mean if the area sample size is zero, a model-based estimator exists. This is an advantage of the model-based approach to small area estimation.

Kackar and Harville (1984) and Harville (1985) showed that if $\hat{\boldsymbol{\psi}}$ is an even function of $\boldsymbol{y}$ (that is $\hat{\boldsymbol{\psi}}(\boldsymbol{y}) = \hat{\boldsymbol{\psi}}(-\boldsymbol{y})$), and $\hat{\boldsymbol{\psi}}$ is a translation invariant function of $\boldsymbol{y}$ (that is $\hat{\boldsymbol{\psi}}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{a}) = \hat{\boldsymbol{\psi}}(\boldsymbol{y})$ for all p-component vectors $\boldsymbol{a}$), then the EBLUP is an unbiased predictor of η provided the EBLUP has finite expectation. Most reasonable estimators of $\boldsymbol{\psi}$ satisfy these assumptions.

2.2.3. Second-order approximation to MSE of EBLUP

To obtain an EBLUP various methods of estimating variance components have been considered. These methods have been carefully reviewed by Rao (2003a, Chapter 6). Customarily, the variance component vector $\boldsymbol{\psi}$ is estimated by some consistent (for large m) estimator $\hat{\boldsymbol{\psi}}$. To derive the EBLUP Prasad and Rao (1990) used method of moments approach via the method of fitting constants, more popularly known as Henderson's method 3, to get variance components estimates (also known as ANOVA estimates). On the other hand, Datta and Lahiri (2000) considered both the maximum likelihood (ML) and the residual maximum likelihood (REML) methods of estimating the variance components in small area estimation setup (see also Cressie, 1992, for REML estimation in an application of small area estimation). Unlike the other methods of estimating variance components to be discussed here, ANOVA estimators have closed-form expressions.

Prasad and Rao (1990) to obtain unbiased estimators for variance components in the nested error regression model first computed $\hat{e}_{ij}$, $\hat{u}_{ij}$, $j = 1, \ldots, n_i$, $i = 1, \ldots, m$, where $\{\hat{e}_{ij},\ j = 1, \ldots, n_i,\ i = 1, \ldots, m\}$ are the residuals from the ordinary least squares regression of $Y_{ij} - \bar{Y}_{is}$ on $\{\boldsymbol{x}_{ij} - \bar{\boldsymbol{x}}_{is}\}$ and $\hat{u}_{ij}$ are the residuals from the ordinary least squares regression of Y_{ij} on $\boldsymbol{x}_{ij}$. Differences $Y_{ij} - \bar{Y}_{is}$ are free from the random effects v_i and thus depend only on the error variance component σ_e^2. If $n_* = n - tr[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\sum_{i=1}^m n_i^2 \bar{\boldsymbol{x}}_{is}\bar{\boldsymbol{x}}_{is}^T]$, and p^* is equal to the number of linearly independent vectors in the set $\{\boldsymbol{x}_{ij} - \bar{\boldsymbol{x}}_{is},\ j = 1, \ldots, n_i,\ i = 1, \ldots, m\}$, the unbiased estimators are

$$\hat{\sigma}_e^2 = (n - m - p^*)^{-1} \sum_{i=1}^{m} \sum_{j=1}^{n_i} \hat{e}_{ij}^2, \quad \text{and} \quad \hat{\sigma}_v^2 = n_*^{-1} \left[\sum_{i=1}^{m} \sum_{j=1}^{n_i} \hat{u}_{ij}^2 - (n - p)\hat{\sigma}_e^2 \right]. \tag{11}$$

For the Fay–Herriot model Prasad and Rao (1990) show that the ANOVA estimator of σ_v^2, an unbiased quadratic estimator, is given by $\hat{\sigma}_v^2 = (m - p)^{-1}\left[\sum_{i=1}^m \hat{u}_i^2 - \sum_{i=1}^m D_i\{1 - \boldsymbol{x}_i^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_i\}\right]$, where $\hat{u}_i = Y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$. Because D_i is approximately known, even if their true values are unknown, it can be checked that if $sup_{i\geq 1} D_i$ is finite and $m^{-1}\sum D_i$ has a limit, this estimator of σ_v^2 is consistent.

If the ANOVA estimator of σ_v^2 for the aforementioned models turns out to be negative, Prasad and Rao (1990) suggested truncating the negative estimate at zero. They have also shown that the probability of having a negative estimator of σ_v^2 goes very fast to zero as $m \to \infty$. In this context, instead of taking the estimator as zero, one may also consider an alternative proposal due to Wang and Fuller (2003, eq. (17)).

Fay and Herriot (1979) suggested the unbiased estimating equation

$$\sum_{i=1}^{m} \frac{(Y_i - \boldsymbol{x}_i^T\tilde{\boldsymbol{\beta}})^2}{D_i + \sigma_v^2} = m - p, \tag{12}$$

to estimate σ_v^2 in their model. The equation is iteratively solved subject to the condition $\sigma_v^2 \geq 0$, where $\tilde{\beta}$ is given earlier. Datta et al. (2005) have studied this estimator extensively. See Torabi (2006) for an extension of Fay and Herriot (1979) estimating function approach in the nested error regression model.

The MSE of an EBLUP measures the accuracy of the point estimator. The MSE of an EBLUP $\hat{\eta}$, denoted by $\mathrm{MSE}(\hat{\eta})$ is given by $E[\hat{\eta}-\eta]^2$. Kackar and Harville (1984) showed under normality that for a translation-invariant estimator of the variance components $\boldsymbol{\psi}$ that depends only on $|\boldsymbol{Y}|$ the MSE of an EBLUP $\hat{\eta}$ can be decomposed as

$$\mathrm{MSE}(\hat{\eta}) = \mathrm{MSE}(\tilde{\eta}) + E[\tilde{\eta}(\hat{\boldsymbol{\psi}}) - \tilde{\eta}(\boldsymbol{\psi})]^2, \tag{13}$$

where the first term on the right-hand side of (13) is the MSE of the BLUP $\tilde{\eta}$ and the second term accounts for the estimation of the variance components. Although the first term has a closed-form expression, the second term usually has no explicit form. Without the normality assumption and the assumption on the variance components estimators, a cross-product term will also appear on the right-hand side of (13).

Using Henderson's (1975) general result on MSE of the BLUP, or from Prasad and Rao (1990) and Datta and Lahiri (2000), it follows that $\mathrm{MSE}[\tilde{\eta}(\boldsymbol{\psi})] = g_1(\boldsymbol{\psi}) + g_2(\boldsymbol{\psi})$, where $g_1(\boldsymbol{\psi}) = \boldsymbol{\lambda}^T\boldsymbol{G}(\boldsymbol{\psi})\boldsymbol{\lambda} - \boldsymbol{\lambda}^T\boldsymbol{G}(\boldsymbol{\psi})\boldsymbol{Z}^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\psi})\boldsymbol{Z}\ \boldsymbol{G}(\boldsymbol{\psi})\boldsymbol{\lambda}$, $g_2(\boldsymbol{\psi}) = [\boldsymbol{h} - \boldsymbol{X}^T\boldsymbol{s}(\boldsymbol{\psi})]^T(\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\psi})\boldsymbol{X})^{-1}[\boldsymbol{h} - \boldsymbol{X}^T\boldsymbol{s}(\boldsymbol{\psi})]$ and $\boldsymbol{s}(\boldsymbol{\psi}) = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\psi})\boldsymbol{Z}\boldsymbol{G}(\boldsymbol{\psi})\boldsymbol{\lambda}$. If $\eta = \mu_i$, for the nested error model Prasad and Rao (1990) show that $g_1(\boldsymbol{\psi}) = (1-\delta_i)\sigma_v^2 = g_{1i}(\boldsymbol{\psi})$ (say), $g_2(\boldsymbol{\psi}) = (\bar{\boldsymbol{X}}_i - \delta_i\bar{\boldsymbol{x}}_{is})^T\times\ (\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\psi})\boldsymbol{X})^{-1}(\bar{\boldsymbol{X}}_i - \delta_i\bar{\boldsymbol{x}}_{is}) = g_{2i}(\boldsymbol{\psi})$ (say). Similarly, for the Fay–Herriot model these two terms are $g_{1i}(\boldsymbol{\psi}) = (1-\delta_i)\sigma_v^2$, and $g_{2i}(\boldsymbol{\psi}) = (1-\delta_i)^2\boldsymbol{x}_i^T(\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\psi})\boldsymbol{X})^{-1}\boldsymbol{x}_i$, respectively.

The second term in (13) is usually of the order $O(m^{-1})$ and a naive approximation to the MSE of EBLUP by ignoring this term may be a serious underestimate of the correct MSE. The leading term in $\tilde{\eta}(\hat{\boldsymbol{\psi}}) - \tilde{\eta}(\boldsymbol{\psi})$ is $\{\boldsymbol{s}(\hat{\boldsymbol{\psi}}) - \boldsymbol{s}(\boldsymbol{\psi})\}^T\boldsymbol{Y}$. Using Taylor's expansion on the leading term, Prasad and Rao (1990) and Datta and Lahiri (2000) showed a second-order accurate approximation to $E[\tilde{\eta}(\hat{\boldsymbol{\psi}}) - \tilde{\eta}(\boldsymbol{\psi})]^2$ is given by

$$E[\tilde{\eta}(\hat{\boldsymbol{\psi}}) - \tilde{\eta}(\boldsymbol{\psi})]^2 = \mathrm{tr}[\mathrm{var}(\boldsymbol{L}(\boldsymbol{\psi})\boldsymbol{Y}^{(1)})\mathrm{var}(\hat{\boldsymbol{\psi}})] + o(m^{-1}), \tag{14}$$

where $o(m^{-1})$ denotes the neglected terms are of lower order than m^{-1}, and $\boldsymbol{L}(\boldsymbol{\psi}) = \mathrm{col}_{1\le d\le q}\boldsymbol{L}_d^T(\boldsymbol{\psi})$ with $\boldsymbol{L}_d(\boldsymbol{\psi}) = \frac{\partial}{\partial\psi_d}\boldsymbol{s}(\boldsymbol{\psi})$. The first term in the right-hand side of (14) is of order $O(m^{-1})$. Denoting this term by $g_3(\boldsymbol{\psi})$, it follows from (13) that

$$\mathrm{MSE}[\hat{\eta}] = g_1(\boldsymbol{\psi}) + g_2(\boldsymbol{\psi}) + g_3(\boldsymbol{\psi}) + o(m^{-1}). \tag{15}$$

Derivations of the second-order approximations mentioned earlier need certain regularity assumptions; see Prasad and Rao (1990) and Datta and Lahiri (2000). Note that while the first two terms of the MSE approximation in (15) mentioned earlier remain the same and do not depend on the variance components estimation method, the last term, namely the $g_3(\boldsymbol{\psi})$ term depends on the estimator of the variance components. Derivation of the MSE approximation requires $\sqrt{m}$-consistent estimators of the variance components. Jiang (1996) showed consistency of the REML estimators. Prasad and Rao (1990) proved $\sqrt{m}$-consistency of the ANOVA estimators. Because Datta and Lahiri (2000) showed that the asymptotic variances of the ML and REML estimators of the variance components are the same up to the order of $O(m^{-1})$ terms, both these methods lead to the same g_3 term. In fact, in the class of consistent estimators of the variance components, since the ML estimators have the "smallest" asymptotic variance, it follows from (14) and (15) that an EBLUP based on the ML/REML estimators of the variance components will have the smallest approximate MSE. This was noted by Datta et al. (2005) in the special case of Fay–Herriot model.

To evaluate the g_3 term for some useful small area models such as the nested error model and the Fay–Herriot model, Prasad and Rao (1990) provided asymptotic variance–covariance matrix of the ANOVA estimators of the variance components. Datta and Lahiri (2000) did the same both for the ML and REML estimators (in this context, see also Das et al., 2004, for a rigorous proof). Prasad and Rao (1990) showed for the nested error regression model that associated with the EBLUP of μ_i in (6), the third term of (15) is $g_{3i}(\boldsymbol{\psi}) = n_i^{-2}(\sigma_v^2 + \sigma_e^2/n_i)^{-3}\mathrm{var}(\sigma_v^2\hat{\sigma}_e^2 - \sigma_e^2\hat{\sigma}_v^2)$. For the Fay–Herriot model for area-level data, under the regularity assumption $0 < \inf_{i\geq 1} D_i \leq \sup_{i\geq 1} D_i < \infty$, the third term of the MSE approximation, from Prasad and Rao (1990) and Datta and Lahiri (2000) is $g_{3i}(\boldsymbol{\psi}) = D_i^2(\sigma_v^2 + D_i)^{-3}\mathrm{var}(\hat{\sigma}_v^2)$. First two terms, g_{1i} and g_{2i}, are provided earlier.

For the Fay–Herriot model, Datta et al. (2005) derived the second-order approximation of the MSE of the EBLUP using the method of moments (MOM) estimator (cf. (12)) of the variance component suggested by Fay and Herriot (1979) in their classic paper. For the Fay–Herriot model Datta et al. (2005) showed that the g_{3i} term is the largest for the Prasad-Rao method, is the smallest for the ML/REML method, and is in between these two for the MOM suggested by Fay and Herriot (1979).

Often we may be interested in the prediction of a linear combination of several small area means γ_i, e.g., we may be interested in $\gamma_1 - \gamma_2$. Although the EBLUP of the linear combination is given by the corresponding linear combination of the EBLUPs of γ_i, the MSE is not simply a function of the component MSEs of the γ_i. To cover this general case we consider prediction of $\boldsymbol{\eta} = \boldsymbol{H}\boldsymbol{\beta} + \boldsymbol{\Lambda}\boldsymbol{v}$, a vector of linear combinations of $\boldsymbol{\beta}$ and $\boldsymbol{v}$. Let $\boldsymbol{\eta}$ be a $u \times 1$ vector, with the ath rows of $\boldsymbol{H}$ and $\boldsymbol{\Lambda}$ are denoted by $\boldsymbol{h}_a^T$ and $\boldsymbol{\lambda}_a^T$, respectively. Let $\tilde{\boldsymbol{\eta}}$ be the BLUP of $\boldsymbol{\eta}$. Then

$$\tilde{\boldsymbol{\eta}}(\boldsymbol{\psi}, \boldsymbol{Y}) = \boldsymbol{H}\tilde{\boldsymbol{\beta}} + \boldsymbol{S}^T(\boldsymbol{Y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}}), \tag{16}$$

where $\boldsymbol{S} \equiv \boldsymbol{S}(\boldsymbol{\psi}) = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\psi})\boldsymbol{Z}\boldsymbol{G}(\boldsymbol{\psi})\boldsymbol{\Lambda}^T$. Let $\hat{\boldsymbol{\eta}}$ be the EBLUP given by $\hat{\boldsymbol{\eta}} = \tilde{\boldsymbol{\eta}}(\hat{\boldsymbol{\psi}}, \boldsymbol{Y})$. Then a second-order approximation of the MSE (the matrix of mean squareds error and the mean products error) of the EBLUP is given by

$$\mathrm{MSE}(\hat{\boldsymbol{\eta}}) = E[(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta})(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta})^T] = \mathrm{MSE}(\tilde{\boldsymbol{\eta}}) + \boldsymbol{G}_3(\boldsymbol{\psi}) + o(m^{-1}), \tag{17}$$

$\mathrm{MSE}(\tilde{\boldsymbol{\eta}}) = \boldsymbol{G}_1(\boldsymbol{\psi}) + \boldsymbol{G}_2(\boldsymbol{\psi})$, $\boldsymbol{G}_1(\boldsymbol{\psi}) = \boldsymbol{\Lambda}\boldsymbol{G}(\boldsymbol{\psi})\boldsymbol{\Lambda}^T - \boldsymbol{\Lambda}\boldsymbol{G}(\boldsymbol{\psi})\boldsymbol{Z}^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\psi})\boldsymbol{Z}\boldsymbol{G}(\boldsymbol{\psi})\boldsymbol{\Lambda}^T$, $\boldsymbol{G}_2(\boldsymbol{\psi}) = [\boldsymbol{H} - \boldsymbol{S}(\boldsymbol{\psi})^T\boldsymbol{X}](\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\psi})\boldsymbol{X})^{-1}[\boldsymbol{H} - \boldsymbol{S}(\boldsymbol{\psi})^T\boldsymbol{X}]^T$, and the (a, b)th element of $\boldsymbol{G}_3(\boldsymbol{\psi})$ is given by

$$g_{3,(a,b)}(\boldsymbol{\psi}) = \mathrm{tr}[\mathrm{cov}(\boldsymbol{L}^{(a)}(\boldsymbol{\psi})\boldsymbol{Y}, \boldsymbol{L}^{(b)}(\boldsymbol{\psi})\boldsymbol{Y})\mathrm{var}(\hat{\boldsymbol{\psi}})]. \tag{18}$$

In the above, $\boldsymbol{L}^{(a)}(\boldsymbol{\psi})$ is obtained from the definition of $\boldsymbol{L}(\boldsymbol{\psi})$ with $\boldsymbol{\lambda}^T$ replaced by $\boldsymbol{\lambda}_a^T$, the ath row of $\boldsymbol{\Lambda}$. Note that $\boldsymbol{L}(\boldsymbol{\psi})$ is defined immediately following (14).

Now, we evaluate the components of (17) and (18) for the Fay–Herriot model for $\boldsymbol{\eta} = (\mu_1, \mu_2)^T$. By simple matrix calculations, we get $\boldsymbol{G}_1 = \sigma_v^2 Diag(1 - \delta_1, 1 - \delta_2)$ (recall $\delta_i = \sigma_v^2(\sigma_v^2 + D_i)^{-1}$. So $\boldsymbol{G}_1$ is a diagonal matrix with the diagonal elements g_{1i}, $i = 1, 2$, given earlier. The matrix $\boldsymbol{G}_2$ typically is not diagonal. Its two diagonal elements g_{2i}, $1 = 1, 2$, are given earlier. The off-diagonal element of $\boldsymbol{G}_2$ is $(1 - \delta_1)(1 - \delta_2)$ $\boldsymbol{x}_1^T(\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{x}_2$. Similarly, it can be shown that $\boldsymbol{G}_3$ is a diagonal matrix with its diagonal elements given by g_{3i} for $i = 1, 2$. Thus, the off-diagonal element of the MSE

matrix is non-zero if and only if the off-diagonal element of $\boldsymbol{G}_2$ is non-zero. More generally, it can be shown that if the components of $\boldsymbol{\eta}$ are small area means, then for the nested error regression model, the Fay–Herriot model and the random regression coefficients model, off-diagonal elements of $\boldsymbol{G}_1$ and $\boldsymbol{G}_3$ are all zeros. The (i, j)th off-diagonal element of the MSE matrix corresponds to the mean product error in prediction for the ith and the jth areas. Lahiri and Rao (1995) obtained an approximate expression of this term in a robust version of the Fay–Herriot model.

2.2.4. Estimator of MSE approximation

In this section, we now obtain second-order unbiased estimators of the MSE of the EBLUP $\tilde{\eta}(\hat{\boldsymbol{\psi}})$ for various methods of estimating the variance components. We say that an estimator $\text{mse}(\tilde{\eta}(\hat{\boldsymbol{\psi}}))$ is second-order unbiased estimator if $E[\text{mse}(\tilde{\eta}(\hat{\boldsymbol{\psi}})) - \text{MSE}(\tilde{\eta}(\hat{\boldsymbol{\psi}}))] = o(m^{-1})$. Let $\boldsymbol{b}(\hat{\boldsymbol{\psi}}; \boldsymbol{\psi})$ be the asymptotic bias of $\hat{\boldsymbol{\psi}}$ up to the order of $o(m^{-1})$. Denote the gradient vector of $g_1(\boldsymbol{\psi})$, namely the vector of partial derivatives of $g_1(\boldsymbol{\psi})$ with respect to $\boldsymbol{\psi}$ by $\nabla g_1(\boldsymbol{\psi})$. From Theorem A.2 of Datta and Lahiri (2000) we get

$$E[g_1(\hat{\boldsymbol{\psi}})] = g_1(\boldsymbol{\psi}) + \boldsymbol{b}^T(\hat{\boldsymbol{\psi}}; \boldsymbol{\psi})\nabla g_1(\boldsymbol{\psi}) - g_3(\boldsymbol{\psi}) + o(m^{-1}). \tag{19}$$

Also, since $g_2(\boldsymbol{\psi})$ and $g_3(\boldsymbol{\psi})$ are of order $O(m^{-1})$, it follows that

$$E[g_2(\hat{\boldsymbol{\psi}})] = g_2(\boldsymbol{\psi}) + o(m^{-1}), \quad E[g_3(\hat{\boldsymbol{\psi}})] = g_3(\boldsymbol{\psi}) + o(m^{-1}). \tag{20}$$

From (19) and (20) it follows that the naive estimator $\text{mse}^N(\tilde{\eta}(\hat{\boldsymbol{\psi}})) = g_1(\hat{\boldsymbol{\psi}}) + g_2(\hat{\boldsymbol{\psi}})$ as well as the "plug-in" estimator $g_1(\hat{\boldsymbol{\psi}}) + g_2(\hat{\boldsymbol{\psi}}) + g_3(\hat{\boldsymbol{\psi}})$ are biased to the order of $O(m^{-1})$ in estimating $\text{MSE}(\tilde{\eta}(\hat{\boldsymbol{\psi}}))$. However, by (15), (19), and (20),

$$\text{mse}(\tilde{\eta}(\hat{\boldsymbol{\psi}})) = g_1(\hat{\boldsymbol{\psi}}) + g_2(\hat{\boldsymbol{\psi}}) + 2g_3(\hat{\boldsymbol{\psi}}) - \boldsymbol{b}^T(\hat{\boldsymbol{\psi}}; \hat{\boldsymbol{\psi}})\nabla g_1(\hat{\boldsymbol{\psi}}) \tag{21}$$

is a second-order unbiased estimator of $\text{MSE}(\tilde{\eta}(\hat{\boldsymbol{\psi}}))$, where $\boldsymbol{b}(\hat{\boldsymbol{\psi}}; \hat{\boldsymbol{\psi}})$ is the estimated bias of $\hat{\boldsymbol{\psi}}$.

In particular, for the ANOVA estimators $\hat{\boldsymbol{\psi}}_A$ of the variance components Prasad and Rao (1990) noted for the important small area models that the asymptotic bias $b(\hat{\boldsymbol{\psi}}_A; \boldsymbol{\psi})$ is zero. From (21) it follows that (see Prasad and Rao, 1990, for details)

$$\text{mse}(\tilde{\eta}(\hat{\boldsymbol{\psi}}_A)) = g_1(\hat{\boldsymbol{\psi}}_A) + g_2(\hat{\boldsymbol{\psi}}_A) + 2g_3(\hat{\boldsymbol{\psi}}_A) \tag{22}$$

is a second-order unbiased estimator of $\text{MSE}[\tilde{\eta}(\hat{\boldsymbol{\psi}}_A)]$.

For the REML estimator $\hat{\boldsymbol{\psi}}_{\text{RE}}$ of $\boldsymbol{\psi}$ Datta and Lahiri (2000) noted for the model given by (3) that the asymptotic bias $\boldsymbol{b}(\hat{\boldsymbol{\psi}}_{\text{RE}}; \boldsymbol{\psi})$ is $o(m^{-1})$. Again from (21)

$$\text{mse}(\tilde{\eta}(\hat{\boldsymbol{\psi}}_{\text{RE}})) = g_1(\hat{\boldsymbol{\psi}}_{\text{RE}}) + g_2(\hat{\boldsymbol{\psi}}_{\text{RE}}) + 2g_3(\hat{\boldsymbol{\psi}}_{\text{RE}}) \tag{23}$$

is a second-order unbiased estimator of $\text{MSE}[\tilde{\eta}(\hat{\boldsymbol{\psi}}_{\text{RE}})]$. Even though the ML and the REML estimators have the same asymptotic variances, using a similar expression obtained from (23) after replacing $\hat{\boldsymbol{\psi}}_{\text{RE}}$ by $\hat{\boldsymbol{\psi}}_{\text{ML}}$ does not hold for the ML estimators since these estimators are biased. For more details we refer to Datta and Lahiri (2000). In particular, for the Fay–Herriot model they have shown that the MLE of the variance component is negatively biased, and use of the expression in (23) by replacing the

REML estimator $\hat{\boldsymbol{\psi}}_{\mathrm{RE}}$ by the MLE $\hat{\boldsymbol{\psi}}_{\mathrm{ML}}$ will underestimate the true MSE. For the MLE it follows from (21) that a second-order unbiased estimator of MSE$[\tilde{\eta}(\hat{\boldsymbol{\psi}}_{\mathrm{ML}})]$ is

$$\mathrm{mse}(\tilde{\eta}(\hat{\boldsymbol{\psi}}_{\mathrm{ML}})) = g_1(\hat{\boldsymbol{\psi}}_{\mathrm{ML}}) + g_2(\hat{\boldsymbol{\psi}}_{\mathrm{ML}}) + 2g_3(\hat{\boldsymbol{\psi}}_{\mathrm{ML}}) - \boldsymbol{b}^T(\hat{\boldsymbol{\psi}}_{\mathrm{ML}}; \hat{\boldsymbol{\psi}}_{\mathrm{ML}})\nabla g_1(\hat{\boldsymbol{\psi}}_{\mathrm{ML}}). \tag{24}$$

Since $\hat{\boldsymbol{\psi}}_{\mathrm{ML}} - \hat{\boldsymbol{\psi}}_{\mathrm{RE}} = \boldsymbol{b}(\hat{\boldsymbol{\psi}}_{\mathrm{ML}}; \hat{\boldsymbol{\psi}}_{\mathrm{ML}}) + o_p(m^{-1})$ (cf. Datta and Lahiri (2000)) the right-hand sides of (23) and (24) differ from each other by $o_p(m^{-1})$.

From Datta et al. (2005) and Torabi (2006, Section 4.6), $O(m^{-1})$ bias terms of MOM estimators of variance components are usually non-zeros. Although Torabi (2006) obtained the first-order bias terms for the nested error regression model, Datta et al. (2005) derived a similar expression for the Fay–Herriot model. While the bias terms are fairly complicated for the nested error model, the first-order asymptotic bias of $\hat{\sigma}^2_{v,\mathrm{FH}}$, the MOM estimator in the Fay–Herriot model is $b\left(\hat{\sigma}^2_{v,\mathrm{FH}}; \sigma_v^2\right) = 2\left[m\sum_{i=1}^m(\sigma_v^2 + D_i)^{-2} - \left\{\sum_{i=1}^m(\sigma_v^2 + D_i)^{-1}\right\}^2\right] / \left\{\sum_{i=1}^m(\sigma_v^2 + D_i)^{-1}\right\}^3 + o(m^{-1})$. For the balanced Fay–Herriot model, where all the sampling variances D_i are equal, the first-order bias reduces to zero. In fact, in this case the MOM estimator and the ANOVA estimator are identical. Using $b(\hat{\sigma}^2_{v,\mathrm{FH}}; \hat{\sigma}^2_{v,\mathrm{FH}})$ in (21) we get the second-order unbiased estimator of the MSE of the EBLUP of η [see Eq. (16) of Datta et al. (2005)].

We will now provide an expression of a second-order unbiased estimator of the MSE matrix of the EBLUP vector $\hat{\boldsymbol{\eta}}$. Denoting the estimator by $\mathrm{mse}(\hat{\boldsymbol{\eta}})$ and generalizing (21), the second-order unbiased estimator is given by

$$\mathrm{mse}(\hat{\boldsymbol{\eta}}) = \boldsymbol{G}_1(\hat{\boldsymbol{\psi}}) + \boldsymbol{G}_2(\hat{\boldsymbol{\psi}}) + 2\boldsymbol{G}_3(\hat{\boldsymbol{\psi}}) - \boldsymbol{Q}(\hat{\boldsymbol{\psi}}), \tag{25}$$

where the (a, b)th element of $\boldsymbol{Q}(\hat{\boldsymbol{\psi}})$ is given by $\boldsymbol{b}^T(\hat{\boldsymbol{\psi}}; \hat{\boldsymbol{\psi}})\nabla G_{1ab}(\hat{\boldsymbol{\psi}})$, with $G_{1ab}(\boldsymbol{\psi})$ denoting the (a, b)th element of $\boldsymbol{G}_1$. On the basis of the discussion in the paragraph following Eq. (18), if components of $\boldsymbol{\eta}$ correspond to small area means, then a second-order accurate expression of the mean product error will involve only the second term in the right-hand side of (25).

Estimation of the MSE of EBLUP outlined earlier is based on Taylor's expansion. Alternatively, a resampling-based approach may be used to estimate the MSE. See Chapter 28 for a discussion.

2.2.5. Multivariate models for small area estimation

A multivariate approach in small area estimation was advocated by Fay (1987) to accurately estimate median income for four-person families for all the U.S. states. Using the Current Population Survey (CPS) estimates of median income for three-, four-, and five-person families, Fay (1987) suggested the multivariate area-level model to produce more accurate estimates for the four-person families at the state level. These state-level median income estimates are obtained from the Annual Demographic Supplement to the March CPS sample. He recommended an empirical Bayes approach and used data from the U.S. decennial census and other administrative records as covariates. Fuller and Harter (1987) considered a finite population approach based on unit-level records to develop EBLUP of the finite population mean vector of multiple characteristics at the small area level. Their model is a multivariate extension of the univariate nested error regression model in Battese et al. (1988). Datta et al. (1999a) also considered this

mulitivariate problem. Although the estimator of the small area mean of a component variable in a multivariate approach is intuitively expected to be more reliable than the corresponding estimator based on a univariate model, Datta et al. (1999a) showed that in the bivariate case the gain actually depends on the strengths and signs of the correlations in the two covariance matrices in the Model M given later.

We here consider a model suitable for unit-level data. A multivariate model for area-level data is similar, and an application of this model is considered in the next section. Suppose there are N_i units in the small area i with $\boldsymbol{Y}_{ij}$ denoting the $s \times 1$ vector of response variables and $\boldsymbol{c}_{ij}$, a d-component vector of fixed covariates associated with the jth unit in the ith small area, $j = 1, \ldots, N_i$, $i = 1, \ldots, m$. Datta et al. (1999a) considered prediction of $\boldsymbol{\gamma}_i = N_i^{-1} \sum_{j=1}^{N_i} \boldsymbol{Y}_{ij}$ for $i = 1, \ldots, m$ based on the following model.

Model M: (a) Conditional on $\boldsymbol{v}_i$, $\boldsymbol{Y}_{ij} \sim N_s(\boldsymbol{B}\boldsymbol{c}_{ij} + \boldsymbol{v}_i, \boldsymbol{\Sigma}_e)$ independently for $j = 1, \ldots, n_i, i = 1, \ldots, m$; (b) $\boldsymbol{v}_i \sim N_s(\boldsymbol{0}, \boldsymbol{\Sigma}_v)$ independently for $i = 1, \ldots, m$, where $\boldsymbol{B} = ((b_{tu}))$ is an $s \times d$ matrix of regression coefficients, $\boldsymbol{v}_i$ is an s-component vector of small area effects, $\boldsymbol{\Sigma}_e(s \times s)$ is a matrix of sampling variance, and $\boldsymbol{\Sigma}_v(s \times s)$ is the variance–covariance matrix of $\boldsymbol{v}_i$.

Let $\boldsymbol{U} = (\boldsymbol{Y}_{11}, \ldots, \boldsymbol{Y}_{1n_1}, \ldots, \boldsymbol{Y}_{m1}, \ldots, \boldsymbol{Y}_{mn_m})$, $\boldsymbol{A} = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_m)$, $\boldsymbol{F}^T = \oplus_{i=1}^m \boldsymbol{1}_{n_i}^T$, and $\boldsymbol{C}^T = (\boldsymbol{c}_{11}, \ldots, \boldsymbol{c}_{1n_1}, \ldots, \boldsymbol{c}_{m1}, \ldots, \boldsymbol{c}_{mn_m})$ where $\oplus_{i=1}^m$ denotes the Kronecker sum of matrices. Using this notation Datta et al. (1999a) expressed (a) and (b) of Model M as a multivariate linear mixed model given by

$$\boldsymbol{U} = \boldsymbol{B}\boldsymbol{C}^T + \boldsymbol{A}\boldsymbol{F}^T + \boldsymbol{E}, \tag{26}$$

where $\boldsymbol{E} = (\boldsymbol{e}_{11}, \ldots, \boldsymbol{e}_{1n_1}, \ldots, \boldsymbol{e}_{m1}, \ldots, \boldsymbol{e}_{mn_m})$, with $\boldsymbol{e}_{ij}$ being iid with $N_s(\boldsymbol{0}, \boldsymbol{\Sigma}_e)$ and independent of $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_m$.

Under Model M, $\boldsymbol{\gamma}_i = \boldsymbol{\mu}_i + \bar{\boldsymbol{e}}_i$ where $\boldsymbol{\mu}_i = \boldsymbol{B}\bar{\boldsymbol{c}}_{i(P)} + \boldsymbol{v}_i$ can be interpreted as the conditional mean vector of the ith small area given the values of covariates $\boldsymbol{c}_{ij}$ and the realized small area effect $\boldsymbol{v}_i$. Here $\bar{\boldsymbol{e}}_i = N_i^{-1} \sum_{j=1}^{N_i} \boldsymbol{e}_{ij}$ and $\bar{\boldsymbol{c}}_{i(P)} = N_i^{-1} \sum_{j=1}^{N_i} \boldsymbol{c}_{ij}$. For large N_i if $\bar{\boldsymbol{e}}_i \approx 0$, a predictor of the mixed effect vector $\boldsymbol{\mu}_i$ may also be appropriate for predicting $\boldsymbol{\gamma}_i$. Datta et al. (1999a) considered prediction of a mixed effect vector $\boldsymbol{B}\boldsymbol{h} + \boldsymbol{A}\boldsymbol{\lambda}$ for given $\boldsymbol{h}$ and $\boldsymbol{\lambda}$ based on the model (26). If $\boldsymbol{Y} = \mathrm{col}_{1 \le i \le m}(\mathrm{col}_{1 \le j \le n_i} \boldsymbol{Y}_{ij})$, $\boldsymbol{e} = \mathrm{col}_{1 \le i \le m}(\mathrm{col}_{1 \le j \le n_i} \boldsymbol{e}_{ij})$, $\boldsymbol{\beta} = \mathrm{col}_{1 \le u \le d}(\mathrm{col}_{1 \le t \le s} b_{tu})$, $\boldsymbol{v} = \mathrm{col}_{1 \le i \le m} \boldsymbol{v}_i$, $\boldsymbol{X} = \boldsymbol{C} \otimes \boldsymbol{I}_s$, $\boldsymbol{Z} = \boldsymbol{F} \otimes \boldsymbol{I}_s$, $\boldsymbol{H} = \boldsymbol{h}^T \otimes \boldsymbol{I}_s$, and $\boldsymbol{\Lambda} = \boldsymbol{\lambda}^T \otimes \boldsymbol{I}_s$, the model in (26) can be written as the linear mixed model in (3)

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{v} + \boldsymbol{e}, \tag{27}$$

with $\boldsymbol{e} \sim N_{ns}(\boldsymbol{0}, \boldsymbol{R}(\boldsymbol{\psi}))$ independently of $\boldsymbol{v} \sim N_{ns}(\boldsymbol{0}, \boldsymbol{G}(\boldsymbol{\psi}))$, where $n = \sum_{i=1}^m n_i$. Here $\boldsymbol{\psi} = (\boldsymbol{\psi}_v^T, \boldsymbol{\psi}_e^T)^T$ is an $s(s+1)$-component vector of variance parameters in $\boldsymbol{\Sigma}_v$ and $\boldsymbol{\Sigma}_e$. Note that $\boldsymbol{R}(\boldsymbol{\psi}) = \boldsymbol{I}_n \otimes \boldsymbol{\Sigma}_e$, $\boldsymbol{G}(\boldsymbol{\psi}) = \boldsymbol{I}_m \otimes \boldsymbol{\Sigma}_v$, $\boldsymbol{B}\boldsymbol{h} = \boldsymbol{H}\boldsymbol{\beta}$, and $\boldsymbol{A}\boldsymbol{\lambda} = \boldsymbol{\Lambda}\boldsymbol{v}$.

We assume the matrix $\boldsymbol{C}$ is of rank d. Then the matrix $\boldsymbol{X}$ is of full column rank ds. Because the prediction of $\boldsymbol{B}\boldsymbol{h} + \boldsymbol{A}\boldsymbol{\lambda}$ is equivalent to prediction of $\boldsymbol{H}\boldsymbol{\beta} + \boldsymbol{\Lambda}\boldsymbol{v}$ based on the model in (27), we can get BLUP and EBLUP using Eq. (16). The MSE of the BLUP and a second-order accurate approximation to the MSE of the EBLUP can be obtained from Eqs. (17) and (18). A second-order unbiased estimator of the MSE matrix may be obtained from (25).

2.3. Cross-sectional time series estimation

In this section we consider small area methods to produce indirect estimates, both time-indirect and domain-indirect, which borrow strength from areas and time. Since many national surveys are repeated in time, it is possible to use both time series and cross-sectional data in production of small area estimates. For example, the CPS in the United States is a monthly household survey that is implemented as a 4-8-4 rotating panel survey. In this survey sampled units are partially replaced every month where a group of households first time selected in the sample remains in the sample for four consecutive months, is eliminated from the sample for the next eight months, enters the sample again for four more consecutive months, and finally replaced in the sample by a group of nearby households. Similarly, the Canadian Labor Force Survey (CLFS) is a monthly household survey where a selected household is retained in the sample for six consecutive months and then is dropped out of the sample. In repeated surveys, considerable gain in efficiency of the small area estimates is possible by borrowing strength across both small areas and time. An early application of cross-sectional and time series model in small area estimation is by Pfeffermann and Burck (1990) where they considered estimation of housing price indices.

All the papers on time series approach to small area estimation deal with area-level model by suitable extension or modification of the Fay–Herriot model to bring in the time series component. Pfeffermann and Burck (1990) used the sampling model

$$Y_{it} = \theta_{it} + e_{it}, \quad t = 1, \ldots, T; \quad i = 1, \ldots, m, \tag{28}$$

where they assumed the sampling error series e_{it} within each area is serially uncorrelated and normally distributed with known sampling variances. The true mean θ_{it} was specified by a linear model with regression coefficient vector $\boldsymbol{\beta}_{it}$, which were allowed to vary with respect to i and t. The regression coefficients were modeled through a state-space approach. An important feature of this model is that corresponding components of $\boldsymbol{\beta}_{it} - \boldsymbol{T}\boldsymbol{\beta}_{i,t-1}$ pertaining to different areas were allowed to be correlated, where $\boldsymbol{T}$ is a state matrix.

Rao and Yu (1992) used the sampling model (28) and suggested the following linking model

$$\theta_{it} = \boldsymbol{x}_{it}^T \boldsymbol{\beta} + v_i + \alpha_{it}. \tag{29}$$

Here Y_{it} is the direct estimator for small area i at time t. Let $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{iT})^T$, $\boldsymbol{\theta}_i = (\theta_{i1}, \ldots, \theta_{iT})^T$, and $\boldsymbol{Y}$ denote the vector obtained by stacking the columns $\boldsymbol{Y}_i$, $i = 1, \ldots, m$. Rao and Yu (1992, 1994) assumed that the sampling error model is given by $\boldsymbol{Y}_i | \boldsymbol{\theta}_i \sim N(\boldsymbol{\theta}_i, \boldsymbol{\Psi}_i)$, $i = 1, \ldots, m$ independently, where $\boldsymbol{\Psi}_i$ is a known sampling variance–covariance matrix. As in the standard Fay–Herriot model, it is assumed that the small area random effects v_i are independent of the sampling error terms and are i.i.d. $N(0, \sigma_v^2)$. The time series aspect of the series θ_{it} is captured through the α_{it} terms that are assumed to be independent of the sampling error terms and the small area effects terms. Rao and Yu (1994) proposed for each small area a common stationary AR(1) model $\alpha_{it} = \rho\alpha_{i,t-1} + \xi_{it}$, $|\rho| < 1$ with $\xi_{it} \overset{iid}{\sim} N(0, \sigma_\xi^2)$. Another time series generalization of the Fay–Herriot model was proposed by Singh et al. (1994).

Datta et al. (2002) and You (1999) used the model of Rao and Yu but replaced the AR(1) time series model by a random walk model. In their application of EB estimation of four-person family state median income Datta et al. (2002) assumed that $\alpha_{i0} = \alpha$ for all i and that there is no intercept term included in $\boldsymbol{x}_{it}^T\boldsymbol{\beta}$. These conditions are needed to make all the model parameters identifiable. Note that under the Rao and Yu (1994) and Datta et al. (2002) model, the variance–covariance matrix of $\boldsymbol{Y}$ is a block diagonal matrix with the ith block given by $\boldsymbol{\Sigma}_i = \boldsymbol{\Psi}_i + \sigma_v^2\mathbf{1}\mathbf{1}^T + \text{var}(\boldsymbol{\lambda}_i)$, where $\boldsymbol{\lambda}_i = (\alpha_{i1}, \ldots, \alpha_{iT})^T$. The variance–covariance matrix of α_i for the Rao–Yu model is a $T \times T$ matrix with the (t, s)th element given by $\sigma_\xi^2\rho^{|t-s|}/(1-\rho^2)$. For the random walk model of Datta et al. this matrix is given by $\sigma_\xi^2\boldsymbol{\Gamma}\boldsymbol{\Gamma}^T$, where $\boldsymbol{\Gamma}$ is a $T \times T$ lower triangular matrix with all the diagonal elements and the non-zero off-diagonal elements being one.

It is interesting to compare the model of Datta et al. (2002) with the univariate version of the model proposed by Ghosh et al. (1996) (details presented in Section 3). Although Datta et al. (2002) and Rao and Yu (1994) models have the small area specific random effects v_i, the model of Ghosh et al. (1996) does not have this term. Also, they assumed the common time series component $\alpha_{it} = \alpha_t$ for all the areas. Thus, their model is subject to overshrinkage problem.

By expressing their model described earlier as a linear mixed model Rao and Yu (1992, 1994) and Datta et al. (2002) obtained the EBLUP or EB predictor of θ_{iT}, the small area mean at the current time T. We denote all the variance components (and, also the autocorrelation parameter in Rao–Yu model) by $\boldsymbol{\psi}$. Assuming $\boldsymbol{\psi}$ known, by applying the BLUP theory described earlier, it is easy to write down the BLUP of θ_{iT}. From Equation (4.2) of Datta et al. (2002) or Eq. (8.3.2) of Rao (2003a) the BLUP is given by $\tilde{\theta}_{iT\text{BLUP}}(\boldsymbol{\psi}) = w_{iT}y_{iT} + (1-w_{iT})\boldsymbol{x}_{iT}^T\tilde{\boldsymbol{\beta}} + \sum_{t=1}^{T-1} w_{it}(y_{it} - \boldsymbol{x}_{it}^T\tilde{\boldsymbol{\beta}})$, where $\tilde{\boldsymbol{\beta}}$ is the GLS estimator of $\boldsymbol{\beta}$. For the random walk model the weights w_{it} are given by $(w_{i1}, \ldots, w_{iT}) = (\sigma_v^2\mathbf{1}_T + \sigma_\xi^2\boldsymbol{g}_T)^T\boldsymbol{\Sigma}_i^{-1}$ and $\boldsymbol{g}_T$ is the T-th column of $\boldsymbol{\Gamma}\boldsymbol{\Gamma}^T$. Although the vector $\boldsymbol{g}_T$ for the random walk model does not depend on any parameter, its counterpart for the AR(1) model in Rao and Yu (1994) depends on the unknown autocorrelation parameter. To get the weights for the Rao–Yu model, we need to replace $\boldsymbol{g}_T$ by $(1-\rho^2)^{-1}(\rho^{T-1}, \rho^{T-2}, \ldots, 1)^T$. Assuming ρ known Rao and Yu (1994) estimated the variance components by the method of fitting constants. For the random walk model Datta et al. (2002) estimated the variance components by the REML method. In both the cases, by applying the MSE approximation results these authors obtained the MSE approximation of the EBLUP of θ_{iT}. Datta et al. (2002) provided a second-order unbiased mse estimator of the EBLUP of θ_{iT}.

Through simulations Rao and Yu (1994) showed that for known autocorrelation parameter there is considerable reduction in the mse of the EBLUP in the cross-sectional and time series model compared to the mse of the EBLUP in the Fay–Herriot model. For a summary of the simulation, readers may refer to Rao (2003a, p. 160).

Datta et al. (2002) used the random walk error model described earlier to estimate the four-person family annual median income estimates for the 50 U.S. states and Washington, DC. They used the annual CPS direct estimates of median income of four-person families for these 51 small areas for 9 years (1981–89) to produce estimates for the year 1989. Once again we emphasize that the state-level median income estimates are obtained annually from the Annual Demographic Supplement to the March CPS sample. The year 1989 was chosen since the corresponding values were available from

the Census that were believed to be very accurate and were used to compare the EBLUP estimates proposed earlier with a few rival estimates. In this example, the covariates consisted of an intercept and the adjusted census median income. Including the intercept term there are only two regression parameters (dimension of $\boldsymbol{\beta}$) in the model.

We compare using the CVs the CPS estimates, the univariate HB time series estimates (HB^1) of Ghosh et al. (1996) (see Section 3), and the EBLUP estimates of Datta et al. (2002) both for ML and REML estimates of the variance components $\boldsymbol{\psi}$. For normal linear model the EBLUP is identical to the EB estimator. The EB estimates emerge as the best estimates among all the rival estimates. Note that because the estimated MSE used to compute the CV of the EBLUP is second-order unbiased, the reported CVs do not underestimate the true CVs. Although the CPS estimates have CV more than 6% for 38 states, the (HB^1) estimates of Ghosh et al. (1996) have four states and EB estimates have zero state with CV more than 6%. The EB estimates produce CV between 2% and 4% for 49 states, whereas the (HB^1) estimates have 10 states and the CPS estimates have only six states in this category.

Datta et al. (2002) compared different estimators on the basis of average absolute relative deviation, average squared relative deviation, average absolute deviation, and average squared deviation described later. Let e_{iTR} denote the true median income for the ith state, and t_i is any estimate of $e_{i\mathrm{TR}}$, $i = 1, \ldots, 51$. Then

$$\mathrm{ARD} = \text{average relative deviation } = (51)^{-1} \sum_{i=1}^{51} |t_i - e_{i\mathrm{TR}}| e_{i\mathrm{TR}}^{-1},$$

$$\mathrm{ASRD} = \text{average squared relative deviation } = (51)^{-1} \sum_{i=1}^{51} (t_i - e_{i\mathrm{TR}})^2 e_{i\mathrm{TR}}^{-2},$$

$$\mathrm{AAD} = \text{average absolute deviation } = (51)^{-1} \sum_{i=1}^{51} |t_i - e_{i\mathrm{TR}}|,$$

$$\mathrm{ASD} = \text{average squared deviation } = (51)^{-1} \sum_{i=1}^{51} (t_i - e_{i\mathrm{TR}})^2.$$

These numbers were computed using the 1989 census median income estimates as the true parameters. Table 1 reports these four measures for CPS estimates, Census Bureau's (BOC) estimates, univariate time series HB estimates (HB^1) of Ghosh et al. (1996), and the EB estimates. It is clear that the EB estimates are better than the CPS and HB^1. The EBLUP estimates compare very well with the Census Bureau's estimates.

Table 1
A Comparison of estimates under four different criteria

Estimate	ARD	ASRD	AAD	ASD
CPS	0.0735	0.0084	2,928.82	13,811,122
Bureau	0.0296	0.0013	1,183.90	2,151,350
HB^1	0.0338	0.0018	1,351.67	3,095,736
EB(ML)	0.0278	0.0014	1,119.00	2,339,959
EB(REML)	0.0291	0.0014	1,125.70	2,368,397

The BOC estimates are composite estimates derived essentially as an EB procedure using an adhoc estimate of the model variance based on the univariate model. The composite estimates were obtained by constraining the EB estimates so that they do not deviate from CPS sample estimates by more than one standard deviation. So the BOC estimates are analogous to the "limited translation estimates" as discussed in Fay and Herriot (1979). It should be noted that the EB/EBLUP estimates based on ML estimates of the variance components performs better than the estimates based on REML estimates. The HB methods that combine both time series and cross-sectional data to produce estimates of small area means will be discussed in the next section. In particular, we review the HB methods of Ghosh et al. (1996) and Datta et al. (1999b) dealing with estimation of annual median income and monthly U.S. unemployment rates, both at the state level, respectively.

2.4. *Empirical Bayes small area estimation in GLMMs*

We have considered so far small area estimation problems only for continuous-valued response variable based on normal linear mixed models. However, sometimes the response variable is categorical or binary in nature. For example, in the SAIPE program the U.S. Census Bureau is interested in estimating the poverty rates among school children at the state and county levels. The response variable in this case is binary: it takes the value 1 if the child is in poverty, and it takes 0 otherwise. More generally, the response variable may take values in multiple categories. In the context of disease mapping the response is a discrete variable counting the number of occurrences of an event. Generalized linear models (GLMs) play an important role in analyzing this kind of data.

The EB approach has played an important role in developing small area estimates for binary and discrete data. Dempster and Tomberlin (1980) and MacGibbon and Tomberlin (1989) obtained small area estimates of proportions based on EB techniques. Dempster and Tomberlin (1980) used logistic regression to estimate proportion of census undercounts. They proceeded by finding first the Bayes predictor (under squared error loss) of the predictand assuming that all the fixed model parameters are known. Then the fixed model parameters are estimated from the marginal distribution of the data and are replaced by their estimates in the Bayes predictor. Rao (2003a, Chapters 5, 9, and 10) has an excellent account of small area estimation in GLMs.

To fix ideas suppose for a unit-level model the response variable Y_{ij} corresponding to the jth unit in the ith area is a binary variable taking values 0 or 1. Let $p_{ij} = P(Y_{ij} = 1)$. Then the finite population mean $\gamma_i = \sum_{j=1}^{N_i} Y_{ij}/N_i$ is equal to P_i, the proportion of units in the ith small area having the particular characteristic. If we let $\boldsymbol{\phi}$ denote all the model parameters, the Bayes predictor $\tilde{P}_i^{\mathrm{B}}(\boldsymbol{\phi})$ of P_i under squared error loss is the conditional expectation of the unsampled response given the observed data in the area, and is given by

$$\tilde{P}_i^B(\boldsymbol{\phi}) = \frac{\sum_{j=1}^{n_i} y_{ij} + \sum_{j=n_i+1}^{N_i} \tilde{p}_{ij}^B(\phi)}{N_i} = f_i \bar{y}_{is} + \frac{1}{N_i} \sum_{j=n_i+1}^{N_i} \tilde{p}_{ij}^B(\phi) \tag{30}$$

where, as before for simplicity of notation, we are denoting the sampled units from the ith area by $y_{i1}, \ldots, y_{in_i}$, and $\tilde{p}_{ij}^B(\phi) = E[Y_{ij}|\boldsymbol{y}_i, \boldsymbol{\phi}]$. Here $\bar{y}_{is}$ is the sample proportion

and $f_i = n_i/N_i$ is the sampling fraction. Associated with the Bayes predictor $\tilde{P}_i^{\mathrm{B}}(\boldsymbol{\phi})$, let $M_{1i}^{\mathrm{B}}(\boldsymbol{\phi})$ denote the MSE of P_i. This MSE is calculated based on the joint distribution of $Y_{i1}, \ldots, Y_{iN_i}$.

In the absence of any covariate, $Y_{i1}, \ldots, Y_{in_i}$ are i.i.d Bernoulli (p_i). Martuzzi and Elliott (1996) assumed an exchangeable prior for the p_i and derived a shrinkage estimator of p_i by minimizing the total squared loss. They estimated the mean and the variance parameter of the distribution of p_i's from the marginal distribution of the data. They applied their method to estimate prevalence of respiratory symptoms in school children in 71 small areas in Huddersfield, Northern England. An alternative to the aforementioned approach is given by assuming p_i coming from a common beta distribution with parameters a and b. Here the prior distribution is conjugate leading to a beta-binomial marginal distribution of $\sum_{j=1}^{n_i} Y_{ij}$. The parameters a and b of the prior distribution are estimated from the marginal distribution of the data. For various methods of estimation of the parameters, one may refer to Rao (2003a, Section 9.4).

In the presence of covariates, to obtain the EB predictor of P_i MacGibbon and Tomberlin (1989) suggested the following logistic regression model for p_{ij}

$$\operatorname{logit}(p_{ij}) = \boldsymbol{x}_{ij}^T\boldsymbol{\beta} + v_i, \tag{31}$$

where $v_i \overset{iid}{\sim} N(0, \sigma_v^2)$ with unit-level covariate vector $\boldsymbol{x}_{ij}$. Here $\boldsymbol{\phi} = (\boldsymbol{\beta}^T, \sigma_v^2)^T$.

Note that under the above mentioned model, unlike in normal linear mixed model, the Bayes predictor $\tilde{P}_i^{\mathrm{B}}(\boldsymbol{\phi})$ does not have a closed-form expression. In fact, as in Rao (2003a, p. 203) for the jth unsampled unit the Bayes predictor of Y_{ij} is given by $\tilde{p}_{ij}^{B}(\phi) = E[p_{ij}\exp(h_i)]/E[\exp(h_i)]$, where the expectations are with respect to a standard normal random variable Z with $h_i = \sum_{j=1}^{n_i} y_{ij}\boldsymbol{x}_{ij}^T\boldsymbol{\beta} + n_i\bar{y}_{is}\sigma_v Z - \sum_{j=1}^{n_i} log[1+\exp(\boldsymbol{x}_{ij}^T\boldsymbol{\beta}+\sigma_v Z)]$ and $p_{ij} = 1/\{1+\exp(-\boldsymbol{x}_{ij}^T\boldsymbol{\beta} - \sigma_v Z)\}$. Thus to evaluate the Bayes predictor we need to evaluate several one-dimensional integrals numerically. The same thing is true for the MSE $M_{1i}^{B}(\boldsymbol{\phi})$ of P_i.

To obtain the EB estimator of P_i it is necessary to estimate the model parameters $\boldsymbol{\phi}$. In the absence of covariates Jiang (1998) and Jiang and Zhang (2001) suggested MOM approach based on estimating equation. Because the marginal distribution of the data does not have a closed-form, the ML method is not straightforward. In this setup numerical quadrature and optimization, the EM algorithm and an MCMC algorithm, play important roles. An EB estimation of P_i in the presence of covariates was discussed by Farrell et al. (1997). See also Jiang and Lahiri (2006).

Because the response variable in many problems may not be binary but may be discrete or categorical in nature it is important to consider models appropriate for such data. To this goal, Ghosh et al. (1998) proposed the following GLMM. In their model specification, in the first step they assumed that Y_{ij}, $j = 1, \ldots, N_i$, $i = 1, \ldots, m$ are conditionally independent with pdf

$$f(y_{ij}|\theta_{ij}, w_{ij}) = \exp[w_{ij}^{-1}\{y_{ij}\theta_{ij} - \psi(\theta_{ij})\} + \rho(y_{ij}; w_{ij})]. \tag{32}$$

The above model is referred to as a generalized linear model (McCullagh and Nelder, 1989, p. 28). If the scale parameters w_{ij} are considered known, which is the case in Ghosh et al. (1998), the distribution in (32) is a one-parameter exponential family expressed in terms of the canonical parameter θ_{ij}. The above family of distributions includes normal, Poisson, Bernoulli, and gamma as special cases.

Ghosh et al. (1998) proposed a linear mixed model for $r(\theta_{ij})$ for some known strictly increasing link function $r(.)$. A special case of their model is given by

$$r(\theta_{ij}) = \boldsymbol{x}_{ij}^T\boldsymbol{\beta} + v_i, \quad j = 1, \ldots, \quad N_i, i = 1, \ldots, m, \tag{33}$$

where v_i's are iid $N(0, \sigma_v^2)$. Equations (32) and (33) together specify a GLMM.

Note that the model given by (30) and (31) is a special case of (32) and (33). The usefulness of GLMM with canonical link (that is, $r(x) = x$) cannot be overemphasized (see McCullagh and Nelder, 1989; McCulloch and Searle, 2001). In particular, the Poisson regression model has been quite popular in disease mapping.

For the GLMM, as in the case of binary response model, once again we may want EB predictor for the finite population γ_i. Again, for known $\boldsymbol{\phi}$ under squared error loss the Bayes predictor $\tilde{\gamma}_i^{\mathrm{B}}$ is

$$\tilde{\gamma}_i^{\mathrm{B}} = f_i\bar{y}_{is} + \frac{1}{N_i}\sum_{j=n_i+1}^{N_i}\tilde{Y}_{ij}(\boldsymbol{\phi}, \boldsymbol{y}_i) = k_i(\boldsymbol{y}_i, \boldsymbol{\phi}), \quad \text{say}, \tag{34}$$

where for $j = n_i + 1, \ldots, N_i$, $\tilde{Y}_{ij}(\boldsymbol{\phi}, \boldsymbol{y}_i) = E[Y_{ij}|\boldsymbol{\phi}, \boldsymbol{y}_i] = E[\psi'(\theta_{ij})|\boldsymbol{\phi}, \boldsymbol{y}_i]$, that is,

$$\tilde{Y}_{ij}(\boldsymbol{\phi}, \boldsymbol{y}_i) = \frac{E\{\psi'(r^{-1}(\boldsymbol{x}_{ij}^T\boldsymbol{\beta} + \sigma_v Z))t_i(\boldsymbol{\phi}, \boldsymbol{y}_i, Z)\}}{E\{t_i(\boldsymbol{\phi}, \boldsymbol{y}_i, Z)\}},$$

where $\psi'(x)$ denotes the derivative of $\psi(x)$ w.r.t. x, and the expectation $E[.]$ is with respect to the distribution of $Z \sim N(0, 1)$ and

$$t_i(\boldsymbol{\phi}, \boldsymbol{y}_i, Z) = \prod_{j=1}^{n_i}\exp[w_{ij}^{-1}\{y_{ij}r^{-1}(\boldsymbol{x}_{ij}^T\boldsymbol{\beta} + \sigma_v Z) - \psi(r^{-1}(\boldsymbol{x}_{ij}^T\boldsymbol{\beta} + \sigma_v Z))\}].$$

Usually, under GLMM the Bayes predictor $k_i(\boldsymbol{y}_i, \boldsymbol{\phi})$ of γ_i, outside the normality setup, has no closed-form expression. Not surprisingly, in this scenario, the MSE of the predictor, which we denote by $M_{1i}(\boldsymbol{\phi}, \boldsymbol{y}_i)$, has no closed-form expression either.

2.4.1. Estimation of MSE of the EB predictor

An EB predictor of γ_i, denoted by $\hat{\gamma}_i^{\mathrm{EB}} = k_i(\boldsymbol{y}_i, \hat{\boldsymbol{\phi}})$, is obtained by replacing $\boldsymbol{\phi}$ by its estimate $\hat{\boldsymbol{\phi}}$ in the Bayes predictor $\tilde{\gamma}_i^{\mathrm{B}}$. The marginal distribution of the data is used to get the estimate $\hat{\boldsymbol{\phi}}$. A naive estimate of the MSE of the EB predictor is given by $M_{1i}(\hat{\boldsymbol{\phi}}, \boldsymbol{y}_i)$. As in the case of the normal linear mixed models, the error in this approximation to the estimate of MSE is of the order $O(m^{-1})$. In general, it is quite challenging as in Section 2.2.3 to derive a second-order accurate approximation to the MSE of the EB predictor by ignoring all $o(m^{-1})$ terms. Similarly, a second-order unbiased estimate of the MSE is also complicated. The MSE of the EB predictor $\hat{\gamma}_i^{\mathrm{EB}}$ can be decomposed as

$$\begin{aligned}\mathrm{MSE}(\hat{\gamma}_i^{\mathrm{EB}}) &= E[\hat{\gamma}_i^{\mathrm{EB}} - \gamma_i]^2 = E[M_{1i}(\hat{\boldsymbol{\phi}}, \boldsymbol{Y}_i)] + E[\hat{\gamma}_i^{\mathrm{EB}} - \tilde{\gamma}_i^{\mathrm{B}}]^2 \\ &= g_{1i}(\boldsymbol{\phi}) + g_{2i}(\boldsymbol{\phi}), \quad \text{say}.\end{aligned} \tag{35}$$

The jackknife method of Jiang et al. (2002), which is computer intensive, can be applied to estimate the MSE. Denoting by $\hat{\boldsymbol{\phi}}_l$, an estimator of $\boldsymbol{\phi}$ obtained by excluding the lth area in its computation, let $\hat{\gamma}_{i,-l}^{\mathrm{EB}} = k_i(\boldsymbol{Y}_i, \hat{\boldsymbol{\phi}}_l)$. A reasonable estimator of $g_{2i}(\boldsymbol{\phi})$

is given by $\hat{g}_{2i} = \frac{m-1}{m}\sum_{l=1}^{m}\left(\hat{\gamma}_{i,-l}^{\mathrm{EB}} - \hat{\gamma}_i^{\mathrm{EB}}\right)^2$. Further, an estimator of g_{1i} is given by $\hat{g}_{1i} = M_{1i}(\hat{\boldsymbol{\phi}}, \boldsymbol{Y}_i) - \frac{m-1}{m}\sum_{l=1}^{m}[M_{1i}(\hat{\boldsymbol{\phi}}_{-l}, \boldsymbol{Y}_i) - M_{1i}(\hat{\boldsymbol{\phi}}, \boldsymbol{Y}_i)]$. Putting these two estimators together a jackknife estimator of the MSE is given by $\mathrm{mse}_J(\hat{\gamma}_i^{\mathrm{EB}}) = \hat{g}_{1i} + \hat{g}_{2i}$. Jiang (1998) and Jiang and Zhang (2001) obtained jackknife estimate of mse of the EB predictor of binary response probability without any covariate. They estimated the parameter $\boldsymbol{\phi}$ by estimating equation. For the binary response model without covariate, using Taylor's expansion, Jiang and Lahiri (2001) obtained a second-order accurate mse estimator of the EB predictor. Ghosh and Maiti (2004) obtained EB predictors of small area means for the general natural exponential family with quadratic variance function (NEF-QVF) family of distributions based on the theory of optimal estimating functions. For a lucid presentation on the estimation of the MSE of the EB predictor of small area mean for binary response data readers may refer to Rao (2003a, Section 9.4). One may also refer to Jiang and Lahiri (2006) for a recent review of the work in this area. For a general discussion on estimation of the parameters in generalized linear mixed models the readers may refer to McCulloch and Searle (2001, Chapter 10).

3. Bayesian approach to small area estimation

The Bayesian inference is based on the predictive (more specifically, posterior predictive) distribution of the unobserved y values given in the observed data. We first provide a Bayesian interpretation of the composite estimator in (1). Corresponding to a positive-valued covariate X, we consider the superpopulation model given by $Y_{ij} \overset{ind}{\sim} N(bx_{ij}, \sigma^2 x_{ij})$, $j = 1, \ldots, N_i, i = 1, \ldots, m$. Under squared error loss the Bayes estimator of γ_i is given by (1). Also, if we use an improper uniform prior for b over $(-\infty, \infty)$, the Bayes estimator of b is given by $\bar{y}_s/\bar{x}_s$, the overall sample mean. The above simple model provides a Bayesian model-based interpretation of the synthetic estimator $\hat{\gamma}_i^C$ in Section 2.1. In the discussion later, various realistic useful Bayesian models are presented for model-based estimation of small area means.

3.1. Hierarchical Bayes small area estimation for unit-level data: Univariate case

We start with the hierarchical Bayesian (HB) model of Datta and Ghosh (1991) that they put forward for unit-level data for all the units in the population. (a) Conditional on $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$, $\boldsymbol{v}, \boldsymbol{\phi} = (\phi_1, \ldots, \phi_t)^T$, and τ, let $\boldsymbol{Y} \sim N(\boldsymbol{X\beta} + \boldsymbol{Zv}, \tau^{-1}\boldsymbol{\Omega})$, (b) conditional on $\boldsymbol{\phi}$ and τ, let $\boldsymbol{v} \sim N(\boldsymbol{0}, \tau^{-1}\boldsymbol{\Delta}(\boldsymbol{\phi}))$, and (c) $\boldsymbol{\beta}$, τ and $\boldsymbol{\phi}$ have a joint prior distribution, proper or improper.

Stages (a) and (b) of the above HB model can be written as a linear mixed model

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{Zv} + \boldsymbol{e}, \tag{36}$$

where $\boldsymbol{e}(N \times 1)$ and $\boldsymbol{v}(b \times 1)$ are mutually independent, with $\boldsymbol{e} \sim N(\boldsymbol{0}, \tau^{-1}\boldsymbol{\Omega})$ and $\boldsymbol{v} \sim N(\boldsymbol{0}, \tau^{-1}\boldsymbol{\Delta}(\boldsymbol{\phi}))$. The matrices $\boldsymbol{X}(N \times p)$ and $\boldsymbol{Z}(N \times b)$ are known design matrices. The matrix $\boldsymbol{\Omega}$ is a known positive definite (p.d.) matrix, and $\boldsymbol{\Delta}(\boldsymbol{\phi})$ is a $q \times q$ p.d. matrix that is structurally known except possibly for some unknown $\boldsymbol{\phi}$. In the above model τ and $\boldsymbol{\phi}$ are related to $t + 1$ variance components denoted by $\sigma_0^2, \sigma_1^2, \ldots, \sigma_t^2$ of the linear mixed model. To simplify representation of the HB model, we denote the

inverses of these variance components by $\tau, \tau\phi_1, \ldots, \tau\phi_t$, respectively. Indeed, in the examples to follow, $\boldsymbol{\phi}$ involves the ratios of the variance components. For simplicity sometimes we will denote $\boldsymbol{\Delta}(\boldsymbol{\phi})$ simply by $\boldsymbol{\Delta}$. In Bayesian approach since all the parameters, be it $\boldsymbol{\beta}$, $\boldsymbol{v}$, τ, or $\boldsymbol{\phi}$, are random, it makes little sense calling (36) a mixed effects model.

We now partition $\boldsymbol{Y}$, $\boldsymbol{X}$, $\boldsymbol{Z}$, and $\boldsymbol{e}$ corresponding to the sampled and nonsampled units. Note that for the sampled units the linear model in (36) corresponds to the general linear mixed model given by (3). We write $\boldsymbol{Y}^{(1)} = \text{col}_{1\le i\le m}\boldsymbol{Y}_i^{(1)}$, where $\boldsymbol{Y}_i^{(1)}$ is the $n_i \times 1$ vector corresponding to the sampled units from the ith small area. Similarly, $\boldsymbol{Y}^{(2)} = \text{col}_{1\le i\le m}\boldsymbol{Y}_i^{(2)}$, where $\boldsymbol{Y}_i^{(2)}$ is the $(N_i - n_i) \times 1$ vector corresponding to the nonsampled units from the ith small area. The Bayesian inference is carried out via the predictive distribution of $\boldsymbol{Y}^{(2)}$ given $\boldsymbol{Y}^{(1)} = \boldsymbol{y}^{(1)}$.

From our discussion in Section 2.3 and Eq. (36) it is immediate that the HB model given here provides an HB model for the nested error regression model and the random regression coefficients model. Datta and Ghosh (1991) have shown that certain cross-classification models, two-stage and multi-stage sampling models with covariates are some of the other important special cases of this HB model.

In our discussion later, we need the density of a multivariate t-distribution. A random vector $\boldsymbol{W}$ is said to have a p-variate t-distribution with location parameter $\boldsymbol{\beta}$, scale metrix $\boldsymbol{\Omega}$, and degrees of freedom ν if the density $f(\boldsymbol{w})$ is given by $f(\boldsymbol{w}) \propto [\nu + (\boldsymbol{w} - \boldsymbol{\beta})^T\boldsymbol{\Omega}^{-1}(\boldsymbol{w}-\boldsymbol{\beta})]^{-(\nu+p)/2}$. To derive the required predictive distribution we will complete the hierarchical model by specifying the prior distribution in stage (c).

(c) We assign independent priors on $\boldsymbol{\beta}$, τ, $\tau\phi_1, \ldots, \tau\phi_t$ with an improper uniform(R^p) prior for $\boldsymbol{\beta}$, and $\tau \sim$ gamma$(a_0/2, g_0/2)$, $\tau\phi_k \sim$ gamma$(a_k/2, g_k/2)$, $k = 1, \ldots, t$, with $a_0 \ge 0$, $g_0 \ge 0$, $a_k > 0$, $g_k \ge 0$, $k = 1, \ldots, t$.

In this discussion, we use the notation gamma(a, b) to denote a gamma density proportional to $\exp(-ax)x^{b-1}$. Allowing $a_0 = 0$ and some of the g_k's 0, some improper gamma distributions are included as priors. In principle, from the prior distribution of $\tau, \tau\phi_1, \ldots, \tau\phi_t$ through transformation it is possible to obtain the joint prior distribution of $\tau, \phi_1, \ldots, \phi_t$, and hence, express the joint prior of $\tau, \phi_1, \ldots, \phi_t$ hierarchically using the conditional distribution of τ given $\phi_1, \ldots, \phi_t$ and the (joint) marginal distribution of $\phi_1, \ldots, \phi_t$. In the latter representation, the parameters $\phi_1, \ldots, \phi_t$ are called hyperparameters, and their prior distribution a hyperprior.

Let $\boldsymbol{\Upsilon} \equiv \boldsymbol{\Upsilon}(\boldsymbol{\phi}) = \boldsymbol{\Omega} + \boldsymbol{Z}\boldsymbol{\Delta}(\boldsymbol{\phi})\boldsymbol{Z}^T$ and partition $\boldsymbol{\Upsilon}$ as $((\boldsymbol{\Upsilon}_{ij}))_{i,j=1,2}$ corresponding to the sampled and nonsampled units. In the following matrices, we suppress that they are functions of $\boldsymbol{\phi}$. Define $\boldsymbol{\Upsilon}_{22.1} = \boldsymbol{\Upsilon}_{22} - \boldsymbol{\Upsilon}_{21}\boldsymbol{\Upsilon}_{11}^{-1}\boldsymbol{\Upsilon}_{12}$, and

$$\boldsymbol{K} = \boldsymbol{\Upsilon}_{11}^{-1} - \boldsymbol{\Upsilon}_{11}^{-1}\boldsymbol{X}^{(1)}\left(\boldsymbol{X}^{(1)T}\boldsymbol{\Upsilon}_{11}^{-1}\boldsymbol{X}^{(1)}\right)^{-1}\boldsymbol{X}^{(1)T}\boldsymbol{\Upsilon}_{11}^{-1}, \tag{37}$$

$$\boldsymbol{M} = \boldsymbol{\Upsilon}_{21}\boldsymbol{K} + \boldsymbol{X}^{(2)}\left(\boldsymbol{X}^{(1)T}\boldsymbol{\Upsilon}_{11}^{-1}\boldsymbol{X}^{(1)}\right)^{-1}\boldsymbol{X}^{(1)T}\boldsymbol{\Upsilon}_{11}^{-1}, \tag{38}$$

$$\boldsymbol{P}^T = \left[\boldsymbol{\Upsilon}_{11}^{-1}\boldsymbol{X}^{(1)}\left(\boldsymbol{X}^{(1)T}\boldsymbol{\Upsilon}_{11}^{-1}\boldsymbol{X}^{(1)}\right)^{-1},\ \boldsymbol{K}\boldsymbol{Z}^{(1)}\boldsymbol{\Delta}\right], \tag{39}$$

$$\boldsymbol{G} = \boldsymbol{\Upsilon}_{22.1} + \left(\boldsymbol{X}^{(2)} - \boldsymbol{\Upsilon}_{21}\boldsymbol{\Upsilon}_{11}^{-1}\boldsymbol{X}^{(1)}\right)\left(\boldsymbol{X}^{(1)T}\boldsymbol{\Upsilon}_{11}^{-1}\boldsymbol{X}^{(1)}\right)^{-1}\left(\boldsymbol{X}^{(2)} - \boldsymbol{\Upsilon}_{21}\boldsymbol{\Upsilon}_{11}^{-1}\boldsymbol{X}^{(1)}\right)^T \tag{40}$$

and $\boldsymbol{T} = ((\boldsymbol{T}_{ij}))_{i,j=1,2}$, where $\boldsymbol{T}_{11} = (\boldsymbol{X}^{(1)T}\boldsymbol{\Upsilon}_{11}^{-1}\boldsymbol{X}^{(1)})^{-1}$, $\boldsymbol{T}_{12} = -\boldsymbol{T}_{11}\boldsymbol{X}^{(1)T}\boldsymbol{\Upsilon}_{11}^{-1}\boldsymbol{Z}^{(1)}$ $\boldsymbol{\Delta} = \boldsymbol{T}_{21}^T$, and $\boldsymbol{T}_{22} = \boldsymbol{\Delta} - \boldsymbol{\Delta}\boldsymbol{Z}^{(1)T}\boldsymbol{K}\boldsymbol{Z}^{(1)}\boldsymbol{\Delta}$.

In the theorem later, we provide the posterior distribution of $\boldsymbol{Y}^{(2)}$ and that of $(\boldsymbol{\beta}^T, \boldsymbol{v}^T)^T$ in two steps. Although the conditional distribution of $\boldsymbol{Y}^{(2)}$ given $\boldsymbol{y}^{(1)}$ and $\boldsymbol{\phi}$ has a closed form, the posterior distribution of $\boldsymbol{Y}^{(2)}$ (integrating out $\boldsymbol{\phi}$) does not have a closed form expression. A derivation of the posterior distribution of $\boldsymbol{Y}^{(2)}$ is given in Datta and Ghosh (1991), and that of $(\boldsymbol{\beta}^T, \boldsymbol{v}^T)^T$ is given in Datta (1992).

THEOREM 1. *Consider the model given in (36) and the prior distribution specified in stage (c). Assume that* $\nu^* = n + \sum_{k=0}^{t} g_k - p > 2$. *Then,*

(i) conditional on $\boldsymbol{\phi}$ *and* $\boldsymbol{y}^{(1)}$, *the distribution of* $\boldsymbol{Y}^{(2)}$ *is multivariate-t with degrees of freedom* ν^*, *location parameter* $\boldsymbol{M}\boldsymbol{y}^{(1)}$, *and scale matrix* $(n+\sum_{k=0}^{t} g_k - p)^{-1}[a_0 + \sum_{k=1}^{t} a_k\phi_k + \boldsymbol{y}^{(1)T}\boldsymbol{K}\boldsymbol{y}^{(1)}]\boldsymbol{G}$;

(ii) conditional on $\boldsymbol{\phi}$ *and* $\boldsymbol{y}^{(1)}$, *the distribution of* $(\boldsymbol{\beta}^T, \boldsymbol{v}^T)^T$ *is multivariate-t with degrees of freedom* ν^*, *location parameter* $\boldsymbol{P}\boldsymbol{y}^{(1)}$, *and scale matrix* $(n + \sum_{k=0}^{t} g_k - p)^{-1}[a_0 + \sum_{k=1}^{t} a_k\phi_k + \boldsymbol{y}^{(1)T}\boldsymbol{K}\boldsymbol{y}^{(1)}]\boldsymbol{T}$;

(iii) the unnormalized posterior density $f(\boldsymbol{\phi}|\boldsymbol{y}^{(1)})$ *of* $\boldsymbol{\phi}$ *is proportional to*

$$|\boldsymbol{\Upsilon}_{11}|^{-1/2}|\boldsymbol{T}_{11}|^{-1/2}\left[\prod_{k=1}^{t}\phi_k^{g_k/2-1}\right]\left[a_0 + \sum_{k=1}^{t} a_k\phi_k + \boldsymbol{y}^{(1)T}\boldsymbol{K}\boldsymbol{y}^{(1)}\right]^{-\nu^*/2}. \tag{41}$$

Using the first two moments of a mulitivariate-t distribution, we get from (i) that $E[\boldsymbol{Y}^{(2)}|\boldsymbol{\phi}, \boldsymbol{y}^{(1)}] = \boldsymbol{M}\boldsymbol{y}^{(1)}$ and $V[\boldsymbol{Y}^{(2)}|\boldsymbol{\phi}, \boldsymbol{y}^{(1)}] = \left\{a_0 + \sum_{k=1}^{t} a_k\phi_k + \boldsymbol{y}^{(1)T}\boldsymbol{K}\boldsymbol{y}^{(1)}\right\}\boldsymbol{G}/\nu^*$. Now, by iterated expectation and variance it follows that

$$E\left[\boldsymbol{Y}^{(2)}|\boldsymbol{y}^{(1)}\right] = E\left[E\left\{\boldsymbol{Y}^{(2)}|\boldsymbol{\phi}, \boldsymbol{y}^{(1)}\right\}|\boldsymbol{y}^{(1)}\right] = E\left[\boldsymbol{M}\boldsymbol{y}^{(1)}|\boldsymbol{y}^{(1)}\right], \tag{42}$$

$$\begin{aligned} V\left[\boldsymbol{Y}^{(2)}|\boldsymbol{y}^{(1)}\right] &= V\left[E\left\{\boldsymbol{Y}^{(2)}|\boldsymbol{\phi}, \boldsymbol{y}^{(1)}\right\}|\boldsymbol{y}^{(1)}\right] + E\left[V\left\{\boldsymbol{Y}^{(2)}|\boldsymbol{\phi}, \boldsymbol{y}^{(1)}\right\}|\boldsymbol{y}^{(1)}\right] \\ &= V\left[\boldsymbol{M}\boldsymbol{y}^{(1)}|\boldsymbol{y}^{(1)}\right] + \frac{E\left[\left\{a_0 + \sum_{k=1}^{t} a_k\phi_k + \boldsymbol{y}^{(1)T}\boldsymbol{K}\boldsymbol{y}^{(1)}\right\}\boldsymbol{G}|\boldsymbol{y}^{(1)}\right]}{\nu^*}. \end{aligned} \tag{43}$$

Conditional on $\boldsymbol{Y}^{(1)} = \boldsymbol{y}^{(1)}$ for suitable known vectors $\boldsymbol{a}_i(n \times 1)$ and $\boldsymbol{b}_i((N-n) \times 1)$ the ith small area finite population mean γ_i can be written as $\gamma_i = \boldsymbol{a}_i^T\boldsymbol{y}^{(1)} + \boldsymbol{b}_i^T\boldsymbol{Y}^{(2)}$. Using (42) and (43) we can easily obtain the posterior mean and the posterior variance of γ_i. For known hyperparameter $\boldsymbol{\phi}$ it follows from the first part of Theorem 3.1 that the Bayes predictor of γ_i, denoted by $\hat{\gamma}_i^{\mathrm{B}}$, under quadratic loss is given by $(\boldsymbol{a}_i + \boldsymbol{M}^T\boldsymbol{b}_i)^T\boldsymbol{y}^{(1)}$. The posterior variance associated with the Bayes predictor of γ_i is $g_i^{\mathrm{B}}(\boldsymbol{\phi})$, where

$$g_i^{\mathrm{B}}(\boldsymbol{\phi}) = \left(n + \sum_{k=0}^{t} g_k - p - 2\right)^{-1}\left\{a_0 + \sum_{k=1}^{t} a_k\phi_k + \boldsymbol{y}^{(1)T}\boldsymbol{K}\boldsymbol{y}^{(1)}\right\}\boldsymbol{b}_i^T\boldsymbol{G}\boldsymbol{b}_i.$$

Note that both the Bayes predictor $\hat{\gamma}_i^{\mathrm{B}}$ and the posterior variance $g_i^{\mathrm{B}}(\boldsymbol{\phi})$ typically involve $\boldsymbol{\phi}$. Datta and Ghosh (1991) showed that for known $\boldsymbol{\phi}$ the Bayes predictor $\hat{\gamma}_i^{\mathrm{B}}$ is also the BLUP of γ_i. The HB predictor of γ_i, denoted by $\hat{\gamma}_i^{\mathrm{HB}}$, is obtained by integrating $\hat{\gamma}_i^{B}$ with respect to the posterior distribution of the hyperparameter $\boldsymbol{\phi}$. In an EB approach

instead of assigning a prior distribution to the hyperparameter $\boldsymbol{\phi}$, it is estimated by some estimator $\hat{\boldsymbol{\phi}}$ using the marginal distribution of the data. The EB predictor is obtained by replacing $\boldsymbol{\phi}$ by $\hat{\boldsymbol{\phi}}$ in $\hat{\gamma}_i^{\mathrm{B}}$. If we denote the EB predictor of γ_i by $\hat{\gamma}_i^{\mathrm{EB}}$, then $\hat{\gamma}_i^{\mathrm{EB}} = \hat{\gamma}_i^{B}(\hat{\boldsymbol{\phi}})$. Because the Bayes predictor $\hat{\gamma}_i^{\mathrm{B}}$ is also the BLUP, it immediately follows that the EB predictor $\hat{\gamma}_i^{\mathrm{EB}}$ is also an EBLUP of γ_i.

If the HB variance of γ_i is V_i^{HB}, then it follows from (42) and (43) that

$$V_i^{\mathrm{HB}} = E\left[g_i^{\mathrm{B}}(\boldsymbol{\phi})|\boldsymbol{y}^{(1)}\right] + V\left(\hat{\gamma}_i^{\mathrm{B}}|\boldsymbol{y}^{(1)}\right). \tag{44}$$

The second term in the right-hand side of the above mentioned expression measures the contribution of unknown $\boldsymbol{\phi}$ to the posterior variance. Its relative magnitude with respect to the first component depends both on the variability of the Bayes predictor $\hat{\gamma}_i^{\mathrm{B}}$ as a function of $\boldsymbol{\phi}$ and on the spread of the posterior density $f(\boldsymbol{\phi}|\boldsymbol{y}^{(1)})$. Although sometimes the contribution of this term to the total posterior variance is negligible, it is not always so. Associated with the EB predictor $\hat{\gamma}_i^{\mathrm{EB}}$ a naive measure of uncertainty is given by $g_i^{\mathrm{B}}(\hat{\boldsymbol{\phi}})$. This measure in comparison with the HB variance V_i^{HB} usually overstates the accuracy because it effectively ignores the second term of (44). Underestimation of the measure of uncertainty can be nonnegligible, if the the second term on the right side of (44) is large. Although there is usually not much difference between the HB predictor and the EB predictor, the naive measure of uncertainty of the EB predictor typically underestimates the true variance.

From the mixed effects model (36) it is immediate that γ_i can be expressed as a linear combination of $\boldsymbol{\beta}$, $\boldsymbol{v}$, and components of the sampling error vector $\boldsymbol{e}$. In predicting γ_i, if the finite population size N_i of the ith area is large, the linear combination of the sampling error vector $\boldsymbol{e}$ is ignored. In that case, the problem of HB prediction of γ_i boils down to the prediction of a linear combination $\boldsymbol{\beta}$ and $\boldsymbol{v}$. The HB solution to this problem follows as before using parts (ii) and (iii) of the theorem mentioned earlier.

Because the finite population mean γ_i for the ith small area involves a linear function of $\boldsymbol{Y}^{(2)}$, using (42) and (44) we can compute the posterior mean and posterior variance of γ_i. It is evident from (41) that the posterior distribution of $\boldsymbol{\phi}$ is fairly complicated and to obtain the posterior moments of γ_i, one needs to perform numerical integration. If the dimension of $\boldsymbol{\phi}$ is small, the numerical integration can be performed via quadrature formula. Otherwise, it is more convenient to use Monte Carlo methods such as importance sampling (Berger, 1985, Chapter 4) or Gibbs sampling (Gelfand and Smith, 1990). For extensive discussion on Gibbs sampling one may refer to Gelman et al. (2004), Carlin and Louis (2000), and Rao (2003a, Chapter 10).

Datta and Ghosh (1991) applied the HB method discussed earlier to analyze several datasets. In particular, they had analyzed the Iowa crop area data of Battese et al. (1988) and another data appropriate for two-stage sampling model in small area estimation. Details are omitted to save space and may be found in Datta and Ghosh (1991). The HB method for area-level data using the Fay–Herriot model is presented in Section 3.2. The reader may also refer to Chapter 10 of Rao (2003a) for some other examples.

3.2. Bayesian multivariate approach to small area estimation

A Bayesian multivariate approach to small area estimation is available both for unit-level data and area-level data. First we consider area-level models. Multivariate HB model for

unit-level data is considered later in this section. Following Fay (1987), Datta et al. (1991, 1996) considered the HB model for the multivariate version of Fay–Herriot model for area-level data. Authors of these papers considered Bayesian estimation of state median income of four-person families for 50 states of the United States and Washington, DC. The U.S. Department of Health and Human Services (HHS) needs estimates of four-person family state median income to implement an energy assistance program to low-income families. The Bureau of the Census (BOC) has provided such estimates for nearly 30 years. Initially, the BOC has used a multiple linear regression by regressing less accurate estimates of median income data from the CPS on auxiliary information on per capita income from the Bureau of Economic Analysis (BEA) and census median income from the last census to produce the estimates for the HHS. Later on the BOC replaced this linear regression methodology by a more sophisticated EB methodology using the Fay–Herriot (1979) model. Because data on three-person and five-person families income by the states are also collected in the CPS, and these variables are strongly correlated with the main variable of interest. Fay (1987) suggested a multivariate modification of the Fay–Herriot EB approach to exploit this strong correlation to generate more accurate estimates of four-person family median income.

Using Fay's (1987) idea described earlier, Datta et al. (1991, 1996) considered both HB and EB analyses for this problem using a multivariate version of Fay–Herriot (1979) models based on area-level data. Fay (1987) considered only point estimates. Datta et al. (1991, 1996) considered both point estimates and measures of accuracy of the estimates. Let $\boldsymbol{\mu}_i$ denote an s-component vector of interest for the ith small area (here state) with the corresponding sample estimate $\boldsymbol{Y}_i$, $i = 1, \ldots, m$, the direct estimate based on the ith small area sample. As in the univariate Fay–Herriot model, for the multivariate model we assume that the sampling variance–covariance matrix associated with $\boldsymbol{Y}_i$ is known and given by $\boldsymbol{D}_i$. Datta et al. (1996) considered the following HB model:

(i) $\boldsymbol{Y}_i|\boldsymbol{\mu}_i \overset{ind}{\sim} N(\boldsymbol{\mu}_i, \boldsymbol{D}_i)$, $i = 1, \ldots, m$;

(ii) $\boldsymbol{\mu}_i|\boldsymbol{\beta}, \boldsymbol{A} = \boldsymbol{a} \overset{ind}{\sim} N(\boldsymbol{X}_i\boldsymbol{\beta}, \boldsymbol{a})$, $i = 1, \ldots, m$;

(iii) Marginally $\boldsymbol{\beta}$ and $\boldsymbol{A}$ are independent apriori with $\pi(\boldsymbol{\beta}, \boldsymbol{a}) \propto 1$.

For the median income estimation problem Datta et al. (1996) considered both bivariate and trivariate hierarchical linear models. In the trivariate case, the basic data is $\boldsymbol{Y}_i = (Y_{i1}, Y_{i2}, Y_{i3})^T$, $i = 1, \ldots, 51$, where Y_{i1}, Y_{i2}, and Y_{i3} are the sample median incomes of four-, three-, and five-person families in state i. The true median corresponding to Y_{ij} is μ_{ij} and $\boldsymbol{\mu}_i = (\mu_{i1}, \mu_{i2}, \mu_{i3})^T$. We also write $\boldsymbol{Y} = (\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_{51})^T$, $\boldsymbol{\mu} = (\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_{51})^T$. Analogously, in the bivariate case using any of three- or five-person or any linear combination of these two with the four-person family, $\boldsymbol{Y}_i$ and $\boldsymbol{\mu}_i$ are two-component vectors.

Let $\boldsymbol{x}_{i1} = (1, x_{i11}, x_{i12})^T$, $\boldsymbol{x}_{i2} = (1, x_{i21}, x_{i22})^T$, $\boldsymbol{x}_{i3} = (1, x_{i31}, x_{i32})^T$, $\boldsymbol{X}_i^{(b)} = (\boldsymbol{x}_{i1}, \boldsymbol{x}_{i2})^T$, and $\boldsymbol{X}_i^{(t)} = (\boldsymbol{x}_{i1}, \boldsymbol{x}_{i2}, \boldsymbol{x}_{i3})^T$, where x_{ijk} is the value of the covariate x_{jk} corresponding to the ith state, $j = 1, 2, 3$, $k = 1, 2$. Here $x_{j1} = \frac{\text{BEA PCI}(c)}{\text{BEA PCI}(b)} \times x_{j2}$, and x_{j2} corresponds to the census median income for jth family size household for the base year b from the most recently available decennial census, $j = 1$ corresponds to three-person household, $j = 2$ corresponds to four-person household, and $j = 3$ corresponds to five-person household. The variables BEA PCI(b) and BEA PCI(c) correspond respectively to BEA per capita income for the base and current years. For the trivariate case, in the

above HB model the design matrix $\boldsymbol{X}_i = \boldsymbol{I}_3 \otimes \boldsymbol{X}_i^{(t)}$. In this case, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_9)^T$ is the vector of regression coefficients. In the bivariate case, $\boldsymbol{X}_i = \boldsymbol{I}_2 \otimes \boldsymbol{X}_i^{(b)}$ and $\boldsymbol{\beta}$ is a six-component vector. In the univariate case, $\boldsymbol{X}_i = \boldsymbol{x}_{i2}^T$ and $\boldsymbol{\beta}$ is a three-component vector.

For the above HB model the Bayes estimates, given by the posterior means, and measures of accuracy of the estimates given by the posterior variances are computed using Gibbs sampling. Datta et al. (1996) obtained estimates of the median incomes for four-person families in 1979 by states, using the 1969 census figures as the base-year figures. We report here three types of HB estimates, namely, HB^1, HB^2, and HB^3, based on univariate, bivariate, and trivariate HB analyses. They also computed the EB^1 estimates, the EB estimates of four-person family based on the univariate, that is, the standard Fay–Herriot model. The estimates based on the univariate setup do not utilize information corresponding to three- and five-person families. Various estimates have been compared by Datta et al. (1996) against the 1980 census figures (corresponding to 1979 income based on a much bigger sample), treating the census figures as the "truth." To perform the comparison, Datta et al. (1996) used the four deviation crietria introduced in Section 2.3.

In addition to the four estimates of four-person family median income given earlier, Datta et al. (1996) also included the sample estimates (direct survey estimates from the CPS data) and the estimates provided by the BOC to the HHS in the comparison. These results are given in Table 2. Under all four criteria, the sample medians from the CPS have the worst performance. Table 2 very clearly demonstrates that any EB or HB procedure that allows borrowing strength from other small areas has a distinct advantage over the direct CPS estimates as well as the BOC estimates. However, among the competing EB and HB procedures there is very little advantage of using one in preference to the other.

Datta et al. (1996) did not report the standard errors associated with EB and HB estimates. However, they discussed that a naive measure of uncertainty that is often used for EB procedure grossly underestimates the standard errors because it fails to incorporate the uncertainty due to estimation of the model variance parameter. The HB methods do not suffer from this drawback. At the same time the HB procedures introduce significant reduction in standard errors when compared to the sample estimates. Also, the multivariate HB models seem to have a distinct advantage over the univariate HB model.

For unit-level data Datta et al. (1998) considered HB analysis using the following multivariate extension of the nested error regression model. In the setup of Section 2.2.5, the authors considered HB prediction of $\boldsymbol{\gamma}_i = N_i^{-1} \sum_{j=1}^{N_i} \boldsymbol{Y}_{ij}$ for $i = 1, \ldots, m$ where $\boldsymbol{Y}_{ij}$

Table 2
Avg. relative deviation, avg. squared relative deviation, avg. absolute deviation, and avg. squared deviation of the estimates

Estimates	BOC	CPS	EB^1	HB^1	HB^2	HB^3
$100 \times$ ARD	3.246	4.984	2.042	2.074	2.044	2.022
$1000 \times$ ASRD	1.65	3.40	0.68	0.69	0.68	0.67
AAD	722.84	1090.41	450.63	458.73	452.47	447.35
ASD	835,710	1,631,203	334,231	346,085	341,070	336,966

is an $s \times 1$ vector of response variables. The multivariate HB model, denoted below by MHB, is specified by defining the prior distribution of the model parameters in Model M given in Section 2.2.5.

Model MHB:

(I) Conditional on $\boldsymbol{B}$, $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_m$, $\boldsymbol{\Sigma}_e$, and $\boldsymbol{\Sigma}_v$, $\boldsymbol{Y}_{ij} \sim N_s(\boldsymbol{B}\boldsymbol{c}_{ij} + \boldsymbol{v}_i, \boldsymbol{\Sigma}_e)$ independently for $j = 1, \ldots, N_i$, $i = 1, \ldots, m$;

(II) Conditional on $\boldsymbol{B}$, $\boldsymbol{\Sigma}_v$, and $\boldsymbol{\Sigma}_e$, $\boldsymbol{v}_i \sim N_s(\boldsymbol{0}, \boldsymbol{\Sigma}_v)$ independently for $i = 1, \ldots, m$;

(III) Marginally, $\boldsymbol{B}$, $\boldsymbol{\Sigma}_v$, and $\boldsymbol{\Sigma}_e$ are independently distributed with a uniform prior on $\boldsymbol{B}$, an inverse Wishart prior $W_a^{-1}(\boldsymbol{\Phi}_v)$ for $\boldsymbol{\Sigma}_v$, and an inverse Wishart prior $W_a^{-1}(\boldsymbol{\Phi}_e)$ for $\boldsymbol{\Sigma}_e$, where $W_a^{-1}(.)$ denotes an inverse Wishart distribution with a degrees of freedom, $\boldsymbol{\Phi}_v$ and $\boldsymbol{\Phi}_e$ are known scales of the respective inverse Wishart distribution.

A random p.d. matrix $\boldsymbol{T}(s \times s)$ has an inverse Wishart distribution with p.d. scale matrix $\boldsymbol{\Delta}(s \times s)$ and degrees of freedom d if it has a density of the form

$$p(\boldsymbol{T}) \propto |\boldsymbol{T}|^{-(d+s+1)/2} \exp\left\{-\frac{1}{2} tr(\boldsymbol{T}^{-1}\boldsymbol{\Delta})\right\}, \; d \geq s.$$

We denote it by $\boldsymbol{T} \sim W_d^{-1}(\boldsymbol{\Delta})$. Note that an inverse Wishart is a conjugate prior distribution for a multivariate normal covariance matrix.

Using the notation and observation of Section 2.2.5, $\boldsymbol{\gamma}_i = \boldsymbol{\mu}_i + \bar{\boldsymbol{e}}_i$ and a predictor of $\boldsymbol{\mu}_i$ will be appropriate to predict $\boldsymbol{\gamma}_i$ if N_i is large. Since $\boldsymbol{\mu}_i$ is a linear function of $\boldsymbol{B}$ and $\boldsymbol{v}_i$, Datta et al. (1998) considered HB predictor of $\boldsymbol{B}\boldsymbol{h} + \boldsymbol{A}\boldsymbol{\lambda}$. The HB predictor and the associated posterior variance of an individual small area mean were computed by Datta et al. (1998) using Gibbs sampling. The corresponding posterior is proper under mild conditions on n, m, s, and d, which are satisfied for the example of crop area estimation for the counties of North Central Iowa. The full conditional distributions to implement the Gibbs sampling are all standard distributions such as multivariate normal and inverse Wishart.

Instead of using proper priors for $\boldsymbol{\Sigma}_v$ and $\boldsymbol{\Sigma}_e$, it is possible to use improper priors as well. Datta et al. (1998) showed that an improper prior of the form

$$\pi(\boldsymbol{B}, \boldsymbol{\Sigma}_v, \boldsymbol{\Sigma}_e) \propto |\boldsymbol{\Sigma}_v|^{-\frac{a_v}{2}} |\boldsymbol{\Sigma}_e|^{-\frac{a_e}{2}} \tag{45}$$

with suitable known a_v and a_e can be used. Again, note that for this prior the full conditional distributions for the Gibbs sampling are all standard distributions, multivariate normal and inverse Wishart. Though for not all choices of a_v and a_e the posterior distribution will be proper, Datta et al. (1998) showed that the posterior distribution will be proper if and only if (i) $n + a_e + a_v - 3s - d - 1 > 0$, (ii) $-t + 2s < a_v < 2$, where $t = \text{rank}(\boldsymbol{C}|\boldsymbol{F}) - \text{rank}(\boldsymbol{C})$, where the matrices $\boldsymbol{C}$ and $\boldsymbol{F}$ appear in (26), and are defined earlier (26).

3.3. HB and EB estimation with correlated sampling errors

The standard Fay–Herriot model that is widely used in small area estimation for area-level data assumes independence of sampling errors corresponding to small areas.

However, in some applications the assumption of independent sampling errors does not hold. One such example is adjustment of the U.S. Decennial Census counts for various cross-classifications based on demographic and georgraphic categories. To adjust the census counts, the U.S. Census Bureau conducted a Post Enumeration Survey (PES) 3 or 4 months after the Census and produced dual system estimates of adjustment factors $\theta_i = T_i/C_i$, where T_i denotes the true count and C_i is the census count for the ith poststratum. Poststrata were defined based on geographical region, race, sex, age, and housing arrangement. The Census Bureau developed a poststratum level model using the PES estimates Y_i and poststratum level covariates $\boldsymbol{x}_i$. The poststrata estimates are ordinarily correlated and the variance-covariance matrix of poststrata estimates variances are estimated and smoothed by variance function modeling. By treating the estimated poststrata variance–covariance matrix as the true variance–covariance matrix, Isaki et al. (1991) discussed EBLUP estimation of adjustment factors using a poststratum level model for the PES estimates Y_i. Many other articles dealing with modeling of the census adjustment factors have been published in the literature. See for example Ericksen and Kadane (1985), Freedman and Navidi (1986), and Cressie (1992).

Following the work of Isaki et al. (1991), Datta et al. (1992) considered HB and EB estimation of census adjustment factors for 84 poststrata based on the 1988 Missouri Dress Rehearsal Data from test sites in Missouri. Of the $m = 84$ poststrata covering two geographical regions, 48 are defined for St. Louis and 36 are for East Central Missouri. The poststrata estimates within St. Louis and within East Central Missouri are correlated but the estimates from St. Louis are assumed uncorrelated with the estimates from the East Central Missouri. Let $Y_i = \mathrm{DSE}_i/C_i$, where DSE_i is the dual system estimate of T_i. Denoting the vector of Y_i by $\boldsymbol{Y}$ and the sampling variance–covariance matrix by $\boldsymbol{D}$, Datta et al. (1992) considered the following HB model. Define $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)^T$.

(I) $\boldsymbol{Y}|\boldsymbol{\theta}, \boldsymbol{\beta}, \sigma_v^2 \sim N(\boldsymbol{\theta}, \boldsymbol{D})$, where $\boldsymbol{D}$ is a known $m \times m$ p.d. matrix;
(II) $\boldsymbol{\theta}|\boldsymbol{\beta}, \sigma_v^2 \sim N(\boldsymbol{X}\boldsymbol{\beta}, \sigma_v^2 \boldsymbol{I})$;
(III) $\boldsymbol{\beta}$ and σ_v^2 are independently distributed with a uniform(R^p) distribution for $\boldsymbol{\beta}$ and a uniform$(0, \infty)$ for σ_v^2.

For the application above the sampling variance–covariance matrix is a block diagonal matrix consisting of two blocks of dimensions 48×48 and 36×36. By variable selection method, from a set of 22 potential explanatory variables a set of 10 explanatory variables were selected that defined the design matrix $\boldsymbol{X}$. For further description of these explanatory variables one may refer to Datta et al. (1992).

If we write $\boldsymbol{\Sigma} = \boldsymbol{D} + \sigma_v^2 \boldsymbol{I}$ and $\boldsymbol{W} = \boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}$, it follows from (I)–(III) that

$$\boldsymbol{\theta}|\boldsymbol{y}, \sigma_v^2 \sim N\left(\boldsymbol{y} - \boldsymbol{D}\boldsymbol{W}\boldsymbol{y}, \boldsymbol{G}_1\left(\sigma_v^2\right) + \boldsymbol{G}_2\left(\sigma_v^2\right)\right), \tag{46}$$

where $\boldsymbol{G}_1\left(\sigma_v^2\right) = \boldsymbol{D} - \boldsymbol{D}\boldsymbol{\Sigma}^{-1}\left(\sigma_v^2\right)\boldsymbol{D}$, $\boldsymbol{G}_2\left(\sigma_v^2\right) = \boldsymbol{D}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{D}$ and the marginal posterior density of σ_v^2 is given by

$$\pi\left(\sigma_v^2|\boldsymbol{y}\right) \propto |\boldsymbol{\Sigma}|^{-1/2}|\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X}|^{-1/2}\exp\left(-\frac{1}{2}\boldsymbol{y}^T\boldsymbol{W}\boldsymbol{y}\right). \tag{47}$$

It follows from the last two equations by iterated formulas for expectation and variance that the posterior mean and posterior variance of $\boldsymbol{\theta}$ are given by

$$\hat{\boldsymbol{\theta}}_{\mathrm{HB}} = E(\hat{\boldsymbol{\theta}}_{\mathrm{B}}|\boldsymbol{y}), \quad \boldsymbol{V}_{\mathrm{HB}} = E\{\boldsymbol{D} - \boldsymbol{DWD}|\boldsymbol{y}\} + V(\hat{\boldsymbol{\theta}}_{\mathrm{B}}|\boldsymbol{y}) \tag{48}$$

where $\hat{\boldsymbol{\theta}}_{\mathrm{B}} = \boldsymbol{y} - \boldsymbol{DWy}$ and the expectations are evaluated with respect to the posterior density (47). Datta et al. (1992) evaluated the HB estimator and the posterior variance by numerical integration.

Instead of integrating out σ_v^2 in the Bayes estimator $\hat{\boldsymbol{\theta}}_{\mathrm{B}}$, if we replace σ_v^2 by a suitable estimate $\hat{\sigma_v^2}$ obtained from the marginal distribution of the data based on (I) and (II) of the aforementioned HB model, the resulting estimator is the EB or EBLUP of $\boldsymbol{\theta}$. Let $\hat{\boldsymbol{\Sigma}} = \boldsymbol{D} + \hat{\sigma_v^2}\boldsymbol{I}$. Plugging $\hat{\boldsymbol{\Sigma}}$ for $\boldsymbol{\Sigma}$ in $\boldsymbol{G}_1$ and $\boldsymbol{G}_2$ will result in an underestimation of the MSE of the EB or EBLUP, where the MSE is computed on the basis of the marginal distribution of the data using (I) and (II). The underestimation is due to ignoring the second term in (48) which accounts for the estimation error in estimating σ_v^2. Datta et al. (1992) also provided a second-order approximation to the MSE of the EB estimator $\hat{\boldsymbol{\theta}}_{\mathrm{EB}}$ given by $\mathrm{MSE}(\hat{\boldsymbol{\theta}}_{\mathrm{EB}}) = \boldsymbol{G}_1(\sigma_v^2) + \boldsymbol{G}_2(\sigma_v^2) + \boldsymbol{G}_3(\sigma_v^2)$, where $\boldsymbol{G}_3(\sigma_v^2)) = \boldsymbol{DW}^3\boldsymbol{D}\mathrm{var}(\hat{\sigma_v^2})$. Rao (2003a, Section 8.2) gives a second-order unbiased estimator of the MSE given by $\mathrm{mse}(\hat{\boldsymbol{\theta}}_{\mathrm{EB}}) = \boldsymbol{G}_1(\hat{\sigma_v^2}) + \boldsymbol{G}_2(\hat{\sigma_v^2}) + 2\boldsymbol{G}_3(\hat{\sigma_v^2})$, where $\hat{\sigma}_v^2$ is an unbiased estimator of σ_v^2.

3.4. Cross-sectional time series estimation

In Subsection 2.3, we review some applications of frequentist approach to small area estimation from time series and cross-sectional data. Here we review a few applications based on an HB approach.

Ghosh et al. (1996) were the first to develop HB estimation of small area means based on time series and cross-sectional data. They proposed their model for estimating the median income of various family sizes for the fifty U.S. states and Washington, DC based on CPS direct estimates, which are annual estimates. Although they considered the more general multivariate approach, for simplicity of presentation and comparison with the EBLUP methods of Datta et al. (2002) we will consider the univariate version of the Ghosh et al. (1996) model.

(I) $Y_{it}|\theta_{it} \overset{ind}{\sim} N(\theta_{it}, \psi_{it})(i = 1, \ldots, m, t = 1, \ldots, T)$ where, as in the other models, the sampling variances ψ_{it} are taken as known;

(II) $\theta_{it}|\boldsymbol{\beta}, \boldsymbol{b}_t, \tau_t \overset{ind}{\sim} N(\boldsymbol{x}_{it}^T\boldsymbol{\beta} + \boldsymbol{z}_{it}^T\boldsymbol{b}_t, \tau_t)(i = 1, \ldots, m, t = 1, \ldots, T)$;

(II) $\boldsymbol{b}_t|\boldsymbol{b}_{t-1}, \boldsymbol{W} \overset{ind}{\sim} N(\boldsymbol{b}_{t-1}, \boldsymbol{W})(t = 1, \ldots, T)$;

(IV) Marginally, $\boldsymbol{\beta}, \tau_1, \ldots, \tau_T, \boldsymbol{W}$ are independently distributed with a uniform(R^p) prior for $\boldsymbol{\beta}$, an inverse gamma prior for τ_t and an inverse Wishart prior for $\boldsymbol{W}$.

In (II) above, $\boldsymbol{x}_{it}$ and $\boldsymbol{z}_{it}$ are known vectors of covariates. The vector $\boldsymbol{x}_{it}$ is the same as the one used by Datta et al. (2002). As pointed out earlier, the above model in comparison with Datta et al. (1999b) or Rao and Yu (1994) model does not include a random small area effects term. Unlike Pfeffermann and Tiller (2006) this model assumes the same vector $\boldsymbol{b}_t$ for all the areas.

Ghosh et al. (1996) used Gibbs sampling to obtain the HB estimators of the state median income and the associated posterior variances. For the above HB model the Gibbs sampling is straightforward because all the full conditional distributions are either

multivariate normal, inverse gamma, or inverse Wishart. The $\mathrm{HB}^{(1)}$ estimator used in the comparison in Table 2 was computed by Ghosh et al. (1996) based on the above HB model. Actually, the HB model of Ghosh et al. (1996) is more general than the one we presented earlier. They considered a multivariate model for estimating three-person, four-person, and five-person family median income. The above model of Ghosh et al. (1996) assumes that the direct CPS estimators are uncorrelated over time. Since as noted before the CPS estimators are highly correlated because of the rotation pattern (4-8-4) of the underlying sampling design. Ghosh et al. (1996) have also considered several AR(1) models for $Y_{it} - \theta_{it}$ with known autocorrelation. The estimates under these models turned out to be quite different from their original model. They have documented comparison of the estimates from various models in Table 8 of their article. They concluded that the independence assumption provides better estimates on an average than the AR(1) models.

Datta et al. (1999b) considered HB estimation of monthly unemployment rates for 49 U.S. states and Washington, D.C. based on time series cross-sectional data on unemployment rates using monthly CPS estimates. Estimates of unemployment rates for 48 months starting in January 1985. This data was also analyzed by Tiller (1992) who used a time series approach and did not borrow strength from the other states. Pfeffermann and Tiller (2006) used a similar model to Tiller (1992) but imposed benchmarking constraints such that resulting estimators borrow strength cross-sectionally as well. Both these articles used a frequentist state-space modeling approach. Since the CPS estimates y_{it} are not seasonally adjusted, unlike Tiller (1992) and Pfeffermann and Tiller (2007) who modeled the seasonal components, Datta et al. (1999b) accounted for the seasonality with year and month effects. In particular, let f_{itu} be an indicator variable for the uth month, defined as $f_{itu} = 1$ if $t = u \bmod 12$, $f_{itu} = 0$, otherwise; and $f_{it12} = 1$ if $t = 12, 24, 36, 48$, $f_{it12} = 0$ otherwise. Similarly, let g_{itw} be an indicator variable for the wth year, defined as $g_{itw} = 1$ if $12(w-1) < t \le 12w$, $g_{itw} = 0$ otherwise. They used the following as part of their hierarchical model:

$$\theta_{it} = x_{it}\beta_i + v_i + \sum_{u=1}^{12} f_{itu}\gamma_{iu} + \sum_{w=1}^{4} g_{itw}\zeta_{iw} + \alpha_{it},$$
$$i = 1, \ldots, 50(= m), \quad t = 1, \ldots, 48(= T), \tag{49}$$

where v_i and β_i are the state specific intercept and slope, respectively, and α_{it} is an error term that is needed to account for the variation not explained by the other components identified in (49). The usual restrictions, $\gamma_{i12} = -\sum_{u=1}^{11} \gamma_{iu}$ and $\zeta_{i4} = -\sum_{w=1}^{3} \zeta_{iw}$ have been imposed to have a full rank linear model. The auxiliary variable x_{it} is a scalar representing the monthly state unemployement insurance claims rate. In this model the random effects v_i account for cross-sectional variation, and the regression coefficients $\boldsymbol{\beta}_i$ are all the same for all the time points.

Let $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{iT})^T$, $\boldsymbol{\theta}_i = (\theta_{i1}, \ldots, \theta_{iT})^T$, $\boldsymbol{\gamma}_i = (\gamma_{i1}, \ldots, \gamma_{i11})^T$, and $\boldsymbol{\zeta}_i = (\zeta_{i1}, \ldots, \zeta_{i3})^T$. With the model for θ_{it} given above, in the first step of the HB model they used the sampling model

$$\boldsymbol{Y}_i | \boldsymbol{\theta}_i \overset{ind}{\sim} N(\boldsymbol{\theta}_i, \boldsymbol{\Psi}_i)(i = 1, \ldots, m), \tag{50}$$

where, as in the other models, sampling variance–covariance matrix $\boldsymbol{\Psi}_i$ for the ith area is taken as known and includes autocorrelation to account for the serial dependence

of the CPS estimators. Alternatively, to account for this sampling error autocorrelation Pfeffermann and Tiller (2006) used an AR(15) model. Datta et al. (1999b) completed the HB model by specifying

$$v_i \overset{iid}{\sim} N\left(v, r_v^{-1}\right), \quad \beta_i \overset{iid}{\sim} N\left(\beta, r_\beta^{-1}\right), \quad \boldsymbol{\gamma}_i \overset{iid}{\sim} N_{11}\left(\boldsymbol{\gamma}, \boldsymbol{W}_1^{-1}\right), \quad \boldsymbol{\zeta}_i \overset{iid}{\sim} N_3\left(\boldsymbol{\zeta}, \boldsymbol{W}_2^{-1}\right), \tag{51}$$

where $v_i, \beta_i, \boldsymbol{\gamma}_i$, and $\boldsymbol{\zeta}_i$ are mutually independent, $i = 1, \ldots, 50$. Specification (51) allows for correlation among y_{it} for different states. They assumed that the error terms in (49), that is, α_{it} follow a random walk model given by

$$\alpha_{it}|\alpha_{i,t-1} \sim N(\alpha_{i,t-1}, r_\alpha^{-1}), \tag{52}$$

where $t = 2, \ldots, 48$, independently for $i = 1, \ldots, 50$. This is in contrast with Rao and Yu (1994), who assumed a stationary model for α_{it}. Finally, the following improper prior distribution is assumed for the hyperparameters:

$$\begin{aligned} f(v, \beta, \boldsymbol{\gamma}, \boldsymbol{\zeta}, r_v, r_\beta, \boldsymbol{W}_1, \boldsymbol{W}_2, r_\alpha) \propto\ & r_v^{\frac{1}{2}b-1} e^{-\frac{1}{2}ar_v} \times r_\beta^{\frac{1}{2}d-1} e^{-\frac{1}{2}cr_\beta} \\ & \times |\boldsymbol{W}_1|^{\frac{k-11-1}{2}} e^{-\frac{1}{2}tr\left(\boldsymbol{S}_1\boldsymbol{W}_1\right)} |\boldsymbol{W}_2|^{\frac{l-3-1}{2}} e^{-\frac{1}{2}tr\left(\boldsymbol{S}_2\boldsymbol{W}_2\right)} \\ & \times r_\alpha^{\frac{1}{2}f-1} e^{-\frac{1}{2}er_\alpha}, \end{aligned} \tag{53}$$

where $\boldsymbol{S}_1 = \Delta_1 \boldsymbol{I}_{11}$, $\boldsymbol{S}_2 = \Delta_2 \boldsymbol{I}_3$, $k = 12$, $l = 4$, Δ_1 and Δ_2 are large positive numbers, $b = d = f = 2$, and a, c, e are small positive numbers.

Datta et al. (1999b) obtained the HB estimates of the current unemployment rates θ_{iT} via Gibbs sampling. The associated posterior standard deviation of the HB estimate was considerably smaller than the sampling standard deviation of the corresponding CPS estimate. For more on the HB analysis of this problem we refer the readers to the article by Datta et al. (1999b).

You et al. (2003) applied HB time series and cross-sectional modeling to estimate Canadian unemployment rates for 62 Census Agglomerations (CA) in Canada. Their setup was similar to the setup of Datta et al. (1999b). Unlike Datta et al., You et al. modeled the α_{it} term by both a stationary AR(1) model with known autoregressive parameter and a random walk model as in Datta et al. (1999b). You et al. used unemployment data from the CLFS for 6 months during January 1999 to June 1999. Their choice of $T = 6$ was motivated by the fact that the serial correlation of the direct estimates weakens after a lag of 6 months due to the rotation pattern of six months in and then out underlying the CLFS design. A comparison of all the models considered by You et al. (2003) based on some model diagonistic criterion showed that the random walk model had the superior performance. For more details readers may refer to the article by You et al. (2003) (for an excellent summary one may read Section 10.8 of Rao, 2003a). Pfeffermann (2002, Section 5) reviews other applications of time series cross-sectional model for small area estimation.

3.5. Hierarchical Bayes small area estimation in GLMMs

We discussed in Section 2.4 the importance of small area estimation methods for discrete and categorical data using GLMM. There we considered an EB approach to this problem

and noticed that second-order unbiased estimation of the MSE of the EB predictor is rather challenging. Availability of fast computing has made it possible to use HB methods for complex models.

Several authors have proposed HB estimation of small area proportion with binary response data. We discussed in Section 2.4 EB estimation in the absence of covariate based on a beta-binomial model. He and Sun (1998) carried out an HB analysis of this model by assigning proper priors on a and b, the parameters of the beta distribution. They used this method to obtain HB estimates of success probabilities of hunting wild turkeys for 115 counties of Missouri.

The beta-binomial model is not quite flexible to include covariates in modeling the small area proportions. To include covariates the logistic regression model has been extended by including a random intercept term that represents a random small area effect. In particular, a logit-normal model by assuming a normal distribution for the small area effects has been used by many authors. Although this type of model given by (31) has been used by MacGibbon and Tomberlin (1989) in an EB approach, Farrell (2000) considered an HB approach for this model. Farrell (2000) used the uniform prior for the fixed logistic regression coefficients and a diffused inverse gamma prior for the small area variance parameter σ_v^2. Using griddy-Gibbs sampler the author applied the HB method to a data selected using 1% sample from a United States Census for estimating local labor force participation rates.

As an alternative to this model, Malec et al. (1997) proposed a two-level model for estimating the proportion of persons at the state or substate level who have visited a physician in the year preceding the interview based on data from the U.S. National Health Interview Survey (NHIS). At the first level of their logistic regression model they assumed a random regression coefficients model given by

$$Y_{ij} \sim \text{Bernoulli}(p_{ij}), \quad \text{logit}(p_{ij}) = \boldsymbol{x}_{ij}^T \boldsymbol{\beta}_i, \quad j = 1, \ldots, n_i, \quad i = 1, \ldots, m, \tag{54}$$

and at the second level $\boldsymbol{\beta}_i$ is modeled as

$$\boldsymbol{\beta}_i = \boldsymbol{Z}_i \alpha + \boldsymbol{v}_i, \tag{55}$$

with $\boldsymbol{v}_i \overset{iid}{\sim} N(\boldsymbol{0}, \boldsymbol{\Sigma}_v)$. In this model, $\boldsymbol{x}_{ij}$ is the unit-level covariate and the $p \times q$ matrix $\boldsymbol{Z}_i$ is the design matrix based on area-level covariate, and α is a $q \times 1$ vector of regression coefficients corresponding to the second level. They used a uniform prior for α and $\boldsymbol{\Sigma}_v$. In another article, Malec et al. (1993) considered the HB estimation of small area proportions using data from NHIS. Nandram and Sedransk (1993) suggested Bayesian predictive inference for binary data from a two-stage cluster sample. Stroud (1991) proposed an HB methodology for small area estimation with binary response data. Subsequently, Stroud (1994) considered a comprehensive Bayesian treatment of binary survey data for various sampling designs.

As mentioned earlier, estimation of MSE of the EB predictors for GLMMs is quite challenging. Given the recent advances in the development of computational algorithm and computational power it makes sense to apply a fully Bayesian approach to the small area estimation problems involving GLMMs. Ghosh et al. (1998) introduced the following HB model for small area estimation problems in a generalized linear

model setup. Let Y_{ij} denote the response of the jth unit in the ith small area. Also, let $\boldsymbol{\theta} = (\theta_{11}, \ldots, \theta_{1n_1}, \ldots, \theta_{m1}, \ldots, \theta_{mn_m})^T$, $\boldsymbol{v} = (v_1, \ldots, v_m)^T$, $R = \sigma^{-2}$, and $R_v = \sigma_v^{-2}$. Their HB model is as follows:

(I) Conditional on $\boldsymbol{\theta}$, $\boldsymbol{\beta}$, $\boldsymbol{v}$, $R_v = r_v$, and $R = r$, the responses Y_{ij} are independent with densities given by (32).

(II) Conditional on $\boldsymbol{\beta}$, $\boldsymbol{v}$, $R_v = r_v$, and $R = r$, $h(\theta_{ij}) \overset{ind}{\sim} N\left(\boldsymbol{x}_{ij}^T\boldsymbol{\beta} + v_i, r^{-1}\right)$.

(III) Conditional on $\boldsymbol{\beta}$, $R_v = r_v$, and $R = r$, $v_i \overset{iid}{\sim} N\left(0, r_v^{-1}\right)$.

(IV) $\boldsymbol{\beta}$, R_v, R are mutually independent with $\boldsymbol{\beta} \sim$ uniform(R^p), $R_v \sim$ gamma $\left(\frac{1}{2}a, \frac{1}{2}b\right)$ and $R \sim$ gamma $\left(\frac{1}{2}c, \frac{1}{2}d\right)$.

Ghosh et al. (1998) were interested in finding the joint posterior distribution of functions $g(\theta_{ij})$'s, given the data $\boldsymbol{y} = (y_{11}, \ldots, y_{1n_1}, \ldots, y_{m1}, \ldots, y_{mn_m})^T$, where $g(.)$ is a known strictly increasing function. In typical applications, $g(\theta_{ij}) = \psi'(\theta_{ij}) = E(Y_{ij}|\theta_{ij})$. The posterior distribution is typically summarized through its moments, such as means, variances, and covariances.

A large number of Bayesian applications involve the use of some type of objective priors, which are often diffused, the idea being the large number of sampled areas a diffused prior will let the data dominate the inference. With the use of improper priors it may happen that posterior distribution is likewise improper. It is thus important to ensure that in the presence of an improper prior, the resulting posterior distribution is proper. Under the assumptions that $a > 0$, $c > 0$, $\sum_{i=1}^m n_i - p + d > 0$, and $m + b > 0$, Ghosh et al. (1998, Theorem 1) proved the propriety of the posterior distribution. In particular, for the variance components σ^2 and σ_v^2 standard objective prior for scale parameter given by $\pi(\sigma^2) \propto \sigma^{-2}$ or $\pi(\sigma_v^2) \propto \sigma_v^{-2}$ will lead to an improper posterior. One may note that $\pi(\sigma_v^2) \propto \sigma_v^{-2}$ would be obtained as the Jeffreys' prior calculated from the distribution of v_i's.

Because the posterior distribution is high-dimensional and it is not a standard distribution, one can get samples from the posterior distribution via MCMC simulations. For the HB model Ghosh et al. (1998) used all the full conditional distributions required to sample in Gibbs sampling are standard distributions such as normal or gamma distribution, except the conditional distribution of θ_{ij} given the other parameters and the data. Under the assumption of a canonical link function, that is $h(z) = z$, the authors showed that this conditional density is log-concave and can be sampled via the adaptive rejection sampling scheme of Gilks and Wild (1992).

We have mentioned earlier that Ghosh et al. (1998) were interested in drawing inference on some functions of the parameters θ_{ij}. However, it is possible to use the posterior sample generated in Gibbs sampling to make predictive inference. In particular, if the goal is to obtain Bayes estimate of the finite population mean $\gamma_i = N_i^{-1}\sum_{j=1}^{N_i} Y_{ij}$, under squared error loss, the Bayes estimate is the posterior mean, which can be computed as follows. We first specify the superpopulation distribution by modifying (I) and (II) of the above HB model.

(I) Conditional on $\boldsymbol{\theta}$, $\boldsymbol{\beta}$, $\boldsymbol{v}$, $R_v = r_v$, and $R = r$, responses Y_{ij} are independent with densities given by (32), $j = 1, \ldots, N_i$.

(II) $h(\theta_{ij}) = \boldsymbol{x}_{ij}^T\boldsymbol{\beta} + v_i + u_{ij}$, where $u_{ij} \overset{iid}{\sim} N(0, r^{-1})$, $j = 1, \ldots, N_i$.

Second, for any unsampled unit Y_{ij}, compute

$$E[Y_{ij}|\boldsymbol{y}] = E[\psi'(\theta_{ij})|\boldsymbol{y}] = E[s(\boldsymbol{x}_{ij}^T\boldsymbol{\beta} + v_i + u_{ij})|\boldsymbol{y}],$$

where $s(x) = \psi'(h^{-1}(x))$. So along with $\boldsymbol{\beta}$, v_i, r_v, and r, one will also need to generate sample for u_{ij} from $N(0, r_u^{-1})$ to get an estimate of $E[Y_{ij}|\boldsymbol{y}]$.

The HB model of Ghosh et al. (1998) is more general than similar GLMM considered by other authors. In step (II) of their HB model they considered $h(\theta_{ij}) = \boldsymbol{x}_{ij}^T\boldsymbol{\beta} + v_i + u_{ij}$, where v_i is the random small area effect and u_{ij} explains possible lack of fit to the model specified by $\boldsymbol{x}_{ij}^T\boldsymbol{\beta} + v_i$. In the GLMM formulation or in logit-normal specification for binary response, many authors such as Breslow and Clayton (1993), Zeger and Karim (1991), MacGibbon and Tomberlin (1989), and Farrell (2000) did not use the error term u_{ij} in their models. Ghosh et al. (1998) model has greater flexibility due to this additional error term and it can account for overdispersion which may not otherwise be captured by a model without this term.

Ghosh et al. (1998) generalized the HB model given above to model data with multiple categories. This is a generalization of the logit-normal model for binary response with two categories to more than two categories. They applied their model to real datasets. One of the datasets has responses to the question "Have you experienced any negative impact of exposure to health hazards in the workplace?" based on a 1991 sample of all persons in 15 geographic regions of Canada. For each region, workers were classified by age (≤ 40 or > 40) and gender. The responses were classified into four categories as follows: (1) yes, (2) no, (3) not exposed, and (4) not applicable or not stated. They estimated the proportion of workers in each of the four categories for the 60 groups (small areas). The HB estimates of the cell probabilities borrow strength from other cells and geographic regions. Shrinkage of the sample cell proportions toward the grand mean was achieved adaptively.

Ghosh et al. (1998) also proposed a spatial generalization of their HB model. In particular, they replaced the iid assumption for the small area effects (the v_i) by a multivariate normal model appropriate to capture spatial dependence. They suggested as a joint distribution of v_i given by the density

$$p(v_1, \ldots, v_m) \propto r_v^{m/2} \exp\left[-\frac{r_v}{2}\sum_{i,l=1}^{m} q_{il}(v_i - v_l)^2\right],$$

where the q_{il} are strictly positive if the areas i and l are contiguous, and zero otherwise. Note that the above density, being a function of only the differences of the v_i, is not a proper density. However, the posterior distribution will be proper. In the context of disease mapping, datasets often show a strong spatial pattern. The above modification is useful to model spatial dependence. We refer to Datta et al. (2000) for a survey of EB and HB approaches to disease mapping. Ghosh et al. (1998) applied this modified HB model to the analysis of a Missouri lung cancer dataset. In disease mapping, the response variable is a count, usually the number of deaths in a region from a particular cause, and is modeled by a Poisson distribution. In their application, Ghosh et al. used a log-linear mixed model with spatial random effects. For a detailed discussion of the analysis of the Missouri lung cancer data, we refer to the paper by these authors.

4. Concluding remarks

In this chapter, we have reviewed some of the recent developments in small area estimation. For other recent reviews of this literature readers may refer to Pfeffermann (2002) and Rao (1999, 2003b, 2006). We conclude this chapter with some brief remarks on some topics that we have not considered so far. We note first that results from a model-based inference depend on the goodness of fit of the model used. Naturally, it is of critical importance to check if the working model is validated by the data and if it produces sensible predictions. These issues are rather broad and are substantively discussed in Rao (2003a). This book considers both frequentist and Bayesian solutions for model selection. We also refer to the work of Ghosh et al. (1998) for model diagnostics of their HB GLMM models in small area estimation. A few other important issues are listed in the following sections.

4.1. Unknown sampling error variances

An important assumption to compute model-based estimates based on area-level data using the Fay–Herriot model and its various univariate, multivariate, and time series extensions, is that the sampling variances and covariances associated with the direct estimators are known. However, these known sampling variances are actually estimated from data. In two recent articles Rivest and Vandal (2003) and Wang and Fuller (2003) obtained the EBLUP of the small area mean when the sampling variances are also estimated. The authors have also obtained approximations to the MSE of the EBLUP and derived estimators of the MSE.

4.2. Confidence intervals

A bulk of the small area estimation research focuses on the estimation of MSE of the EBLUP and EB point estimates. However, the construction of EB confidence intervals has been very sparse. Even conventional t-based confidence intervals utilizing a second-order unbiased estimator of MSE do not always achieve target coverage probabilities that are accurate up to second-order. Such an interval is not second-order accurate because the the expansion of its coverage probability differs from the target coverage by a $O(1/m)$ term. Confidence intervals that utilize a naive MSE estimator tend to be too short because of the underestimation of the MSE, and will usually fail to meet the target coverage probabilities. An early solution to produce second-order accurate confidence intervals is due to Smith (2001). Datta et al. (2002b) provided a rigorous derivation of second-order accurate confidence intervals for the balanced Fay–Herriot model. Nandram (1999) developed confidence intervals for small area means for a type of unit-level small area model not considered here. Ghosh and Maiti (2008) developed empirical Bayes confidence intervals for small area means based on Edgeworth expansions. For a parametric bootstrap approach to confidence intervals one may refer to Chapter 28.

In a fully Bayesian approach (HB approach) one can obtain a Bayesian credible interval based on MCMC samples. Using suitable sample percentile points from the MCMC sample one can construct a credible interval satisfying approximately the targeted posterior coverage probability, also called credible level. However, the frequentist coverage

probability of a Bayesian credible interval may be quite different from the credible level. Datta et al. (2002b) obtained a rigorous expansion of the frequentist coverage probability of a Bayesian credible interval for the balanced Fay–Herriot model.

4.3. Constrained EB/HB estimation

EBLUP, EB, and HB methods have been extensively used to develop estimates of small area means. However, one may be interested in subgroup analysis where the problem is not only to estimate individual small area means, but also to identify small areas whose true means are either below or above certain cut-off point. One may also be interested in generating the histogram of the small area means, or estimating the ordered small area means or the ranks of the individual small area means. For detailed descriptions and solutions to these problems we refer to Ghosh (1992) and Shen and Louis (1998). Typically, the EBLUP, EB, or HB predictors overshrink the direct estimates of small area means to their regression estimates. This results in a histogram of these model-based small area estimates (for example, the HB estimates) that is too concentrated compared to the histogram of the true small area means. Indeed, the sample variance of the HB estimates (under squared error loss) of the small area means is smaller than the posterior expectation of the sample variance of the true small area means. (Note that the first sample moment of the HB estimates agrees with the posterior expectation of the sample mean of the true small area means.) One way to rectify this deficiency is to modify the HB estimates so that the first two sample moments agree with the corresponding posterior expectations of the first two sample moments of the true small area means by putting appropriate constraint (hence the name). A similar argument has been made for EB/EBLUP estimates. The resulting constrained estimates put more weight to the direct estimates (and hence, less weight to the regression estimates) compared to the usual model-based (HB, EB/EBLUP) estimates.

A practically important issue in which some sense is related to constrained estimation is benchmarking, where it is often required that the small area estimates when aggregated, agree with direct estimate in a broader area where it can be trusted (for example, at the national level where the sample size is sufficiently large. This is important for consistency of publications. It also warrants some model robustness. with the national level direct estimate. For comments related to Bayesian and frequentist approaches to this problem, one may refer to Singh (2006, Section 2.5). One may also refer to Pfeffermann and Barnard (1991) and Pfeffermann and Tiller (2006).

4.4. Conditional MSE

In our discussion of EBLUP or EB predictors of small area means, we considered MSE of these estimators based on the distribution specified jointly by the sampling model and the linking model. Second-order approximations to the MSE and second-order unbiased estimation of the MSE are based on the marginal distribution of the data (e.g., the direct estimators). For the Fay–Herriot model, Rivest and Belmonte (2000) to evaluate certain shrinkage estimators (such as EBLUP or EB estimator) of small area mean θ_i, derived MSE of a small area estimator conditional on $\boldsymbol{\theta}$. They also obtained an exact unbiased estimator of the MSE by applying Stein's identity. Fuller (1989) suggested conditional MSE of an estimator of the area mean θ_i conditional on the direct estimator Y_i. This

MSE measure is a compromise between unconditional MSE and the posterior variance in the HB approach. In this context, we refer the reader to Booth and Hobert (1998), Singh et al. (1998) and an unpublished report by this author with two collaborators.

4.5. Pseudo EBLUP and pseudo HB estimators of small area means

For unit-level models, for example, the nested error regression model, the EBLUP or the HB predictor of the small area mean completely ignores the survey weights w_{ij} attached to the jth unit in the ith small area. These estimators are not design-consistent and may be highly biased if the linking model is not correct. Design-consistent model-based small area estimators are appealing to survey practitioners (Kott, 1989; Prasad and Rao, 1999), because they provide protection against model failures as the small area sample size, n_i, increases. Using the survey weighted induced area-level model, Prasad and Rao (1999) and You and Rao (2002a) obtained pseudo EBLUP of the small area mean (see also Kott, 1989). Similarly, as an alternative to the HB predictor, You and Rao (2003) developed a pseudo HB predictor. Both pseudo EBLUP and pseudo HB predictors are design-consistent. For a different approach of incorporating the survey weights see Section 6.3 of Chapter 39.

4.6. Small area estimation with covariates measured with error

One or more auxiliary variables in model-based small area estimation may be subject to measurement error. In two recent articles Ghosh and Sinha (2007) and Ghosh et al. (2006) considered estimation of small area means for the nested error regression model with a single covariate subject to measurement error. They have considered both the functional and structural measurement error models. In their nested error model they assumed that the the true value of covariate (say, x_i) remains the same for all the units in a small area, and it is measured with error as X_{ij} for the jth sampled unit in the ith small area. In the functional approach, they proposed a one-way ANOVA model with fixed effects for the X_{ij}. In the structural approach, they proposed a one-way ANOVA model with random effects for the X_{ij}. To obtain the EB or HB predictor of the small area means first they obtained, assuming the model parameters known, the usual Bayes predictor using the predictive distribution of the unsampled y values given the values of the sampled y. They obtained their EB predictors by estimating the unknown model parameters from the joint distribution of the sampled y and the sampled X values. Noting that the Bayes predictor of Ghosh et al. (2006) did not condition also on the sampled X values, this author with two collaborators suggested a modification of Ghosh et al. (2006) by conditioning on the X values as well which resulted in a better predictor. The new EB predictor has a smaller MSE than the MSE of the EB estimator by Ghosh et al. (2006). In a different context for the area-level data, Ybarra and Lohr (2008) in a recent article considered small area estimation for the Fay–Herriot model with certain covariates measured with error. Details are omitted due to lack of space.

4.7. Small area estimation with unmatched sampling and linking models

For area-level data in the Fay–Herriot model given by (4), the direct estimator Y_i estimates μ_i and a normal regression model on μ_i is proposed. In this model, typically,

for the sampling error a normal distribution is used. However, sometimes, a normal regression model on $g(\mu_i)$ for some nonlinear function $g(\cdot)$ is more appropriate. As an example, in estimating a small area proportion μ_i, a logistic regression model for the probability of success is better suited than a linear regression model. Another example is a log-linear model in estimating small area counts. These are examples of unmatched linking model where the nonlinear linking model does not match with the sampling model for the direct estimator. You and Rao (2002b) used an unmatched linking model in estimating Canadian census undercoverage by using the HB approach. Although an unmatched linking model is more computer intensive than a standard Fay–Herriot model with a matched linking model, the former model provides a more realistic model. For further discussion one may refer to Rao (2003a, p. 243). It should be noted that the unmatched sampling and linking models are analogous to a Bayesian model where one uses a nonconjugate prior distribution.

4.8. Comparison of the frequentist and Bayesian approaches

In this chapter we have reviewed both frequentist and Bayesian approaches to model-based small area estimation. Although the frequentist approach is still more popular among practitioners, the Bayesian approach is also gaining popularity and acceptability. Though a frequentist or a Bayesian approach is a matter of personal choice, the Bayesian approach is based on probability calculus and conceptually straightforward where one uses conditional probability calculations to update a prior probability to a posterior probability in light of the data.

The difficulty in the Bayesian approach is prior specification and computation. Although the former is still a difficult issue, we have made enormous progress in recent years on computational issues. It is worthwhile to point out that frequentist solutions based on jackknife or bootstrap are also computer intensive. To resolve the issue of prior specification one need to study sensitivity analysis (see for example, Berger, 1985). Bayesian analysis is often performed based on objective priors. Objective Bayesian procedures, as they are called, often possess good frequentist properties. There is a large literature on this issue (see for example, Datta and Mukerjee, 2004). In small area estimation context Datta et al. (2005) obtained a Bayesian solution where the posterior variance is also a second-order unbiased estimator of the MSE. Such dual interpretation makes a Bayesian method very desirable. For an extension of the result of Datta et al. (2005) for the Fay–Herriot model, one may refer to Ganesh and Lahiri (2008).

One advantage with the Bayesian approach is that it automatically incorporates all sources of uncertainty associated with an inference problem. For example, the estimation error for the unknown hyperparameters (which are often variance components in small area estimation) is automatically taken into account. On the other hand in an EB or an EBLUP approach one needs to be careful to account for the estimation error associated with the hyperparameters. Laird and Louis (1987) suggested a bootstrap approach to provide a more accurate measure of uncertainty, measured by the MSE, associated with the EB estimator. Second-order accurate estimator of the MSE of the EBLUP is already extensively discussed in Section 2, and to achieve better approximation one needs heavy algebraic manipulations. A fully Bayesian approach is more flexible in handling complicated models such as GLMMs. Using Gibbs sampling, and intensive computing depending on the complexity of the Bayesian model, it is relatively routine to

obtain numerous copies of samples from the posterior distribution. From the posterior samples, it is relatively straightforward to calculate the posterior means, variances, and Bayesian credible intervals (based on the ordered posterior sample values) of the small area means. Although the Bayesian approach is more flexible than its frequentist counterpart, its relative disadvantage is in specifying the joint prior distribution on the model parameters.

Acknowledgements

Research of Datta was partially supported by NSF Grant SES-0241651. The author is grateful to Professor Malay Ghosh for many helpful discussions and his help in the preparation of this manuscript. He is also grateful to Professor Danny Pfeffermann, Professor J.N.K. Rao and two referees for many useful suggestions on the first version of this chapter.

Sample Surveys: Inference and Analysis, Vol. 29B
ISSN: 0169-7161
© 2009 Elsevier B.V. All rights reserved
DOI: 10.1016/S0169-7161(09)00233-8

Design and Analysis of Surveys Repeated over Time

David Steel and Craig McLaren

1. Overview of issues for repeated surveys

Many surveys are repeated on several occasions, and the associated estimates are used to analyze changes in variables over time. Major social and economic surveys, such a labor force and retail trade surveys (RTS), are conducted monthly or quarterly to identify changes in the level or rate of change of variables, including turning points. Repeated surveys can produce time series of estimates, which will be analyzed using estimates for several time periods. Examination of the change in estimates between two consecutive periods or the same period in the previous year is common. Seasonal effects can be estimated and removed to produce seasonally adjusted estimates. Seasonally adjusted series can be volatile and to assist in analyzing the underlying pattern of change moving averages or some form of trend estimation may be applied.

Many repeated surveys involve overlap in the sample between different time periods. The sample overlap induces a correlation structure in the sampling errors of the time series of estimates, which affects the analysis of changes in them and may be exploited in producing estimates. The correlation of the sampling errors affects the variability of the time series of survey estimates and seasonally adjusted and trend estimates produced from them. Population changes can contribute to the change in variables over time and so it is important that the population sampling frame is updated to incorporate changes in the population as quickly and regularly as possible.

The design of a sample over time needs to consider the frequency of sampling and the pattern of inclusion of selected units over time. The frequency of sampling depends on the purpose of the survey, how quickly changes are likely to occur and associated decisions are needed. Common frequencies for surveys are monthly, quarterly, and annual, although more frequent sampling may be adopted, for example in opinion polls leading up to an election or monitoring television ratings. A key design issue is whether to use overlapping or nonoverlapping samples over time. For overlapping samples, the precise pattern of overlap must be designed.

Repeated, panel and longitudinal surveys, rotating panel surveys, split panel surveys, and rolling samples are surveys that are designed to permit analysis over time.

In a panel or longitudinal survey, an initial sample is selected and at each occasion that the survey is conducted an attempt is made to include all the members of the initial sample, even if they move. Longitudinal surveys are developed to permit analysis of changes at the individual level. These surveys can provide estimates of change at the population level for variables for which information is collected at each occasion, provided strategies are used to keep the sample representative of the population at each time period. Analysis of net change using aggregate estimates may hide important gross changes occurring at the individual level, which may be revealed from longitudinal data. Longitudinal analysis can help determine the relationships between variables and looking at causes of change through examining the temporal sequences of events. Longitudinal survey data can be used for a variety of analyses, including survival analysis, event history analysis, and analysis of transition probabilities. Multilevel models that take account of the repeated nature of the data are being used increasingly used (Skinner and Holmes, 2003). Analysis of longitudinal studies is discussed in Chapter 34 of this volume.

In a repeated survey, there is not necessarily any overlap of the sample for the different occasions. When the emphasis is on estimates for the population and major subpopulations, an independent sample may be used on each occasion, which is often the case when the interval between the surveys is large. An alternative is to try to use the same sample at each occasion, with some additions to ensure that the sample estimates refer to the current population. For regular monthly or quarterly surveys, the sample is often designed so that there is considerable overlap in the sample between successive surveys. This can be done using rotating panel surveys that use a sample that is followed over time, but a proportion of sample units is removed from the survey at some time periods and replaced by other units. Usually units that move location are not followed. Having overlap in the sample will reduce the sampling variance of estimates of change. Cost savings often arise because on the first time a person or business is included in the survey there are higher setting-up costs than on subsequent occasions. Sampling variances of estimates of change are reduced because the variation due to including different units is reduced. The reduction in variances depends on the correlation of the variable at the individual level over time and the degree of sample overlap. If the correlation is low, then the reduction is small. The correlation needs to be positive for a reduction to apply. A negative correlation will increase sampling variances, although such cases are not common.

These considerations would lead to maximizing the sample overlap at each time period, with the only change in the sample arising from the need to update it to represent units moving in and out of the population. However, such a design would lead to selected units being included in the survey indefinitely. In practice, a limit needs to be placed on how many times a person or business is surveyed, to spread the reporting load and maintain response rates and the quality of the reported data. In deciding the degree of sample overlap these considerations need to be balanced. The degree of sample overlap between any two time periods is determined by the rotation pattern, which is the pattern of selected units' inclusion in the survey over time. A rotation sampling design can be implemented using rotation groups and panels. The sample will consist of several rotation groups. A panel is the set of selected units that enter and leave the sample at the same time. When a panel leaves the sample it is replaced by one from the same rotation group.

In a rotating panel survey, the focus is on aggregate estimates of change. However, any overlapping sample can also be used to analyze change at the micro-level. For example, a table can be produced from the matched sample showing the change of a variable between two time periods. An important example is when a table of change in status is produced, which is referred to as a Gross Flows table. Longitudinal data can be created from rotating panel surveys, but the length of the total time period and the time interval between observations are determined by the rotation pattern used. Also, the resulting sample of individuals for which a longitudinal data are available will be biased away from people who move permanently or are temporarily absent.

An alternative to a rotating panel survey is a split panel survey that involves a panel survey supplemented on each occasion by an independent sample. This design permits longitudinal analysis from the panel survey for more periods than would be possible in a rotating panel design but also produces cross-sectional estimates from the entire sample.

In deciding on the sample design, in general the three dimension of space, time, and variables need to be considered (Kish, 1987, 1998). A survey may be conducted continuously, but the sample size in any time period may not be sufficient to provide reliable estimates for that period, at least for subnational estimates. However, by cumulating the sample over several time periods reasonably reliable estimates may be produced. In this approach sample overlap is detrimental. The sample design can be developed so that it is a rolling sample with nonoverlapping samples that over time cover many areas and, eventually, all areas. This approach can be useful in producing subnational and small area estimates. A major example of this approach is the American Community Survey (Alexander, 2002). A related approach is rolling estimates. For example in the U.K. Labor Force Survey, a nonoverlapping sample is interviewed in each week of a quarter. Each month estimates based on an average of the latest 13 weeks are produced (Caplan et al., 1999; Steel, 1997). Further discussion of the issues associated with the design of surveys over time is given in Chapter 5 of this volume.

In a repeated survey, the estimates for a particular period can be calculated using only data for that period using standard sample weighting methods, such as calibration or generalized regression estimation (see Chapter 9 and Chapter 25 in this volume). When there is sample overlap it is possible to exploit the correlation structure for different rotation groups to produce estimates of levels and changes with smaller sampling variances using composite estimators and Best Linear Unbiased Estimates (BLUEs).

Repeated surveys can provide estimates for each time period, $y_t, t = 1, \ldots N$. When a monthly or quarterly survey has been conducted for several years then a time series can be produced and analyzed. Seasonally adjusted estimates are often produced to help interpretation of the time series, giving the series $SA_t, t = 1, \ldots N$. To assess the underlying pattern of change trend estimates can also be produced, which raises the question of what is trend? In some cases, it is taken to be $\Delta^{(s)} y_t = y_t - y_{t-s}$ or $\Delta^{(s)} SA_t$ and for s = 1 these will be volatile. The Australian Bureau of Statistics (ABS) publishes trend estimates using Henderson moving averages applied to the seasonally adjusted series (Australian Bureau of Statistics, 1993). Other options are available and are discussed in Section 8. Even if trend estimates are not calculated an informal analysis of trend in a monthly series may involve examining $\Delta^{(s)} y_t$ or $\Delta^{(s)} SA_t$ for $s = 1, 2, 3, 6, 12$.

Smith (1978) distinguishes between primary analysis that uses the individual sample observations, y_{it}, for $i \in s_t, t \in \tau_i$ and secondary analysis, which uses survey estimates

of the population total or mean, $y_t, t = 1, \ldots N$. Here, s_t is the sample at time t,and τ_i is the set of time periods for which population unit i is included. Analysis may also be based on elementary estimates calculated at a level below the overall population based on subsamples, where typically the subsamples correspond to panels in the sample.

At each time period, complex sample design may be used, possibly involving stratification, multistage sampling, and unequal selection probabilities. The estimates may be calculated using weighting to account for different selection probabilities and incorporate adjustments for nonresponse and calibration to known population data (See Chapter 25 in this volume). The effects of the complex design and estimation are usually taken into account in the estimation of sampling variances for levels and simple changes in survey estimates for the population using standard methods of variance estimation. An alternative approach is to include the population structure in the analysis through techniques, such as multilevel modeling if primary survey data are available. However, for analysis of the time series of estimates, there are different approaches to taking the complexities of the sample design and estimation into account and the effects of sample overlap. The approach will depend on the level of analysis (primary, elementary, secondary), the approach to the time series analysis, and the targets of inference. Primary analysis is rarely undertaken in repeated surveys, although it is used in longitudinal surveys.

Elementary estimates may be used for several reasons. They can be used in the analysis in ways that automatically estimate and reflect the variance and covariances of the sampling errors. This means that the structure of the sampling errors is directly included in the analysis. This can be done in a design-based approach using BLUEs and composite estimation (see Section 4) or in a model-based approach using state-space models (SSMs) (see Section 7). Also, use of the elementary estimates allows direct exploitation of the different correlation structure of different panels at each time point using weighted estimates that reflect the possible complex sample design and account for missing data, and changes in population composition through standard approaches, whereas a primary analysis would have to incorporate these in the analysis and modeling.

In this chapter, we assume that we are not directly interested in analysis of changes at the individual or micro level, even if we are using individual level data. We are concerned with situations in which the targets of inference are at the population level. We may perform a primary analysis but only because it offers some advantage in our inferences about population level behavior.

Duncan and Kalton (1987), Binder and Hidiroglou (1988), Kalton and Citro (1993), and Steel (2004) reviewed the issues in the design and analysis of repeated surveys. Holt and Skinner (1989) considered the various components that affect estimates of change obtained from repeated surveys. This chapter will focus on developments since Binder and Hidiroglou (1988). We will also consider the key design issue of deciding on the rotation pattern to use in a rotating panel survey, as well as approaches to analysis. Section 2 sets out some of the basic theory, Section 3 reviews rotation patterns, and Section 4 considers BLUE and composite estimation. Correlation models for survey errors are reviewed in Section 5 and the impact of different rotation patterns on the variances of key estimates are described in Section 6. Time series methods for estimation are described in Section 7. Seasonally adjusted and trend estimates are briefly described in Section 8 and variance estimation for these estimates are described in Section 9.

The issue of the effect of the design of rotation patterns on seasonally adjusted and trend series is explored in Section 10.

2. Basic theory of design and estimation for repeated surveys

Corresponding to the survey estimate for time t there is the population value Y_t and we can write

$$y_t = Y_t + e_t, \tag{1}$$

where e_t is the sampling error. If the survey estimate is unbiased, then $E[e_t \mid Y_t] = 0$. Properties of y_t and $\Delta^{(s)} y_t$ can be obtained using the sampling or randomization distribution to give the design-based expectations and variances.

A major value of repeated surveys is their ability to provide estimates of change. The simplest analysis of change is the estimate of one period change, $y_t - y_{t-1}$. In a monthly survey, this corresponds to 1-month change, and for a survey conducted annually, it corresponds to annual change. In general, the change s time periods apart can be estimated, using $y_t - y_{t-s} = \Delta^{(s)} y_t$. The focus is often on $s = 1$, but for a survey repeated on a monthly basis changes for $s = 2, 3, 12$ are also commonly examined. Having sample overlap at lag s will usually lead to a positive correlation between the estimates. Since

$$\mathrm{Var}(\Delta^{(s)} y_t) = \mathrm{Var}(y_t) + \mathrm{Var}(y_{t-s}) - 2\sqrt{\mathrm{Var}(y_t)}\sqrt{\mathrm{Var}(y_{t-s})}\ \mathrm{Corr}(y_t, y_{t-s}) \tag{2}$$

this overlap reduces the variance of $\Delta^{(s)} y_t$ compared with having no sample overlap.

If comparisons are made with time periods for which there are no sample units in common, then the variance of the estimate of change will be the sum of the variances, which will often be approximately twice the variance of the estimate of the level for a particular time period. These considerations result in designing the sampling so that there is overlap between the samples for time periods between which the movements are of major interest. So if there is strong interest in monthly movement, then there should be high sample overlap between successive months. If there is also interest in changes 12 months apart, then consideration should be given to designs that induce sample overlap at this lag. However, for many variables the individual level correlation 12 months apart may not be high enough for there to be appreciable gains from doing so. If there is interest in 3 month change, then sample overlap at 3 months lag is desirable. There may also be interest in changes in the rate of change, such as $\Delta^{(s)} y_t - \Delta^{(s)} y_{t-k} = y_t - y_{t-s} - (y_{t-k} - y_{t-k-s})$. If s = k, this becomes $y_t - 2y_{t-s} + y_{t-2s}$.

Equation (1) shows that for the estimation of the change between two time periods $t - s$ and t a key factor is the correlation between the two estimates, $\mathrm{Corr}(y_t, y_{t-s})$. In general, for complex designs and estimators, this correlation will also be complex and will depend on the design and the correlation between values for the same unit over time. In the traditional design-based approach, the variance under consideration is the sampling variance due to sampling errors only and so is the correlation. In Section 7, we also include and exploit variability due to the stochastic process generating the time series of population means or totals.

If the samples are independent between the two time periods, then $\mathrm{Corr}(y_t, y_{t-s}) = 0$.

In general, there will be overlap between the samples and the degree of overlap is a factor influencing the correlation. Consider the simple situation of a stable population (i.e., no births and deaths), and simple random sampling with negligible sampling fractions. Let n_t be the sample size at time t, $n_{t,t-s}$ is the size of the sample in common between the two time periods, $k_{t,t-s} = \frac{n_{t,t-s}}{n_t}$ is the proportion of the sample at time t that is common between periods t and $t-s$, and $k_{t-s,t} = \frac{n_{t,t-s}}{n_{t-s}}$ is the proportion of the sample at time $t-s$ that is common between periods t and $t-s$. Then

$$\mathrm{Corr}(y_t, y_{t-s}) = \sqrt{k_{t,t-s}k_{t-s,t}}r_{t,t-s},$$

where $r_{t,t-s}$ is the individual level correlation between values at time t and $t-s$. The correlation between the estimates will be zero if $n_{t,t-s} = 0$ irrespective of the individual level correlation or if $r_{t,t-s} = 0$ irrespective of the sample overlap and equals $r_{t,t-s}$ only if $k_{t,t-s} = k_{t-s,t} = 1$. Tam (1984) and Laniel (1987) gave more general results.

Many rotation patterns are designed so that $k_{t,t-s} = k_{t-s,t} = k(s)$, giving $\mathrm{Corr}(y_t, y_{t-s}) = k(s)r_{t,t-s}$. If the patterns of changes at the individual level do not vary over time, then $r_{t,t-s} = r(s)$. If $\mathrm{Var}(y_t)$ does not change, that is, there are no major changes to the sample design or the population structure, then the sampling errors are weakly stationary. Under these conditions

$$\mathrm{Var}(\Delta^{(s)}y_t) = 2\mathrm{Var}(y_t)(1 - k(s)r(s)). \tag{3}$$

Usually, we would expect the unit level correlation to be positive. These results suggest that the higher the sample overlap the higher the correlation between the estimates and this leads to designs with high sample overlap between periods for which the change is of interest. For analysis of one period changes high overlap between adjacent periods is desirable. More complex models for the correlation of the sampling errors are considered in Section 5.

The conditions for weak stationarity of the sampling errors involve characteristics of the population, namely $r_{t,t-s} = r(s)$ and relevant population variances, such as stratum or cluster population variances, and aspects of the design, such as sample overlap and sample size, and allocation to strata or stages in a multistage design. In a long running survey, there can be major redesigns of the sample that change the sampling variance. Periodically, a completely new sample may be selected, for example, after a census provides new information on the population, leading to a break in the sampling error series.

Moving averages and rolling estimates can be applied to the estimates produced from a repeated survey. This is a method of analysis that can be applied to any design, although it is particularly suited to rotation patterns that result in no sample overlap for the periods over which the averages are calculated. Consider 3 month moving averages in a monthly survey. There are two variants, nonoverlapping and overlapping. The nonoverlapping approach estimates 3 month on 3 month change. For example, it compares the average of December, January, and February with the average of March, April, and May. This is equivalent to the average of the 3 months changes. In general, the nonoverlapping approach calculates

$$\frac{y_t + y_{t-1} + y_{t-2}}{3} - \frac{y_{t-3} + y_{t-4} + y_{t-5}}{3} = \frac{1}{3}[(y_t - y_{t-3}) + (y_{t-1} - y_{t-4}) + (y_{t-2} - y_{t-5})].$$

The overlapping approach estimates average change over the last 3 months, for example, it compares the average of February, March, and April with the average of March, April, and May, which is equivalent to the difference between May and February divided by three. In general, the overlapping approach calculates

$$\frac{y_t + y_{t-1} + y_{t-2}}{3} - \frac{y_{t-1} + y_{t-2} + y_{t-3}}{3} = \frac{1}{3}(y_t - y_{t-3}).$$

This approach produces more up-to-date information than quarterly release of quarterly averages, but it is not as up-to-date as analysis of monthly estimates.

More generally, moving averages of the form $\tilde{y}_t^{(2k+1)} = \frac{y_{t+k}+\ldots y_t+\ldots y_{t-k}}{2k+1}$ may be applied to smooth the often volatile series of estimates that are obtained from a repeated survey. Such averages can be regarded as crude trend estimates, but better methods of trend analysis are available as discussed in Section 8.

Positive correlations increase the variance of averages over time. For example, the average of three consecutive months would have variance

$$\mathrm{Var}\left(\frac{y_{t+1} + y_t + y_{t-1}}{3}\right) = \frac{1}{9}\begin{bmatrix} \mathrm{Var}(y_{t+1}) + \mathrm{Var}(y_t) + \mathrm{Var}(y_{t-1}) \\ +2\mathrm{Cov}(y_{t+1,}\, y_t) + 2\,\mathrm{Cov}(y_{t,}\, y_{t-1}) \\ +2\mathrm{Cov}(y_{t+1,}\, y_{t-1}) \end{bmatrix}.$$

If the sample overlap depends only on the gap between two periods and the series of sampling errors are weakly stationary so that $\mathrm{Corr}(y_t,\, y_{t-s}) = k(s)r(s) = R(s)$, then

$$\mathrm{Var}\left(\frac{y_{t+1} + y_t + y_{t-1}}{3}\right) = \frac{\mathrm{Var}(y_t)}{9}[3 + 4R(1) + 2R(2)].$$

Similar reasoning for the change in nonoverlapping quarterly averages gives

$$\mathrm{Var}\left(\tilde{y}_t^{(3)} - \tilde{y}_{t-3}^{(3)}\right) = \frac{\mathrm{Var}(y_t)}{9}[6 + 6R(1) + 0R(2) - 6R(3) - 4R(4) - 2R(5)]$$

and for the change in adjacent overlapping averages,

$$\mathrm{Var}\left(\tilde{y}_t^{(3)} - \tilde{y}_{t-1}^{(3)}\right) = \mathrm{Var}\left(\frac{y_{t+1} - y_{t-2}}{3}\right) = \frac{\mathrm{Var}(y_t)}{9}[2 - 2R(3)].$$

These results suggest that for analysis based on three period averages and the changes in them having no sample overlap at lags 1 and 2 so that $R(1)$ and $R(2)$ are zero or small and having overlap at lag 3 so $R(3)$ is appreciable would be beneficial.

Averaging of estimates can be used to produce more stable estimates when the original estimates have high sampling variances, for example, for small subgroups in the population, such as estimates for small geographic areas. However, averaging over time changes the length of the time period to which the estimate refers and will hide any variation within the period over which the average is calculated. Time series methods are available to help combine data across time and space to produce small area estimates from rotating panel surveys. Small area estimation is considered in Chapter 31 and in Chapter 32 of this volume.

In general, for an estimator that is a linear combination of the estimates, $\mathbf{l}'\mathbf{y}_N$, where $\mathbf{y}_N = (y_1, \ldots y_N)'$ is the vector containing the values of the time series up to time N, and $\mathbf{l} = (l_1, \ldots, l_N)'$ is a vector of fixed coefficients, $\mathrm{Var}(\mathbf{l}'\mathbf{y}_N) = \mathbf{l}'\mathrm{Var}(\mathbf{y}_N)\mathbf{l}$. For example, the change in the most recent nonoverlapping 3 month averages corresponds to

$\mathbf{l} = \left(0, 0, \ldots, 0, -\frac{1}{3}, -\frac{1}{3}, -\frac{1}{3}, \frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)'$. In an evaluation of options for a U.K. Monthly Labor Force Survey, Steel (1996) used this approach to consider the impact of different rotation patterns for estimates of (i) the current month's level, (ii) the changes between months for s = 1, 3, 12, (iii) the 3 month averages (iv) 1 and 3 month changes in 3 month averages, (v) 12 month averages, and (vi) 1 and 12 month changes in 12 month averages. In general to determine the variance of $\mathbf{l}'\mathbf{y}_N$ we need to know, estimate or model $\mathrm{Var}(\mathbf{y}_N)$, which is considered in Section 5. For simple averages and differences, we only need to know those elements of $\mathrm{Var}(\mathbf{y}_N)$ corresponding to nonzero coefficients in $\mathbf{l}$.

3. Rotation patterns

An overlapping sample design can be implemented using a rotation pattern to manage the sample over time. Rotation patterns can be developed that have the same proportion of the sample in common between any two time periods the same time apart and the same proportion of sample rotated out and into the sample at each period. The rotation sample design should ensure that the cross-sectional estimates are unbiased and reduce the cost of the survey and the sampling variances on important estimates of change. A rotation pattern would usually ensure that at each time point the sample is balanced according to the number of times a person or business has been included in the survey because of the effect that the number of times a person has been included in the survey has on the data reported (Bailar, 1975).

Overlap in the sample may occur at different stages in a multistage design. Rotation is often carried out within primary sampling units (PSUs) for cost reasons. This can produce a small, secondary correlation between estimates even when there are no sample units in common. For example, in the Australian Labor Force Survey, the PSUs are allocated to eight rotation groups. In a particular month, the dwellings in one of the rotation groups are rotated out of the survey and replaced by a sample of dwellings in the same PSU.

A further aspect of the design is the level of information collected, which is the number of time periods for which information is collected on a particular occasion. For example, in a monthly survey, information may be collected from a unit for the current month and the previous month.

There are many different rotation patterns in use and more that can be considered. Consider a monthly survey. The simplest rotation pattern is when a unit is included for a months. Rao and Graham (1964) considered rotation patterns in which units remain in the sample for a time periods, leave for b, and then return for a further a time periods. This pattern is repeated until the unit is included for a total of m months. This can be denoted as an a-b-$a(m)$ rotation pattern. They found that for a composite estimators using $a = 2$ and $b = \infty$ gave maximum gain for estimating levels, but for estimating change a should be as large as possible, which suggests no rotation. This illustrates the trade-off between rotation pattern and target of analysis. For example, the U.S. Current Population Survey (CPS) uses a 4-8-4(8) rotation pattern, whereas the Australian Labor Force Survey uses an in-for-8 rotation pattern, which can be denoted 8(8). The Canadian Labor Force Survey uses an in-for-6 rotation pattern. These surveys use one level, so that information is collected referring to 1 month. More generally, a pattern of the form a_1-b_1-a_2-b_2-... a_p(m) can be considered, where the number of months included and excluded from the survey varies.

Different rotation schemes lead to different overlap patterns. An in-for-m scheme leads to a 1-s/m overlap between samples s months apart, for $s = 1, \ldots, m$-1 and no overlap for months m or more months apart. Unless m exceeds 12, there will be no sample overlap for months a year apart. The 1-2-1 (m) pattern leads to no sample overlap between months 1 or 2 months apart, but an overlap of 1-s/3m for $s = 3, 6, \ldots, 3m$. The sample overlap between months a year apart is 1-4/m provided m is five or more. The 4-8-4(8) rotation scheme leads to sample overlap of 1-s/4 for months s months apart, for $s = 1, 2, 3$. For $s = 12$, the overlap factor is 4/8 and the overlap is 4/8-abs(s-12)/8 for $s = 9, \ldots, 15$. The 6-6-6(12) scheme leads to sample overlap of 1-s/6 for months s months apart, s $= 1, \ldots, 5$. For $s = 12$, the overlap factor is 6/12 and the overlap is 6/12-abs(s-12)/12 for s $= 7, \ldots, 17$. Chapter 5 gives more details of rotating panel surveys.

Yansaneh and Fuller (1998) considered the impact of the 4-8-4-(8), in-for-8 and in-for-6 rotation patterns and composite estimators for the current level of the series and changes up to 12 periods. The U.S. unemployed person and civilian labor force series were considered. A similar study was conducted by Cantwell and Caldwell (1998) who examined the revisions in U.S. monthly retail and wholesale surveys under different rotation patterns. Bell (1999) considered the effect of the 4-8-4(8), 1-2-1(8), 2-2-2(8), and in-for-8 rotation patterns on direct and composite estimators of monthly level and movement, quarterly level and movement in the original series and the level and movement of the Henderson moving average-based trend. McLaren and Steel (2000) considered the impact of the following rotation patterns on the level and 1-month movement in seasonally adjusted and trend estimates: 1-2-1(m), $m = 5, 8$; 1-1-1-(6); 2-2-2(8); 2-10-2(4); 3-3-3 (6), 4-8-4(8), 6-6-6(12), in-for-m, $m = 6, 8$. Further rotation patterns were considered in Steel and McLaren (2002). In a study considering a monthly retail survey, Steel and McLaren (2000a) considered the following: 1-2-1(m), $m = 4, 8, 12$; in-for-m, $m = 1, 2, 3, 6, 12, 24, 36$ and quarterly rotation patterns in which selected businesses were included for between 3 and 36 months. Section 6 provides comments on the impact of different rotation patterns on various estimates.

Park et al. (2001) presented a general class of rotation patterns that are balanced in terms of the time in sample and rotation groups and an algorithm to construct these designs for surveys in which respondents provide data for only the current period, which they call a two-way balanced one-level design. Park et al. (2003) extended this approach to also include balance on recall time when respondents provide data for l periods to produce three-way balanced l-level designs.

4. Best linear and composite estimation

Smith (1978), Binder and Hidiroglou (1988), and Binder and Dick (1989a) reviewed estimation methods for repeated surveys under a classical approach where population means or totals are considered to be fixed quantities and a time series approach where the population means or totals are considered to be random variables generated by some stochastic process. Fuller (1990) also provided a review of some of the issues associated with estimation in repeated surveys. In this section, we will consider the classical approach and consider the time series methods in Section 7.

The sampling variance of the survey estimates can be reduced by exploiting the correlations over time between the estimates from each rotation group through various forms of least squares and composite estimation methods. These methods enable us to exploit the data from previous time periods through the correlation, effectively increasing the sample on which the estimate is based.

Estimators can be developed that use data for all N time periods and the correlation structure induced by the rotation pattern. We focus on rotation designs in which the sample in any period consists of G panels. When a panel is rotated out of the survey it is replaced by another panel. The set of panels related in this way is referred to as a rotation group. In a particular period, elementary estimates can be calculated from each panel in the survey as discussed by Gurney and Daley (1965), Smith (1978), and Wolter (1979). As in Yansaneh and Fuller (1998), we consider rotation patterns that are balanced in terms of the number of times units have been included in the survey.

Suppose we can calculate an estimate from each rotation group at time t, y_{tg}, which is unbiased for Y_t, the finite population value of interest, so that $E(y_{tg}) = Y_t$. Stacking the G rotation group estimates gives the data vector $\mathbf{y}^{RG} = (y_{11}, \ldots, y_{N1}, \ldots, y_{1G,}, \ldots y_{NG})'$. Then $E(\mathbf{y}^{RG}) = \mathbf{X}\mathbf{Y}_N$, where $\mathbf{Y}_N = (Y_1, \ldots, Y_N)'$ is the vector containing the series of population values and $\mathbf{X} = 1_G \otimes \mathbf{I}_N$ with $1_G = (1, \ldots, 1)'$, $\mathbf{I}_N$ is the N $\times$ N identity matrix, and $\otimes$ indicates the Kronker product, so that $\mathbf{X} = [\mathbf{I}_N, \ldots, \mathbf{I}_n]'$.Var$(\mathbf{y}^{RG}) = \mathbf{V}$ which depends on the rotation pattern and the population correlations.

The BLUE of $\mathbf{Y}_T$ is then $(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$, which has variance $(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}$. We need to determine or estimate $\mathbf{V}$ to calculate the BLUE. If we use a matrix of fixed values $\mathbf{W}$ instead of $\mathbf{V}$, in the calculation of the BLUE, then its variance is $(\mathbf{X}'\mathbf{W}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}^{-1}\mathbf{V}\mathbf{W}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{W}^{-1}\mathbf{X})^{-1}$. In practice, $\mathbf{V}$ has to be estimated or a working value has to be used.

The BLUE involves storing elementary estimates for the entire length of the series and inversion of a NG $\times$ NG matrix. So for a monthly series of 40 years involving eight rotation groups the matrix is 3640 $\times$ 3640 and computation of the BLUE becomes complicated as the number of periods increases. Methods that approximate the full BLUE can be used.

Yansaneh and Fuller (1998) developed a recursive regression estimator to produce an estimate equivalent to the BLUE based on estimates for m periods that avoids the complexity of the direct BLUE approach. Another way to reduce the computational burden associated with the BLUE approach is to restrict the calculation to the last m periods. They compared the sampling variances of the CPS composite estimation with the restricted BLUE with $m = 12$ and 16 and the recursive estimation for estimates of level and s month change, for $s = 1, \ldots, 12$ for employment and unemployment. Several rotation patterns were considered, including 4-8-4(8), in-for-6 and in-for-8 rotation patterns. They commented that $m = 36$ gave virtually the same efficiency as the recursive restricted estimator. The results suggest that for 4-8-4(8) rotation pattern useful gains can be made using BLUEs and that $m = 16$ is almost as efficient as the recursive regression estimator.

Bell (1999) also considered restricted window BLUEs with $m = 7$ and 1-2-1(8), 2-2-2-(8), and 4-8-4(8) rotation patterns. He found that $m = 7$ gave nearly all the available gains but that smaller windows gave noticeable higher standard errors. The evaluation considered a range of estimators for employment and unemployment: level, 1 month movement, quarterly average, nonoverlapping movement in quarterly average,

seasonally adjusted, and 1 month movement in seasonally adjusted estimates, trend estimates at the end of the series, 1 month movement in the trend estimates at the end of the series and revision of the movement in trend. The evaluation was based on a correlation model for employment that allowed correlations within PSUs. Although the model was used to construct the elements of the matrix $\mathbf{V}$, the evaluation used a delete a group jack-knife applied to real survey data to estimate variances, so any inefficiencies due to errors in the value of $\mathbf{V}$ were reflected in the evaluation.

In applying the BLUE or restricted window version, the standard survey weighting has to be applied to calculate the rotation group or panel estimates. This can lead to problems as the sample size in each rotation group will be smaller, possibly resulting in instability in the generalized regression estimator (GREG). One approach is to use less covariates in the GREG, for example, by using broader categories. Bell (2001) suggested an alternative using panel estimates obtained by applying the inverse of the selection probabilities as weights, which is the Horvitz–Thompson estimator, and then applying the standard GREG weighting to the these values. This method has been introduced for the Australian Labor Force Survey (Australian Bureau of Statistics, 2007). The revision strategy also has to be considered, since as data for additional periods are added the estimates for previous periods change.

Composite estimation also avoids the complexity of storing all the elementary estimates and inversion of large matrices. In composite estimation, the sample for the previous time periods is used along with the sample for the current period. In its simplest form, the estimate for the current period is obtained by updating the estimate of the previous period using an estimate of the change in which the matched and nonmatched samples are given different weights. However, issues of time in survey bias need to be considered (Bailar, 1975).

Composite estimation methods have been investigated extensively. Jensen (1942) considered sampling on two successive occasions with partial overlap. Yates (1949) extended the approach to sampling on more than two occasions. Patterson (1950) generalized this approach and found that matching 50% of the sample on successive occasions gave optimal gain in efficiency for estimating level and 75% was good for estimating change and still gave good efficiency for estimation of the mean. Ecker (1955) considered higher order rotation designs where information is collected at a particular occasion for two or more successive occasions. Gurney and Daley (1965) generalized the results of Patterson (1950) to a linear model framework and obtained the minimum variance linear unbiased estimators.

Composite estimation methods have mainly been applied in monthly labor force surveys. Wolter (1979) developed composite estimation for the two-level schemes formerly used in the U.S. RTS in which selected businesses report every 3 months giving data for the current and the previous months.

The basic approach is to consider two estimates for the current month; the first is the estimate calculated from the data collected in the month, y_t and the second is obtained by taking the estimate from the previous month, y_{t-1}^K and adding an estimate of change based on the panels that have not been rotated between the two periods, $y_{t-1}^K + \left(y_t^M - y_{t-1}^M\right)$, where y_t^M is the estimate for time t calculated from the matching sample. A weighted average of the two estimates is then calculated,

$$y_t^K = (1 - K)y_t + K\left(y_{t-1}^K + \left(y_t^M - y_{t-1}^M\right)\right). \tag{4}$$

This is the composite estimator which was initially used in the U.S. CPS with $K = 0.5$. An additional term is also added to further reduce variance and ameliorate the impact of the times in survey effect, which is the difference between the estimates for the current month based on the new panel, y_t^{UM}, and the panels matching to the previous month, y_t^M giving the AK estimator.

$$y_t^{AK} = (1 - K)y_t + K\left(y_{t-1}^K + \left(y_t^M - y_{t-1}^M\right)\right) + A\left(y_t^{UM} - y_t^M\right) \tag{5}$$

with $K = 0.4$ and $A = 0.2$. (Cantwell and Ernst, 1993; Gurney and Daley, 1965; Huang and Ernst, 1981).

Bailar (1975) found that for estimates of month-to-month change in the original series the use of a composite estimator gave smaller variance on average than a ratio estimator. Bailar (1975, 1978) studied the effect of rotation group bias on the estimates of level and change in the U.S. CPS. Huang and Ernst (1981) extended these results to 4-8-4(8) and 3-9-3(6) rotation patterns for the CPS for two different composite estimators. They assumed a constant variance and covariance over all observations and calculated that an optimal AK composite estimator had greater efficiency than a simple composite estimator for the level and 1 month change in the original series and also annual average.

The recursive nature of the composite estimator means that it is implicitly using the elementary estimates for the entire length of the series and therefore using some information from samples that have been rotated out of the sample. Although the implicit weights given to each elementary estimate are not the same as would be obtained from the BLUE, the efficiency can be close.

The values of A and K can be chosen to minimize the variance of y_t^K. In general, they depend on the variable and compromise values are chosen based on analysis of the impact of different values on key estimates. Relevant variance formulas are given in Cantwell (1990). Higher values are better for estimating employment levels because of the higher correlation over time. Lent et al. (1996) suggested $K = 0.4$ and $A = 0.3$ for estimating unemployment and $K = 0.7$ and $A = 0.4$ for estimating employment and the civilian labor force.

A refinement is composite weighting in which the values of A and K are chosen separately for the estimation of employment and unemployment to produce marginal totals. The standard GREG weights are then adjusted so that the estimates agree with the margins obtained from the AK estimator (Lent et al., 1996).

The generalized composite estimator (Breau and Ernst, 1983; Cantwell, 1990) is

$$y_t^{\mathrm{GCE}} = \sum_{g=1}^{G} a_g y_{tg} - \omega \sum_{g=1}^{G} b_g y_{t-1,g} + \omega y_{t-1}^{\mathrm{GCE}}. \tag{6}$$

The coefficients are constrained so that $\mathbf{1}'\mathbf{a} = \mathbf{1}'\mathbf{b} = 1$. This estimator allows different coefficients for the rotation group estimates, which gives more scope for variance reduction. Park et al. (2001) presented the general theory for choosing $\mathbf{a}$ and $\mathbf{b}$ for fixed ω when estimating four types of quantities; the current level, the change between levels, aggregates of several time periods, and changes in aggregates. Kim et al. (2005) considered generalized composite estimator for l-rotation system in which selected units report data for the one most recent periods.

The Canadian LFS also uses a version of composite estimation that can be implemented in a standard GREG estimation system by clever supplementation of the standard demographic auxiliary variables already used with new additional auxiliary variables. The method is referred to as modified regression estimation (Singh et al., 2001). Implementation is described by Gambino et al. (2001). See also Fuller and Rao (2001) and Bell (2001).

There are two versions corresponding to two choices of the auxiliary variables. For two consecutive time periods, t and $t-1$, let $M_{t,t-1}$ denote the theoretically matching sample and $U_{t,t-1}$ is the unmatched sample. The choice of auxiliary variables correspond to (i) $z_i^{(1)} = y_{i,t-1}$ for $i \in M_{t,t-1}$ and $z_i^{(1)} = \bar{y}_{t-1}^c$ for $i \in U_{t,t-1}$, where $\bar{y}_{t-1}^c$ is the composite estimate of the population mean for $t-1$or and (ii) $z_i^{(2)} = y_{i,t} + k^{-1}(y_{i,t-1} - y_{i,t})$ for $i \in M_{t,t-1}$ and $z_i^{(2)} = y_{i,t}$ for $i \in U_{t,t-1}$, where $k = \frac{\sum_{i \in M_{t,t-1}} w_i}{\sum_{i \in S_t} w_i} \approx \frac{5}{6}$ for the in-for-6 rotation pattern used. The control total used in the GREG is last month's estimate. Imputation has to be used for units in $M_{t,t-1}$ for which data on both months is not available.

Use of $z_i^{(1)}$ is good for estimating level and use of $z_i^{(2)}$ is good for estimating change. Fuller and Rao (2001) noted that use of $z_i^{(2)}$ can lead to a drift problem where the modified regression estimator can deviate from the direct survey estimate over a long time period and suggest a compromise choice of $z_i = (1-\alpha)z_i^{(1)} + \alpha z_i^{(2)}$. The value $\alpha = \frac{2}{3}$ is used as a compromise for estimating level and change for the key estimates.

The gains from using composite estimation need to be evaluated for any particular survey and variables and may be small if the sample overlap is high or the unit level correlation is low. Attention is usually focused on the estimate of level for the most recent period and the movement between the two most recent time periods, although other estimates should be considered. The gains for estimates of levels are greatest when the degree of overlap is moderate and the correlation is high. For estimates of movement, high sample overlap is still preferred.

5. Correlation models for sampling errors

Development and estimation of realistic autocorrelation models for the sampling error series taking into account the sample design is an important issue. For a particular survey using a specified rotation pattern, the issue is how to account for the sampling error in estimation and analysis. When considering different options for the rotation pattern at the design stage, we need models for the sampling correlations that allow the impact of different rotation patterns to be assessed.

The correlation structure of the sampling error series can be directly estimated from the primary survey data, if that is available. If a standard variance estimation system is available, then it may be used to calculate $\hat{\text{Var}}(y_t - y_{y-s})$ and estimate $\text{Cov}(y_t, y_{t-s})$ by $\frac{1}{2}(\hat{\text{Var}}(y_t - y_{t-s}) - \hat{\text{Var}}(y_t) - \hat{\text{Var}}(y_{t-s}))$. The resulting correlation estimates include both the effect of the rotation pattern used as well as the correlation of the population values over time. This approach was used by Lee (1990) for example.

If the sample is composed of rotation groups from which elementary estimates can be calculated, then it is useful to distinguish the correlation over time for those estimates when no rotation has occurred (first-order correlations) and those that are present when rotation has occurred (second-order correlations). The second-order correlations arise

because rotation will usually take place within a PSU or geographic area. Kumar and Lee (1983) and Park et al. (2001) showed that the second-order correlations should not be ignored when considering the variance of estimators.

Adam and Fuller (1992) used an analysis of variance approach using data for replicates to estimate the sampling autocorrelations. Pfeffermann et al. (1998) used panel estimates to calculate pseudo errors defined as $\hat{e}_t^{(j)} = y_t^{(j)} - y_t = e_t^{(j)} - e_t$ for panel j at time t, where $y_t^{(j)}$ is the elementary estimate for panel j and $e_t^{(j)}$ is the associated sampling error.

Once the correlations of the sampling errors have been estimated they can be used in conjunction with the variance estimates to produce an estimate of $\mathrm{Var}(\mathbf{y}_N)$, which can then be used to estimate the variance of any linear function of $\mathbf{y}_N$. By estimating the first- and second-order correlations separately, it is possible to assess the impact of different possible rotation patterns.

Empirical evidence shows that the autocorrelations are higher for employment than unemployment and the autocorrelations decay over time, that is as s increases. Although there can be slight peaks as $s = 12$ because of seasonal effects.

The estimates of the autocorrelations and considerations of the lags where the sample overlap occurs for the rotation pattern can be used to suggest and estimate the parameters of convenient models for the sampling errors. The analysis of sampling error is simplified if the series of sampling errors has a stable autocorrelation structure. Bell and Wilcox (1993) noted that the sample overlap occurs for finite time and if the nonoverlapping sample are independent, then the sampling errors can be approximated by a moving average model of appropriate order. Often, it is assumed that the survey errors follow a stationary autoregressive moving average (ARMA) (p, q) process of the form $\phi(B)e_t = \theta(B)a_t$, where a_t is a white noise process and B is the backshift operator, $Be_t = e_{t-1}$, and $\phi(B)$ and $\theta(B)$ polynomials of degree p and q, respectively (see Hillmer and Trabelsi, 1987). This model can be generalized to allow for a seasonal component in the survey errors. The choice for the parameters will depend on the series being investigated. This method concentrates on the estimation of the parameters of the model and includes an effect for the rotation pattern used in the surveys.

Cholette and Dagum (1994) summarized some of the ARMA models used for the sampling errors in various surveys. Hillmer and Trabelsi (1987) used $(1 - 0.8B)e_t = a_t$ for the U.S. monthly retail sales of hardware stores An ARMA(3, 6) model was used by Binder and Dick (1989b) for the in-for-6 rotation pattern used in the Canadian LFS,

$$(1 - 0.2575B + 0.3580B^2 + 0.6041B^3)e_t = (1 + 0.1847B + 0.5873B^2 - 0.3496B^3 - 0.0647B^4 - 0.0982B^5 - 0.0347B^6)a_t.$$

Trabelsi and Hillmer (1990) considered a model that accounts for the rotation pattern and composite estimation for the U.S. RTS:

$(1 - \phi_1 B)(1 - \phi_3 B^3)(1 - \phi_{12} B^{12})e_t = (1 - \theta B)a_t$, with $\phi_1 = 0.75$ to account for the composite estimation used for this survey. The remaining parameters depend on the series, which can be estimated from estimates of the sampling autocorrelation. They considered $\phi_3 = 0.3,\ 0.6$ and $\phi_{12} = 0.3,\ 0.6$. Bell and Hillmer (1990) estimated $\phi_3 = 0.635,\ 0.664$, $\phi_{12} = 0.723, 0.714$, and $\theta = -0.130, -0.134$ for retail sales of eating and drinking places, respectively, after a log transformation. Bell and Wilcox (1993)

also considered a model for the U.S. RTS $(1-0.75B)(1-0.70B^3)(1-0.75B^{12})e_t = (1+0.10B)a_t$. These models account for the use of composite estimation and the rotation design for the RTS, which involved three panels each providing data every 3 months leading to sample overlap at lags at multiples of 3 months. The parameters can be estimated from the autocorrelation using least squares methods. Hausman and Watson (1985) considered a survey error model ARMA(1,15) for the U.S. CPS based on the 4-8-4(8) rotation pattern and use of composite estimation. Pfeffermann and Tiller (2006) used an autoregressive (AR)(15) process for the sampling errors in the U.S. CPS to approximate the ARMA(2, 15) process that results from the sum of an moving average (MA)(15) and AR(2) process, the former arising from the rotation pattern leading to sample overlap up to lag 15 and the latter arising from the rotation of the sample within the same census tracts.

To examine the impact of different rotation patterns, a model for the autocorrelation that explicitly includes the sample overlap is required so that the impact of different overlap factors can be gauged. For a single-stage sample, Steel (1996) used the model $\mathrm{Corr}(y_t, y_{t-s}) = R(s) = k(s)h(s)\rho(s)$, where $k(s)$ is the theoretical sample overlap between the samples at t and $t-s$. The factor $h(s) = \alpha\beta^s$ reflects the reduction in overlap because of nonresponse and movement of households and $\rho(s)$ is the unit level correlation, which was estimated from gross flows tables using matching households in the U.K. LFS for employment and unemployment. The assumption that the variances and autocorrelation of the sampling error series are constant implies that no major changes in the sample design or population structure occur over the length of the series. The model implies no correlation when there is no overlap of the sample at the household level.

Assume that the estimator at time t is, at least approximately, the average of the estimates from each rotation group $y_t = \frac{1}{G}\sum_{g=1}^{G} y_{tg}$ and the estimates from different rotation group are independent. Panels from the same rotation group will not necessarily be independent because rotation occurs with the same PSU. Let $\mathrm{Corr}(y_{tg}, y_{(t-s)g}) = R^{NR}(s)$ if no rotation has occurred and $\mathrm{Corr}(y_{tg}, y_{(t-s)g}) = D(s)$ if rotation has occurred, then $\mathrm{Corr}(y_t, y_{t-s}) = R(s) = D(s) + k(s)\left(R^{NR}(s) - D(s)\right)$. Scott et al. (1977) gave a similar model.

Bell (1999) estimated $R^{NR}(s)$ and $D(s)$ for the Australian LFS using the panel estimates. For values of $s > 8$ he used a model obtained by least squares estimation for estimates of employment and unemployment. Lee (1990) provided values of $R^{NR}(s)$ and $D(s)$ for the Canadian LFS using panel estimates. Adam and Fuller (1992) and Gunlicks et al. (1997) gave results for the U.S. CPS.

Changing variance and covariance have been considered by Bell and Hillmer (1990), Tiller (1992), Cholette and Dagum (1994), and Bell and Kramer (1999). For example, changing variances can be accommodated by starting with a series of survey errors $\mathbf{e}_N^* = (e_t^*, \ldots, e_N^*)'$ with constant variance, and the same autocorrelation structure as $\mathbf{e}_N = (e_1, \ldots, e_N)'$. The original stable survey errors can be considered transformed $e_t = g_t e_t^*$, so that $\mathrm{Var}(\mathbf{e}_N) = \mathbf{G}\mathrm{Var}(\mathbf{e}_N^*)\mathbf{G}'$, where $\mathbf{G} = \mathrm{diag}(g_t)$. Breaks in the sample due to the introduction of a new sample imply zero correlation between estimates before and after the break and is reflected in a block diagonal structure for $\mathrm{Var}(\mathbf{e}_N)$. McLaren (1999) investigated variation in the unit level correlation at lag 1 using gross flows from the Australia LFS, based on the relationship that for estimates of a proportion using a simple random sample $\mathrm{Cov}(y_t, y_{t-1}) = P_{t,t-1} - P_t P_{t-1}$, where P_t is the proportion in

the category of interest as time t and $P_{t,t-1}$ is the proportion in the category at times t and $t-1$. For employment, the correlation varied between 0.86 and 0.92, and for unemployment, it varied from 0.57 to 0.69 over a 13-year period.

6. Rotation patterns and sampling variances

In general, the higher the sample overlap between two time periods the lower the standard error on estimates of change between them. For averages of estimates, positive correlation between the survey estimates involved will increase the sampling variance. It is better to average uncorrelated estimates, which can be obtained from independent or nonoverlapping samples. If both averages and differences of estimates are of interest, then the relative importance of each type of estimate has to be considered and the impact of different options assessed on both types. To assess the impact of different rotation patterns on various estimates, some information or assumptions about the covariances involved are needed.

In looking at different rotation schemes and the resulting overlap patterns, a range of estimates that may be used by analysts can be considered. For example, in a monthly survey, the following may be of interest:

- Monthly levels, y_t.
- The change in the monthly level for months s months apart $y_t - y_{t-s}$. Particular interest might be in $s = 1$, 3 and 12.
- Average of 3 months' level estimates, $\tilde{y}_t^{(3)}$.
- Change between 3 monthly averages with centres s months apart, $\tilde{y}_t^{(3)} - \tilde{y}_{t-s}^{(3)}$. Setting $s = 1$ gives change in the overlapping "rolling" estimates, s = 3 gives the change in nonoverlapping 3 monthly periods, and $s = 12$ this gives the change between 3 monthly averages a year apart.
- Average of 12 months' level estimates, $\tilde{y}_t^{(12)}$.
- Change in the 12 month averages a year apart $\tilde{y}_t^{(12)} - \tilde{y}_{t-12}^{(12)}$.

The variance of these different estimates will be determined by the overlap pattern and the correlations between estimates.

Blight and Scott (1973) looked at optimal design in terms of minimizing the consecutive difference of level estimates and found that complete overlap was best. Based on where the sample overlap is concentrated, we should expect that the in-for-8 rotation pattern would be good for estimating changes for s = 1, 2, 3. The 4-8-4(8) pattern should be good for changes when s = 1, 12 and better than in-for (8) for quarterly averages. A 1-2-1(8) pattern would be poor for changes when s = 1, 2, but good for changes when s = 3, 6 and reasonable for change for s = 12, provided the individual level correlation at lag 12 is high. Results suggest the lag 12 correlation is low for unemployment and moderate for employment. This pattern is very good for three average monthly and changes in them using overlapping or nonoverlapping averages. The 2-10-2(8) rotation pattern gives theoretical overlap of k(s) = 0.5 for s = 1,12 and this level of monthly overlap is good for composite estimation of level. Bell (2001) noted that 2-2-2(8) provides a compromise between estimating short-term movements and analyses looking at medium or long-term movements as measured by quarterly averages of trend estimates. McLaren and Steel (2000) gave similar results.

To compare the effects of different rotation schemes on each type of estimate Steel (1996, 1997) gave the ratio of the variance of each estimate relative to the variance of the estimate of level for the different rotation schemes considered for a monthly UK LFS for estimates of unemployment and employment, respectively. These results show that for estimating monthly change the higher the monthly overlap the better, although the further gains diminish as the overlap increases and, because of the higher monthly correlation, the gain from having monthly overlap is higher for employment estimates. For estimating quarterly averages and the changes in them the 1-2-1(m) patterns are better than the *in-for-m* patterns because they result in independent monthly samples within the quarter and the overlap is concentrated at a 3 month lag. The 4-8-4(8) design is worse than the in-for-6 design for monthly change, but better for annual changes in monthly estimates.

To decide on a rotation pattern we must decide on the relative importance of different estimates. Steel (1997) compared the in-for-6, 1-2-1(5), and 4-8-4(8) in more detail. For the monthly change in unemployment, the 1-2-1(5) and 4-8-4(8) schemes had variance 2.30 and 1.13 times larger than the in-for-6 scheme. For the 3 month change in the quarterly average unemployment, the in-for-6 and 4-8-4(8) schemes had variance 1.98 and 2.3 times larger than the 1-2-1(5) scheme. For the 12 monthly change in unemployment, the 1-2-1(5) and in-for-6 schemes had variance 1.10 and 1.16 times larger than the 4-8-4(8) scheme.

The impact of rotation pattern and BLUE or composite estimation can be considered simultaneously. See results from Yansaneh and Fuller (1998), Bell (1998, 2001), and McLaren and Steel (2001).

7. Time series methods for estimation in repeated surveys

Composite estimation and BLUE seek to exploit the correlation structure in the sampling errors and treat the finite population values Y_t as fixed. In some situations, there may be reasons to go beyond treating the population values Y_t as fixed and postulate a stochastic time series model for them. Even if there is no direct interest in time series analysis, using a time series model may help in estimation of Y_t and changes in these values, such as $\Delta^{(s)}Y_t$, which is the focus of this section.

Blight and Scott (1973), Scott and Smith (1974), and Scott et al. (1977) developed the time series approach to estimation of Y_t from repeated surveys. Jones (1980) unified the time series approach into a general form for the minimum mean squared estimator (MMSE) using elementary estimates, which involves large matrix calculations. A key result is that if $E[\mathbf{Y}_N] = \boldsymbol{\mu}_N$, then the MMSE of $\mathbf{Y}_N$ based on $\mathbf{y}^{RG}$ is $\mathbf{Y}_N = \boldsymbol{\mu}_N + [\mathbf{X}'\mathbf{V}^{-1}\mathbf{X} + \mathrm{Var}(\mathbf{Y}_N)]^{-1}\mathbf{X}'\mathbf{V}^{-1}(\mathbf{y}^{RG} - \mathbf{X}\boldsymbol{\mu}_N)$. Binder and Hidiroglou (1988) simplified this approach by using SSMs and the Kalman filter to allow efficient computation.

In time series analysis, the usual approach is to consider the population values to follow a basic structural model (BSM) with components corresponding to the trend-cycle, L_t, seasonal effects, S_t, and an irregular term, I_t. For an additive model, this can be written as

$$Y_t = L_t + S_t + I_t, \tag{7}$$

which leads to

$$y_t = L_t + S_t + I_t + e_t. \tag{8}$$

Multiplicative models can also be used, especially for economic data. A detailed description of the BSM and associated Kalman filter is given in Harvey (1989). See also Binder and Dick (1989a). Feder (2001) reviewed the application of the state-space approach to repeated surveys. The trend is often modeled by

$$L_t = L_{t-1} + R_{t-1} + \eta_t, \text{ where } \eta_t \sim N(0, \sigma_\eta^2) \tag{9}$$

$$R_t = R_{t-1} + \zeta_t, \text{ where } \zeta_t \sim N(0, \sigma_\zeta^2). \tag{10}$$

Which describes a local linear trend with local rate of change R_t. Seasonality can be modeled as either a trigonometric model or dummy variable model. All stochastic terms in the population BSM are assumed to be independent of one another and serially independent. Dependence between the population values is reflected in the Eqs (9) and (10).

The general multivariate formulation of a SSM consists of an observation equation relating the vector of estimates $\mathbf{y}_t$ to an unobserved state vector $\boldsymbol{\alpha}_t$:

$$\mathbf{y}_t = Z_t \boldsymbol{\alpha}_t + \boldsymbol{\varepsilon}_t. \tag{11}$$

And a transition equation that describes how the state vector evolves over time:

$$\boldsymbol{\alpha}_t = T_t \boldsymbol{\alpha}_{t-1} + \boldsymbol{\eta}_t, \tag{12}$$

where $\text{Var}(\boldsymbol{\varepsilon}_t) = \mathbf{H}_t$ and $\text{Var}(\beta_t) = \mathbf{Q}_t$. The error vectors are assumed to be serially uncorrelated, which is an important feature of the approach. For a repeated survey with sample overlap the sampling errors are autocorrelated and hence to use the SSM approach the sampling errors are usually included in the state vector. For secondary analysis based on aggregated series the form of the model for the sampling errors will be based on consideration of the rotation pattern and estimation and modeling of variances and correlations as described in Section 5. For elementary estimates, the panel estimates can be used and the state vector includes the panel sampling errors (see Pfeffermann, 1991).

Once a model has been expressed as a SSM, it can be analyzed by applying the Kalman filter and the Kalman smoother. The Kalman filter provides the optimal estimator of the state vector using data up to time t. The Kalman smoother provides the optimal estimator of the state vector for previous time periods.

Equation (11) has been written in terms of the estimates and since the observation errors are assumed to be independent over time, one approach is to include the sample errors in the state vector. For example, Feder (2001) considered the case when the sampling error follows an AR(1) process, so that $e_t = \rho e_{t-1} + \delta_t$. For a univariate quarterly series, the estimate is a scalar, the state vector is $\boldsymbol{\alpha}_t = (L_t, R_t, S_t, S_{t-1}, S_{t-2,} e_t)'$,

$$Z_t = (1, 0, 1, 0, 0, 1)', \boldsymbol{\eta}_t = (e_t, \eta_t, \zeta_t, 0, 0, \delta_t)', \text{ and } T_t = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & -1 & -1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \rho \end{bmatrix}.$$

The form of model for the sampling errors will depend on the overlap pattern, the sample design and the individual level covariances, as discussed above. AR(1) models were used by Blight and Scott (1973) and AR(2) by Pfeffermann et al. (1998). Feder (2001) gave a multivariate example in which the estimates of the number of households in four size categories from the Canadian LFS are of interest and the sampling errors are assumed to follow a vector AR(3) process. Again, the survey errors are included in the state vector.

The state-space approach produces estimates of each of the components in (7) from which estimates can be produced of the population value, $\hat{Y}_t = \hat{L}_t + \hat{S}_t + \hat{I}_t$. They can also be to produce estimated of the local trend, $\hat{L}_t$ and the seasonally adjusted estimates, $\hat{S}A_t = \hat{Y}_t - \hat{S}_t = \hat{L}_t + \hat{I}_t$. In each case, the estimated sampling error has been removed. Also, the Kalman filter produces estimates of the variances of each of the estimated components. Pfeffermann and Tiller (2005) noted that replacing parameters by estimates in the theoretical prediction mean square error (PMSE) of state vector predictors underestimates the true PMSE and propose parametric and nonparametric bootstrap methods that account for the use of estimated model parameters.

A SSM is used to produce monthly employment and unemployment estimates by the U.S. Bureau of Labor Statistics (Tiller, 1992). The population model is expanded to include a covariate term $\mathbf{x}_t'\beta_t$, where β_t is modeled as a random walk. The sampling error is assumed to be an AR(15) process, which is an approximation of the sum of a MA(15) process and an AR(2) process. This model arises from the design of the CPS in which there is overlap at lags 1–3 and 9–15. The AR(2) process arises from the fact that rotation of the sample occurs within the same census tract. Harvey and Chung (2000) considered an example of estimating the number of people unemployed using the count of people claiming unemployment benefit as a covariate.

In some situations independent information is available, for example, from a census or large annual survey, which can be used to used to benchmark a monthly or quarterly survey (see Durbin and Quenneville, 1997; Hillmer and Trabelsi, 1987). Pfeffermann and Tiller (2006) consider benchmarking of monthly model-based estimates for states and census divisions obtained from the U.S. CPS to annual design-based estimates for the census divisions.

Bell (2004) described a general time series model that involves a regression term and several unobserved components called a RegComponent model, $y_t = \mathbf{x}_t'\beta + \sum_{c=1}^{C} h_{ct}\varepsilon_{ct}$, where h_{ct} are known scale factors, the independent unobserved components ε_{ct} follow ARIMA models of the form $\phi_c(B)\delta_c(B)z_{ct} = \theta_c(B)a_{ct}$, δ_c is a difference operator, and a_{ct} is a white noise process. This model includes as special cases the regression ARIMA model, (C = 1) and the BSM (C = 3). Sampling error can be included as a further component but the parameters of the sampling error component are estimated separately from estimated sampling variances and autocorrelations. These parameters are then held fixed as the RegComponent model is fitted to the observed data. A similar approach is suggested by Pfeffermann et al. (1998) because of the potential identification problem between the sampling error components and other components in the BSM. The U.S. Bureau of the Census's REGCMPT computer program uses a state-space formulation for estimation of this model.

The use of panels in a survey induces the correlation structure of the sampling errors but can also be used to estimate and remove their effect. Pfeffermann et al. (1998) used the idea of pseudo panel-survey errors to estimate and then remove the effect of

the correlated sampling error. They noted that the use of a panel design may introduce spurious short-term trends in the observed series. Applying this method to estimates, the autocorrelations of the sampling errors for the Australian LFS data they decided that an AR(2) model is a good approximation to the survey errors. The standard X11 trend filter, which is based on assumption of uncorrelated sampling errors, reacts to the short-term trends that such a process produces. Pfeffermann et al. (1998) proposed a two-step procedure. The first step estimates the parameters of the model for the survey errors by using the sampling variances and autocorrelations from the panel estimates and then solving the Yule–Walker equations to estimate the parameters of the AR(2) process and the associated residual variance. In the second step, the parameters of the BSM for the population values are estimated by maximum likelihood using the Kalman filter with the parameters of the survey error model held fixed at their estimated values. This approach enables the separation of the autocorrelation structure of the sampling errors and the true series by using the sample to estimate the former. This approach is used rather relying on the standard estimation approach for BSM that includes the sampling errors because the form of the autocorrelation of the sampling errors is close to that associated with the evolving trend, which can lead to problems for identification of parameters. The empirical results show how the resulting trend series is smoother and not affected by the spurious short terms trends induced by the correlated survey errors. Pfeffermann and Tiller (2006) developed an alternative filtering algorithm for the SSM, which allows for correlated measurement error by directly including the covariance matrix of the errors in the procedure. For filter-based methods, McLaren and Steel (2001) suggested modifications that directly incorporate the covariance matrix of the sampling errors.

Most surveys will produce a range of estimates leading to multivariate methods. A particular case is when there is interest in a set of estimates that are categories of the variable, such as employment status, so that the estimates are proportions that sum to one. Separate analysis of the component series ignores this constraint. Brunsdon and Smith (1998) and Silva and Smith (2001) considered this issue and developed methods based on the additive log transformation. Brunsdon and Smith (1998) used a Vector ARIMA (VARMA) modeling approach and Silva and Smith (2001) used BSM and VARMA approaches.

There is a strong thread relating the approaches discussed thus far. When the population values are fixed, the autocorrelation of the sampling errors can be exploited to produce efficient estimates of these values using the BLUE. To avoid the large matrix inversion and storage of a large number of estimates, recursive methods, such as composite estimation, recursive regression estimation, or restricted window BLUE, can be used. When the true values are regarded as random variables extra components of variability are included, this leads to the MMSE. The use of SSM allows the estimation to be undertaken using recursive methods in a way that incorporates the autocorrelation structure of the sampling errors. There are four approaches proposed to account for the autocorrelations in the sampling errors: (i) incorporate the sampling errors in the state vector in a secondary analysis based on $\mathbf{y}_N$ using a model for the sampling errors; (ii) incorporate the panel sampling errors in the state vector in an analysis based on elementary estimates; (iii) separately estimate the autocorrelations from panel estimates and hold them fixed in an analysis using $\mathbf{y}_N$; and (iv) place the covariance matrix of the sampling errors directly in the estimation algorithm.

In this section, we have assumed that Y_t is the target of inference. More generally, L_t and SA_t will be of interest as discussed in Section 8.

8. Seasonal adjustment and trend estimation

This section considers situations in which there is direct interest in the time series structure of the series $Y_t, t = 1, \ldots N$ and if we had these values we would have undertaken time series analysis of them. The sampling error is then a source of measurement error with a correlation structure arising from the overlapping sample design, which if ignored can lead to biases in the estimates of the parameters and components of the time series model.

For monthly or quarterly surveys, seasonal adjustment may be used to remove the impact of regular systematic- and calendar-related influences. Producing seasonally adjusted estimates helps assess the underlying direction or trends in the series by allowing comparability from month-to-month, identify turning points as well as assisting in short-term forecasting and in relating time series to other series or extreme events. The assessment of trends is usually based on seasonally adjusted series and may be done informally or through the calculation of trend estimates.

Decomposing a time series as in the BSM given by (7) can highlight important features of the data, helps in monitoring time series, aids in forecasting and making policy decisions, as we can separate out the different components and gauge how each term is contributing. Decomposition will not necessarily be unique and can include other components, such as trading day components.

The trend component reflects the underlying movement in a time series. It can be due to influences, such as population growth, price inflation, and general economic development and contains the long-term business cycle. There is no unique definition of trend.

Seasonality is any effect that is reasonably stable with respect to annual timing, consistent in direction and of predictable magnitude. It may be caused by the timing of public holidays, calendar events, and weather. Seasonality can evolve slowly over time due to social and economic changes and government policy, and it can also evolve abruptly over time, resulting in a seasonal break.

Time series values tend to oscillate around a general trend level. The irregular component consists of short-term fluctuations, neither systematic nor predictable. On occasions, the degree of irregularity is unusually large, resulting in extreme values. Irregularity can be caused by real world events in relation to different holidays, inclusion of additional supplementary surveys. If the sample error is ignored, the estimated irregular can be affected by the nature of sampling, which may be related to the rotation pattern.

There are four decomposition models which are generally used: additive, log-additive, multiplicative, and pseudoadditive. The model which gives the more stable seasonal component is generally the more appropriate one to fit to the series. We will focus on the additive decomposition.

The additive model assumes that the seasonal and irregular components are independent of the trend component. The trend of an additive series can fluctuate, but the magnitude of seasonal spikes remains about same. The seasonal component remains stable from year-to-year and seasonal fluctuations average out to zero over the year. It is

used if the seasonal effects are the same from year-to-year or change slowly over time. In a multiplicative model, as trend increases, the amplitude of the seasonal influences increases and the variance of irregular component is directly proportional to the seasonal and trend cycles.

The seasonally adjusted data can be produced by estimating and removing the systematic calendar-related effects from the original data. In the additive case, the estimated seasonally adjusted value $\hat{S}A_t = Y_t - \hat{S}_t \approx L_t + I_t$, if the seasonal factor has been estimated well by $\hat{S}_t$. Seasonally adjusted estimates should contain only trend and irregular influences. Trend estimates are closely related to the seasonally adjusted data but have tried to eliminate or reduce the influence of the irregular influences.

There are two broad approaches to seasonal adjustment and estimating trends; model- or filter-based methods. Model-based approach involve the BSM, ARIMA modeling, and SSMs. These approaches are implemented in the TRAMO-SEATS programs (Gomez and Maravall, 1997), the STAMP program (Koopman et al., 2000), and REGCOMPT (Bell, 2004). Nonparametric filter-based methods as embedded in X11 and X12 (Findley et al., 1998).

In model-based methods, the unobserved trend, seasonal and irregular components of the original series are modeled. Parameter estimates for each of the components can be estimated simultaneously, and the irregular component is assumed to be white noise in the model-based approach and the trend component follows a local linear model. Business cycles can also be modeled with the trend. The seasonal component is a stochastic process with its own noise and allows evolving seasonality. The correlation structure of the components is well described by the model. We can explicitly include particular models for the impact of the sampling error produced by the rotation pattern and other known effects, as considered in Section 7.

Model-based seasonal adjustment is based on formulating a model, such as an ARIMA model or SSM. Often, trigonometric expressions are used for seasonality. The ARIMA-based approach to seasonal adjustment was developed by Burman (1980), and Hillmer and Tiao (1982) who gave the details of the canonical decomposition of ARIMA models into component form. This results in the model $y_t = \mathbf{x}_t'\beta + z_t$, where the independent unobserved components z_t follow ARIMA models of the form $\phi(B)\delta(B)z_t = \theta(B)a_t$ and δ is a difference operator and a_t is a white noise process. This is the basis of the signal extraction in ARIMA time series (SEATS) program, which decomposes the observed series into trend, seasonal, and irregular components. SEATS is usually used in conjunction with the time series with ARIMA Noise, Missing Observations and Outliers (TRAMO) program. These are used extensively by EUROSTAT and the European Central Bank.

Seasonal adjustment is often carried out by government statistical agencies using X11 (Shiskin, 1967), or its extensions, X11ARIMA (Dagum, 1988) and X12ARIMA (Findley et al., 1998). The ABS produces and publishes trend estimates, obtained by applying Henderson moving averages, for all its major series (ABS, 1993). It encourages users to base their analysis of the series on these estimates (Linacre and Zarb, 1991). Other government agencies produce trend estimate using a variety of method (Knowles, 1997). However, many statistical agencies do not publish trend estimates because of the revisions that may be made to them as estimates for later time periods are added to the series.

In filter-based methods, filters (moving averages) are applied in an iterative way to decompose the original series into its trend, seasonal, and irregular components and

usually ignore the sampling error. The estimated irregular component does not always display white noise characteristics particularly if the data comes from a survey with sample overlap. Pfeffermann (1994) exploited this to produce estimates of the variance of the seasonally adjusted estimates.

In the filter-based approaches, the trend component is defined as having cycles longer than a certain length. The seasonal component is the band around the seasonal harmonics. Seasonal adjustment aims to remove all spectral power at the seasonal frequencies, and the irregular component is defined as the residual or what is left over from the original once the trend and seasonal components have been removed.

Filter-based approaches use standard ratio to moving average approach. The basic steps are initial estimate of the trend, remove the trend leaving seasonal and irregular, then estimate the seasonal component. Seasonality cannot be identified until the trend is known. An estimate of the trend cannot be found until the series has been seasonally adjusted and so an iterative approach is adopted, as in the X11 methodology.

Extremes can distort the estimation of the components of the time series. Correcting extremes improves the estimation of the time series components of trend, seasonal, and irregular. There are different types of extremes that are represented in a time series. Variability associated with sampling can cause extreme values to occur for estimates based on small samples. In practice, extremes caused by sampling error may occur for more than one time period. This is because of sample rotation where that sample remains in the survey for a number of time periods.

A key issue is that the time series model for the population model will involve autocorrelation and the sample design, in particular the pattern of sample overlap, will lead to autocorrelation in the sampling errors. If the autocorrelation in the sampling error is ignored, it will affect the modeling of the population model and may partly appear in the estimated trend component. Hence there is a need to account for the correlation structure of the sampling error in the time series analysis, especially if sampling error is substantial. In model-based approaches, this can be done by including a component for the sampling error, the parameters for which are estimated from the sampling variances and autocorrelations, or incorporated in the state vector in a SSM. In X11, the sampling error is often ignored, although Pfeffermann et al. (1998) suggested separately estimating the sampling errors using a SSM and subtracting them for the original series and then applying X11.

Although X11 was developed as a somewhat ad-hoc, nonparametric approach, Maravall (1985) found that X11 could be approximated under a BSM with appropriate parameters. Burridge and Wallis (1985) also considered a Kalman filter approximation to X11 and used this to calculate the variance of seasonally adjusted series. Knowles and Kenny (1997) compared a Kalman filter-based approach to X11 for trend estimation and found that by appropriate choice of parameters the Kalman filter could approximate the Henderson trend filters closely. At the ends of the series, the Kalman filter gave a higher noise variance and they concluded that the Henderson moving average (HMA) were more appropriate as trend filters.

9. Variance estimation for seasonally adjusted and trend estimates

Variance estimates are needed to measure the uncertainty for any type of estimate and produce confidence intervals. There is well-established theory for calculating variance

estimates for original estimates derived from a sample survey. We also need to consider the variance of estimates obtained by time series analysis, in particular seasonally adjusted estimates and trend estimates.

Variances are easily obtained from model-based approaches, for example, under the BSM and the Kalman filter, the variance of the time series component estimates are produced as part of the estimation process. Many agencies use a filter-based approach and so much of the attention has been focused on methods for calculating variances for this type of approach.

For filter-based seasonal adjustment, such as X11, a simple approach treats the population values as fixed. If the X11 seasonal adjustment process is approximated by a set of weights, then $\hat{S}A_{t|N} \approx \mathbf{w}'_{t|N}\mathbf{y}_N$, where $\hat{S}A_{t|N}$ is the seasonally adjusted estimate to time **t** based on a time series of data going to time T, where $t \leq N$. This estimate can be treated as an estimate of $\mathbf{w}'_{t|N}\mathbf{Y}_N$, which is the seasonally adjusted value that would be obtained if there was no sampling error. This implies that the error in $\hat{S}A_{t|N}$ is taken to be $\mathbf{w}'_{t|N}(\mathbf{y}_N - \mathbf{Y}_N) = \mathbf{w}'_{t|N}\mathbf{e}_N$ and the sampling variance is $\mathrm{Var}(\hat{S}A_{t|N}) \approx \mathbf{w}'_{t|N}\mathrm{Var}(\mathbf{y}_N)\mathbf{w}_{t|N}$. A practical issue is the estimation of the variance matrix $\mathrm{Var}(\mathbf{y}_N)$, which under this approach is the covariance matrix of the sampling errors as discussed in Section 5. Wolter and Monsour (1981) used this approach for the sampling variance of the seasonally adjusted series, and it can be applied to trend estimates by using suitable definitions of the weights. McLaren and Steel (2000) used this approach to assess the impact of different rotation patterns for both seasonally adjusted and trend estimation. In this framework, it is simple to consider variance estimates for different measures, for example, for 1 month movement, by redefinition of the weight matrix. There are many approaches to linearization of the X11 process, including Young (1968), Wallis (1974), Cleveland and Tiao (1976), Ghysels and Perron (1993), and Dagum et al. (1996).

This approach is estimating the sampling variance of the $\hat{S}A_{t|N}$ and does not include the effect of irregular terms or other components of Y_t or revisions as estimates for later time periods become available. More generally, the variance arising from the seasonal and trend components can also be included. Bell and Kramer (1999) also included the revision error when assessing the variations of $\hat{S}A_{t|N}$.

Pfeffermann (1994) considered the variance of $\hat{S}A_t - SA_t$, $\hat{S}A_t - L_t$, and $\hat{L}_t - L_t$, where $SA_t = Y_t - S_t$ and $\hat{S}A_t = y_t - \hat{S}_t$, assuming $\hat{S}_t$ and $\hat{L}_t$ are unbiased, so that they extract the corresponding components with no error. The method uses a linear approximation to seasonal adjustment that developed an estimate of sampling error directly from the estimated time series using the estimated irregulars, which includes the variability due to the sampling error and also the irregular component of the time series. This approach does not explicitly model the time series of the sampling errors and has the advantage that it does not require estimation of the sampling covariances, only the sampling variance. Scott et al. (2004, 2005) provided a summary of the approach and empirical results. For trend estimation, the same approach can be used.

Bell (2005) reviewed issues and approaches to variances for seasonal adjustment and considered alternative definitions for model- and X11-based seasonal adjustment. When the stochastic nature of the components of the time series are recognized, then the relevant definition of the error associated with seasonal adjustment needs to be considered and include $\hat{S}A_t - (y_t - S_t) = S_t - \hat{S}_t$ and $SA_t - (\hat{L}_t + \hat{I}_t)$.

A spectral approach has also been considered by Chen et al. (2003).

Despite the progress that has been made the most appropriate way to achieve standard errors for seasonally adjusted and trend estimates for filter based is still being researched. Currently, no government agency which uses the filter-based approach is publishing variance estimates for seasonally adjusted or trend estimates.

10. Rotation patterns and seasonally adjusted and trend estimates

Different rotation patterns produce different correlation structures in the sampling errors over time, which can affect the properties of seasonally adjusted and trend estimates. Using linear approximations to the seasonally adjusted and trend estimates produced by X11 and X11 ARIMA McLaren and Steel (2000), considered the impact of different rotation patterns on the sampling variance of seasonally adjusted and trend estimates obtained by applying Henderson moving averages to the seasonally adjusted estimates. Steel and McLaren (2002) provided further evaluation of the impact of different rotation patterns. They found that the popular rotation designs that focus on obtaining high sample overlap between adjacent time periods are good for the estimates of changes between consecutive periods in the original and seasonally adjusted estimates. However, for the level and 1 month and 3 month change in trend estimates and the level and 3 month change in seasonally adjusted estimates, rotation pattern with little or no monthly overlap, such as 2-2-2(8) and 1-2-1(8), give considerable gains. Steel and McLaren (2000b) reached similar conclusions for the mean squared error of the revision of trend estimates. These results arise because trend estimation effectively involves averaging over several time periods and examining changes over more than the just the two latest periods. We noted in Section 2 that it is beneficial to have no little or no sample overlap within the relevant window used to calculate an average and that for examining change over longer periods the sample overlap is best concentrated at longer lags.

The Henderson moving averages were derived assuming that the original series has an independent error structure. McLaren and Steel (2001) developed an approach for constructing trend filters that takes into account the correlation of structure of the series and showed that the benefit of such designs for trend estimates is still applied.

Sample Surveys: Inference and Analysis, Vol. 29B
ISSN: 0169-7161
© 2009 Elsevier B.V. All rights reserved
DOI: 10.1016/S0169-7161(09)00234-X

The Analysis of Longitudinal Surveys

Gad Nathan

1. Introduction

In the past few decades, economists, sociologists, and other social scientists have become increasingly interested in longitudinal surveys and their analysis, in their attempts to understand the dynamics of economic and social processes. Previously, single-time cross-sectional surveys and their analysis formed the primary basis for empirical investigations in the social sciences. Longitudinal surveys, in which the same units are investigated on several occasions, over extensive periods of time, are expensive undertakings and are complex operationally and methodologically. Even when longitudinal data are theoretically available, for example, from repeated panel surveys, not designed for longitudinal analysis, such as the Labor Force Survey in the United Kingdom and the Current Population Survey in the United States, technical and methodological problems considerably reduce their usefulness for longitudinal analysis at the micro- or individual level. In addition, their overall time span for a single unit is relatively short (usually no more than about 2 years), and their panel design is primarily geared to increase the efficiency of cross-sectional estimates and estimates of change at the aggregate or macro level, rather than to that of the analysis of gross changes and flows or of other developments over time at the individual level.

However, in recent years, longitudinal surveys have become of prime importance as a basis of empirical research in the social sciences. They are now being used increasingly for longitudinal analysis, and in many cases, longitudinal surveys are carefully designed to permit the derivation of sophisticated analyses of the long-term dynamics of social and economic processes. Thus, the Panel Study of Income Dynamics (PSID), carried out by the Survey Research Center, Institute for Social Research, University of Michigan, since 1968, is a longitudinal study of a representative sample of U.S. individuals and the family units in which they reside. It emphasizes the dynamic aspects of economic and demographic behaviors. Similarly, the British Household Panel Survey (BHPS), carried out on a continuous basis since 1991 by the Economic and Social Research Council (ESRC) Research Centre on Micro-Social Change at Essex University, has as its aims "to further understanding of social and economic change at the individual and household level in Britain, and to identify, model and forecast such changes and their causes and consequences in relation to a range of socio-economic variables" – University of Essex

(2006). Similar household panel surveys are the U.S. Survey of Income and Program Participation (SIPP), the Canadian Survey of Labor and Income Dynamics (SLID), and the European surveys conducted under the European Union Statistics on Income and Living Conditions (EU-SILC) regulations – further details on the design of these surveys are provided by Kalton (Chapter 5). For some insight on the various analytical uses of panel surveys, in the social sciences see Solon (1989) and Heckman and Robb (1989).

Although in some cases, longitudinal surveys are designed to also produce cross-sectional estimates for each point of time, we consider, in this chapter, only the analysis of the dynamic aspects of longitudinal surveys, under a model-based approach. It should be noted that the terms "longitudinal surveys" and "panel surveys" are often used interchangeably (sometimes due to differences in U.S. and U.K. usage). In this chapter, we consider longitudinal analysis as that which relates to any data collected for the same units over a series of time points (or even continuously), usually over a considerable length of time. Although the emphasis is on the analysis of data from sample surveys, we shall also consider methods of analysis developed or used for other types of data (e.g., administrative data, census data, or experimental data), in as far as they can be applied also for the analysis of survey data. Other chapters in this volume deal with various related aspects of longitudinal surveys and panel surveys. Thus, Kalton (Chapter 5) deals with the design and analysis of panel surveys, primarily those with short periods of repetition and rotating panels, and their use for cross-sectional estimation and for estimation of change, under a design-based approach. The chapter also deals with special issues of panel surveys, such as the effects of changing modes of collection and weighting adjustments for attrition and wave nonresponse. Steel and McLaren (Chapter 33) consider inference over time on the basis of repeated survey data, primarily under a design-based approach and using time series methods. Finally, Singh (Chapter 35) describes some approaches and recent developments in the analysis of longitudinal categorical data, under a model-based approach, as well as joint modeling for cross-sectional and longitudinal analysis of categorical data.

In the following section, we consider the various types of longitudinal surveys and the problems they pose. Section 3 reviews general and specific models used for the analysis of longitudinal data, primarily those applicable to survey data. In Section 4, we consider some model-based methods for the treatment of wave nonresponse, attrition, and misclassification errors. Finally, Section 5 deals with the effects of complex and informative sample design on longitudinal analysis and their treatment for purposes of analysis.

2. Types and problems of longitudinal surveys

Longitudinal data can be derived from a wide range of sources and by a variety of collection methods. In the following, we shall examine the main modes by which longitudinal data are obtained and examine the possible ramifications of the methods used on the way in which the data can be analyzed.

The prevalent method of collection of data for a longitudinal survey is *prospective* measurement, in which a sample of respondents is followed forwards in time and data are collected on their current situation at a series of points in time. This is the method used in the important set of *birth cohort studies*, such as the U.K. 1970 British Cohort

Study (BCS70). This was based on a sample of all births in Great Britain occurring during one week in 1970, followed up in four subsequent waves over a period of 20 years, to study the medical, physical, educational and social developments of the cohort and to investigate the forces and patterns that shape their lives – Butler et al. (1997). This pioneering project has now been replaced, since 2000, by the similarly designed Millennium Cohort Study (MCS) – Smith and Joshi (2002). The major disadvantage of the prospective method of collection is the long lead time required until a sufficient body of longitudinal data are available for analysis. Other practical problems involve the difficulties of follow-up for dynamic populations and the inherent cumulative attrition, with its implications for nonresponse bias – see Nathan (1999).

The *retrospective method of measurement* collects information at the current time on past events, based on recollection or records. This is the widely used method for case-control studies, sometimes termed *retrospective sampling*, in which subjects are recruited according to their disease status and their past exposure to risk factors is examined – see Scott and Wild (Chapter 38). However, *retrospective measurement* has also been increasingly used in longitudinal sample surveys, often in conjunction with prospective measurement. Thus, in the sixth wave of the BHPS of 1996, all respondents were asked about their family structure during childhood (i.e., whether they lived with one or both parents or other family members during childhood) – Francesconi (2005). Although the retrospective method of collection overcomes the problem of attrition, it is associated with possible acute effects of response error, when it relies on memory. Several studies have indicated the serious problems associated with memory effects in retrospective surveys. For instance, Kazemian and Farrington (2005) compared the validity of retrospective reports with that of prospective reports and official records on the age of onset for criminal offences and found that retrospective reporting is unsuitable for a wide range of research questions. Similarly, Smith and Thomas (2003) found, by test-retest reliability methods, that the quality of long-term recall reports on migration histories may be poor, though they do propose steps that can be taken to improve it.

Although the prospective and retrospective methods differ with respect to their nonsampling errors, standard methods for their longitudinal analysis are basically the same. Thus, a classic result of Prentice and Pyke (1979) shows that prospective and retrospective logistic models applied to case-control data give equivalent results. Similar results for Bayesian analysis are obtained by Seaman and Richardson (2004). However, when observations are clustered, such as in case-control studies with covariate variables obtained from family members, Neuhaus et al. (2002) showed that in some cases the prospective and retrospective analyses differ.

An important category of longitudinal studies is that of *observational* studies, that is, experimental studies in which no random assignment to treatments is possible. These could be retrospective observations on covariate data, obtained from historical administrative or medical records, to supplement data obtained from a prospective longitudinal sample survey. For instance, in the MCS, mentioned previously, interview reports on children were linked to birth register and hospital maternity records (after requesting informed consent – obtained for 92% of respondents), to investigate relationships between current status and birth data – see Tate et al. (2006). In other cases, the observational study is carried out prospectively on a sample of patients, for which case-control studies are difficult or impossible to implement. In particular, this has been the favored

research mode for a wide range of longitudinal studies of the effects of therapeutic interventions and environmental factors on the progression of human immunodeficiency virus infection, in which the natural histories of cohorts of those infected are observed over a length of time – see, for instance, Ko et al. (2003). Since treatments in these studies are not randomly assigned and to overcome the problem of the confounding of treatment-response relationships by time-varying variables, specialized methods of analysis, such as those based on marginal structural models, intensity scores, and inverse probability weighting are required – see, for instance, Gill and Robins (2001), Brumback et al. (2003), and Hogan and Lee (2004).

Another form of longitudinal study in which the effects of different treatments are studied is that of *intervention* studies. In this type of study, an intervention for a medical or social process is initiated after the start of the longitudinal data collection, and subjects are selected for the new intervention on the basis of their previous measurements. Thus, in a study described by Lin and Hughes (1997), historical data on a marker for disease progression define whether subjects are chosen to receive a new treatment. A similar longitudinal intervention study in the economic area is described by Heckman and Robb (1985). They considered the analysis of the effect of training on earnings when enrollment into training is the outcome of a nonrandom selection process.

3. General models for analysis of longitudinal data

3.1. Repeated measures models and generalized estimating equations

The predominant method of analysis for longitudinal data has long been based on the application of generalized linear models (GLMs) – McCullagh and Nelder (1999) – to repeated measures and the use of generalized estimating equations (GEEs) to estimate the model parameters – see, for instance, Diggle et al. (1994). The GLM describes the conditional distribution of the outcome, given its past, where the distribution parameters may vary across time and across units as a stochastic process, according to a mixing distribution. Two different approaches to longitudinal analysis are dealt with by means of similar GLMs. In the "subject-specific" approach, sometimes referred to as the *random effects model*, the heterogeneity between subjects is explicitly modeled, whereas in the "population-average" approach, sometimes referred to as the *marginal model*, the average response is modeled as a function of the covariates, without explicitly accounting for subject heterogeneity. To set ideas, consider the following set-up:

Let y_{it} be the value of the outcome random variable and $\mathbf{x}_{it}$ be a $p \times 1$ vector of fixed covariates for unit i at time t. The times, t, are not necessarily equally spaced, but it is assumed that all units are observed for all time periods. Let $\mathbf{z}_{it}$ be a fixed $q \times 1$ vector of covariates associated with the random effect vector, $\mathbf{b}_i$. Let $u_{it} = \mathrm{E}(\mathrm{y_{it}}|\mathbf{b}_i)$ and $\mu_{it} = E(y_{\mathrm{it}})$ be the conditional and unconditional expectations of the outcome variable, respectively. Then, the mixed GLM, under the *subject-specific approach*, is defined by

$$h(u_{it}) = \mathbf{x}'_{it}\boldsymbol{\beta} + \mathbf{z}'_{it}\mathbf{b}_i; \quad \mathrm{var}(y_{it}|\mathbf{b}_i) = g(u_{it}) \cdot \phi, \tag{1}$$

where $\mathbf{b}_i$ is independently distributed with the distribution, F, and the functions h and g are the link and variance functions, respectively. Under the *population-average*

approach, the marginal expectation is modeled, without the random effect, as

$$h^*(\mu_{it}) = \mathbf{x}'_{it}\boldsymbol{\beta}^*; \quad \text{var}(y_{it}) = g^*(\mu_{it}) \cdot \phi^*. \tag{2}$$

Estimation of the model parameters is obtained, under both models, by solving appropriate generalized estimating equations – see details in Zeger et al. (1988), which includes an example of the analysis of longitudinal data from the Harvard Study of Air Pollution. Further extensions of these models to Markov transition models and examples of their application, primarily in the health sciences, may be found in Diggle (1994).

3.2. Multilevel models

Frequently, longitudinal sample surveys deal with hierarchical populations, such as individuals within households or employees within establishments, for which multilevel modeling is appropriate. On the other hand, Goldstein et al. (1994) considered the analysis of repeated measurements using a two-level hierarchical model, with individuals as second levels and the repeated measurements as the first levels. Thus, denoting y_{it} as above, they proposed the following two-level model:

$$y_{it} = \sum_{k=1}^{p} x_{itk}\beta_k + \sum_{\ell=1}^{p_2} z_{2\ell it}e_{2\ell i} + \sum_{m=1}^{p_1} z_{1mit}e_{1mit}, \tag{3}$$

where the first term denotes fixed effects and the last two terms denote random effects at the higher level (individuals) and at the lower level (measurements), respectively. This is similar to the subject-specific model (1), with the addition of the measurement (time) random effect. Assuming multivariate normality and standard assumptions on covariances, they obtained maximum likelihood estimates of the parameters by the use of an Iterative Generalized Least Squares algorithm. The results are extended to discrete first-order and second-order autoregressive time series models and to continuous time models. A small data set of nine height measurements for each of a sample of boys over 5 years, with age (and its exponents) as the fixed effect, provides an example of the analysis.

Skinner and Holmes (2003) considered a random effects model, in which permanent individual random effects, u_i, are the higher level effects and transitory random effects, v_{it}, are the lower level effects, which may be correlated over time. As an example they set up a basic hierarchical model for the log earnings, y_{it}, of individual i at wave t, for data from the BHPS, as:

$$y_{it} = \beta_t + u_i + v_{it}, \quad t = 1, \ldots, T, \tag{4}$$

where the transitory random effects, v_{it}, follow a first-order autoregressive model AR (1):

$$v_{it} = \rho v_{it-1} + \varepsilon_{it}, \quad t = 1, \ldots, T. \tag{5}$$

The random variables u_i and ε_{it} are assumed to be mutually independent with $E(u_i) = E(\varepsilon_{it}) = 0$ and $\text{var}(u_i) = \sigma^2$; $\text{var}(\varepsilon_{it}) = \sigma^2_\varepsilon$.

Similar models are used by Rao and Yu (1994) for small area estimation, where the unit, i, is the small area.

Fitting of the models and estimation of the parameters can be carried out by two alternative methods. The first is a covariance structure approach, in which the observations on the T waves are treated as a multivariate outcome with individuals as "single level" units. The second approach treats the data as hierarchical, with the lower level units as the waves, $t = 1, \ldots, T$, and the higher level units as the individuals, i.

Feder et al. (2000) considered a model that encompasses both the hierarchical nature of many human populations and the time series relationships between repeated measurements and random effects of higher-level groups. In the following, higher-level groups will be called "households" and lower level units "individuals." The proposed model combines standard multilevel mixed linear models (Goldstein, 1986, 1995), operating at given points in time with a state-space model that represents the time series relationships of the random group effects and the individual measurements. Basic notation and assumptions are as follows:

Let y_{hjt} define the value of the response variable at time $t = 1, \ldots, T$, for individual $j = 1, \ldots, n_h$, belonging to household $h = 1, \ldots N$. The measurements y_{hjt} are assumed to follow the hierarchical two-level linear model:

$$y_{hjt} = \mathbf{x}'_{hjt}\mathbf{b}_t + \mathbf{z}'_{ht}\mathbf{v}_t + \mathbf{z}'_{ht}\mathbf{u}_{ht} + e_{hjt}, \tag{6}$$

where $\mathbf{x}_{hjt}$ is a p-dimensional vector of individual level explanatory variables values; $\mathbf{z}_{ht}$ is a q-dimensional vector of household level explanatory variables; $\mathbf{b}_t$ and $\mathbf{v}_t$ are fixed vector coefficients of appropriate orders; $\mathbf{u}_{ht}$ is a $(q \times 1)$ vector of household level random effects and e_{hjt} is an individual level random residual. The individual and household level random errors are assumed to follow independent first-order autoregressive models,

$$\mathbf{u}_{ht} = \mathbf{A}\mathbf{u}_{ht-1} + \mathbf{d}_{ht}; \quad \mathbf{d}_{ht} \sim \mathrm{N}(\mathbf{0}_q, \mathbf{D}) \tag{7}$$

$$e_{hjt} = \rho e_{hjt-1} + \varepsilon_{hjt}; \quad \varepsilon_{hjt} \sim \mathrm{N}\big(0, \sigma_\varepsilon^2\big). \tag{8}$$

For convenience, A and D are assumed to be diagonal, implying independence of the random group level effects. We also assume $|\mathbf{A}_{ii}| < 1$ and $|\rho| < 1$ to ensure stationarity. It follows from (7) and (8) that for a given time, t, the marginal distributions are as follows:

$$\mathbf{u}_{ht} \sim \mathrm{N}(\mathbf{0}_q, \mathbf{D}^*) \tag{9}$$

$$e_{hjt} \sim \mathrm{N}\big(0, \sigma_e^2\big); \quad \sigma_e^2 = (1 - \rho^2)^{-1}\sigma_\varepsilon^2 \tag{10}$$

where $\mathbf{D}^* = (1 - \mathbf{A}^2)^{-1}\mathbf{D}$.
Thus, the models operating at given time points are standard multilevel models with the above variances for the random first- and second-level effects.

Although the likelihood of this model is easily constructed by using the time series properties of the combined model, the large number of parameters to be estimated results in unstable estimates, if direct maximization of the likelihood is used. Rather a two-stage estimation procedure is proposed. At the first stage, a separate two-level model is fitted for each time point, yielding estimates for the fixed effects and for the variances. At the second stage, the time series likelihood is maximized to yield estimates of the time series model parameters.

The methods are illustrated by a simulation study and an empirical application to data from the Israeli Labor Force Survey. In this application, the outcome variables are weekly hours worked, whereas years of education and gender serve as individual

level explanatory variables and the number of employed persons in the household as a household level explanatory variable.

3.3. Other methods of analysis

Path analysis has long been a preferred method of modeling complex relationships between large numbers of variables in cross-sectional analysis of structured data sets in the social sciences. Its generalization to modeling longitudinal data has been primarily by means of *Graphical Chain Modeling* (GCM) and *Structural Equation Modeling* (SEM). Both approaches provide pictorial representations of the association between variables which are ordered, usually temporally, with the aim of identifying the direct and indirect effects of one variable on another. Although the GCM approach builds up a model for the complete system by fitting a sequence of sub-models, the SEM approach specifies a single model for the complete system of variables being studied.

The GCM approach is based on the construction of a causal diagram which represents the investigator's understanding of the major causal influences among the measurable quantities involved. A basic conditional independence graph is constructed to characterize the conditional independence structure of the data. Each vertex of the graph represents a variable and two vertices are connected if there is a direct association between the variables, whereas unconnected vertices represent variables that are conditionally independent, given all the other variables. The graphs may be used to formulate research hypotheses about indirect relations in an association structure, under the assumption that the set of direct relations is sufficient to understand all associations in the system and that it cannot be further reduced without destroying such association. For further details on how graphical chain models help identify analogies and equivalences between different models and to provide a unifying concept for many statistical techniques used in the analysis of longitudinal data, see Wermuth and Lauritzen (1990). For an interesting example of the application of GCM to the study of the determinants of neonatal and postneonatal mortality in Malaysia, see Mohamed et al. (1998). The method allows both the examination of the effects of direct association of each determinant on mortality and the pathways by which intermediate socio-economic determinants affect mortality.

The SEM approach extends standard regression models to include multiple outcomes, sometimes called endogenous variables, and unobservable latent variables. The basic structural model is a set of regression equations relating each endogenous variable with other endogenous variables and with exogenous variables or covariates. A second component of the SEM is a measurement model, which relates observed study variables to unobservable underlying constructs, represented by one or more latent variables. For a thorough review of the SEM approach, its relationship to latent variable models for multivariate outcomes and to measurement theory, as well as applications to environmental epidemiology, see Sánchez et al. (2005). An interesting application of structural equation modeling to longitudinal data from the U.K. National Child Development Study, which studies simultaneously six different pathways hypothesized to link education and health to other variables, is given by Chandola et al. (2006). They find by applying SEM methods that the association can be explained by a combination of mechanisms, such as adolescent and adult health behaviors and adult and parental social class.

Among a variety of other models used for the analysis of longitudinal data, the role of *Antedependence Models* in dealing with nonstationarity deserves special attention.

The idea of antedependence, first formulated by Gabriel (1962), relates to a set of ordered variables, such as longitudinal observations, which are defined as being s-th order antedependent if each variable, given at least s immediate antecedent variables, is independent of all other preceding variables. Núñez-Antón and Zimmermann (2000) considered unstructured and structured antedependence models for longitudinal data. The unstructured normal model is defined by

$$y_1 = \mathbf{x}'_1\boldsymbol{\beta} + \varepsilon_1; \;\; y_t = \mathbf{x}'_t\boldsymbol{\beta} + \sum_{k=1}^{s^*} \varphi_{tk}\left(y_{t-k} - \mathbf{x}'_{t-k}\boldsymbol{\beta}\right) + \varepsilon_t; \quad (t = 2, \ldots, T), \quad (11)$$

where $s^* = \min(s, t-1)$, ε_t are independent normal random variables with mean zero and possibly time-dependent variances, $\sigma_t^2 > 0$ and $\{\varphi_{tk}\}$ are unrestricted parameters. The model is unstructured in the sense that the $(s+1)(2T-s)/2$ parameters $\{\varphi_{tk}\}$ and $\{\sigma_t^2\}$ cannot be expressed as functions of a smaller set of parameters.

Structured antedependence models follow the same basic model above, but relationships are assumed between the parameters, resulting in more parsimonious models. An example is a model in which correlations over the same time lags are equal and are just monotonic decreasing functions of the time lag. Núñez-Antón and Zimmermann (2000) used several empirical data sets to compare structured and unstructured antedependence models with unstructured covariance models, Autoregressive Integrated Moving Average (ARIMA) models, and random coefficient models.

Another widely used method for the analysis of data from longitudinal surveys and from epidemiological studies is that of *event history analysis*, which models the movement of individuals between states. Basically, this extends survival analysis to several types of failure or competing risks with transitions between them. A flexible framework for statistically modeling such problems is given by *multivariate counting processes* – see, for example, Keiding (1999). Further methods for the modeling of change and of event occurrence are surveyed in Singer and Willet (2003).

4. Treatment of nonresponse

The problems posed by nonresponse in longitudinal surveys have much in common with those occurring in cross-sectional surveys. However, there are some special aspects of nonresponse in longitudinal data, which must be considered. On the one hand, the fact that the same individuals or households are repeatedly requested to provide information on repeated occasions obviously leads to attrition and wave nonresponse, due to fatigue and to difficulties in tracing sample units which are often highly mobile, see, for example, Nathan (1999). On the other hand, the existence of observations for some points in time for the same unit suggests that this information can assist in dealing efficiently with the effects of nonresponse, by considering plausible relationships over time between individual measurements. In the following, we focus on the treatment of missing data resulting from wave nonresponse, where data are available for some points in time and missing for others, rather than complete nonresponse, which can be dealt with similarly to the ways used for dealing with nonresponse in cross-sectional surveys. Different patterns of wave nonresponse to be considered are attrition (no observations from some time point onwards), missing for a single time or for a continuous period and intermittent

dropout. The relationships between the missing data mechanism and the missing and observed data need to be specified. An important distinction is between the mechanisms of missing completely at random (MCAR), missing at random (MAR), and not missing at random (NMAR) or informative missingness – Little (1995), Little and Rubin (2002).

The design-based treatment of wave nonresponse in panel surveys has been addressed in papers by Kalton (1986) and Lepkowski (1989) and in this volume by Kalton (Chapter 5). The methods proposed use imputation and weighting based on regression models, incorporating known auxiliary data, including response to other waves, and taking into account cross-sectional and longitudinal interrelationships.

Model-based treatment of nonresponse in longitudinal data has been considered primarily in the context of experimental science applications. Thus, Diggle and Kenward (1994) proposed a modeling framework for longitudinal data with informative dropouts, which explicitly considers MCAR and MAR dropout as sub-models. Under a general multivariate normal model, they specify a logistic regression model for the dropout process, which allows dependence of the dropout probability on missing observations. They show how to construct likelihood for the unknown model parameters from parametric specifications of the measurement process and of the dropout process. The model parameters are then estimated by maximum likelihood and examples are given for applications to data from milk protein trials, for milk yields, and from multicentre clinical trials in the study of depression. Rotnitzky et al. (1998) considered the use of semiparametric regression for the treatment of informative nonresponse. They proposed a class of augmented inverse probability of response-weighted estimators of the model parameters, which are consistent and asymptotically normal under parametric modeling of the response probabilities. Their estimation procedure can be viewed as an extension of the GEE method that allows for informative nonresponse.

In the sample survey context, Skinner and Holmes (2003) considered the effects of nonresponse in a longitudinal survey, under the models described in Section 3, Eqs (4) and (5), by modifying the standard estimator of the finite population covariance matrix, so that it is based on all "attrition samples," s_t, those responding until and including time t. However, this does not deal with the problem of informative nonresponse. Miller et al. (2001) developed a method for analyzing the categorical outcomes obtained from longitudinal survey samples, with outcomes subject to multiple-cause nonresponse, within the framework of weighted GEEs. They assumed a model that combines different multivariate link functions to permit fitting Markov models to an outcome with categories represented by a mixture of ordinal success states and multiple failure states. They extended the missing data approach of Rotnitzky et al. (1998) to the use of multiple-logit models, to model the probability of multiple reasons for missing success or failure outcomes, assuming that the probability of nonresponse depends only on observed responses and on covariates specified in the missing data. Taylor series and jackknife variance estimators are developed for parameters estimated from these models and are presented within the context of adjusting for survey considerations and multiple-cause nonresponse. The results are applied to disability data obtained from the U.S. Longitudinal Study of Aging (LSOA). A similar approach is proposed by Gong et al. (2003).

Pfeffermann and Nathan (2001) used the time series structures with hierarchical modeling, described in Section 3, Eqs (6)–(10), to deal with informative nonresponse in longitudinal surveys. They considered two methods based on these models, an augmented regression method and one based on a state-space model. The augmented regression

prediction extends standard regression prediction by adding a correction term that takes into account the existing correlations between the observed and the missing data, so that imputation of missing data is based on all observations for all the time periods, as follows:

Let $\tilde{\mathbf{Y}}_{hj} = (y_{hj1}, \ldots, y_{hjT})'$ represent the generic vector of complete values (observed and missing) for individual j in household h, with variance-covariance (V-C) matrix, $\mathbf{S}_h$ (which is a known function of the unknown parameters contained in $\mathbf{A}, \mathbf{D}$, ρ, σ_e^2, but does not depend on j). Let $\mathbf{Q}_{hj}$ define the response indicator matrix of size $t_{hj} \times T$, corresponding to unit hj, (t_{hj} is the number of times that unit hj is observed), such that the observed values are $\mathbf{Y}_{hj} = \mathbf{Q}_{hj}\tilde{\mathbf{Y}}_{hj}$. Similarly, denote by $\overline{\mathbf{Q}}_{hj}$ the indicator matrix for the missing values, of size $\bar{t}_{hj} \times T$, ($\bar{t}_{hj} = T - t_{hj}$), such that the missing values are $\mathbf{Y}_{hj}^{(m)} = \overline{\mathbf{Q}}_{hj}\tilde{\mathbf{Y}}_{hj}$.

The imputed values, based only on data for the same individual are the augmented Best Linear Unbiased (BLU) regression predictions (Pfeffermann, 1988):

$$\hat{\mathbf{Y}}_{hj}^{(m)} = \overline{\mathbf{Q}}_{hj}\tilde{\mathbf{Y}}_{hj}^{(p)} + \overline{\mathbf{Q}}_{hj}\mathbf{S}_h\mathbf{Q}_{hj}'\left(\mathbf{Q}_{hj}\mathbf{S}_h\mathbf{Q}_{hj}'\right)^{-1}\left(\mathbf{Y}_{hj} - \mathbf{Q}_{hj}\tilde{\mathbf{Y}}_{hj}^{(p)}\right), \tag{12}$$

where $\tilde{\mathbf{Y}}_{hj}^{(p)} = \left(y_{hj1}^{(p)}, \ldots, y_{hjT}^{(p)}\right)$ is the complete vector of simple regression predictions, defined by $y_{hjt}^{(P)} = \mathbf{x}_{hjt}'\mathbf{b}_t + \mathbf{z}_{ht}'\mathbf{v}_t$, $\overline{\mathbf{Q}}_{hj}\mathbf{S}_h\mathbf{Q}_{hj}' = \mathrm{Cov}\left(\mathbf{Y}_{hj}^{(m)}, \mathbf{Y}_{hj}\right)$, and $\mathbf{Q}_{hj}\mathbf{S}_h\mathbf{Q}_{hj}' = V(\mathbf{Y}_{hj})$. Similar augmented regression predictions can be based on all the observed data for all the individuals in the household.

The state-space method is based on the formulation of the model. (6)–(10) in a state-space form, as follows:

The observation equation is defined as

$$[\mathbf{Q}_{ht}\mathbf{Y}_{ht}] = [\mathbf{Q}_{ht}\tilde{\mathbf{X}}_{ht}]\underline{\tilde{\beta}}_t + [\mathbf{Q}_{ht}\tilde{\mathbf{Z}}_{ht}]\underline{\alpha}_{ht} \tag{13}$$

and the transition equation is defined as

$$\underline{\alpha}_{ht} = \mathbf{T}_h\underline{\alpha}_{h,t-1} + \underline{\upsilon}_{ht}, \tag{14}$$

where $[\mathbf{Q}_{ht}\mathbf{Y}_{ht}]$ denotes the observed values for household h at time t ($\mathbf{Y}_{ht}$ defines the generic vector of complete values for all the individuals in household h, of order n_h and $\mathbf{Q}_{ht}$ is the corresponding response indicator matrix), $[\mathbf{Q}_{ht}\tilde{\mathbf{X}}_{ht}]$ and $[\mathbf{Q}_{ht}\tilde{\mathbf{Z}}_{ht}]$, with $\tilde{\mathbf{X}}_{ht} = \begin{pmatrix} \mathbf{x}_{h1t}' & \mathbf{z}_{ht}' \\ \vdots & \vdots \\ \mathbf{x}_{hn_ht}' & \mathbf{z}_{ht}' \end{pmatrix}$ and $\tilde{\mathbf{Z}}_{ht} = \begin{pmatrix} \mathbf{z}_{ht}' & \\ \vdots & \mathbf{I}_{n_h} \\ \mathbf{z}_{ht}' & \end{pmatrix}$, are the design matrices of the explanatory variables; $\underline{\tilde{\beta}}_t = \left(\mathbf{b}_t', \mathbf{v}_t'\right)'$ is the $(p+q)\times 1$ vector of fixed parameters; $\underline{\alpha}_{ht} = \left(\mathbf{u}_{ht}', \mathbf{e}_{ht}'\right)'$ is the $(q+n_h)\times 1$ state vector with $\mathbf{e}_{ht} = \left(e_{h1t}, \ldots, e_{hn_ht}\right)'$; $\mathbf{T}_h = \mathbf{A}\oplus\rho\mathbf{I}_{n_h}$ is the transition matrix (a block-diagonal matrix with $\mathbf{A}$ and $\rho\mathbf{I}_{n_h}$ as the two blocks), and $\underline{\upsilon}_{ht} = \left(d_{ht}', \underline{\varepsilon}_{ht}'\right)'$ is a vector of random errors with V-C matrix: $V(\underline{\upsilon}_{ht}) = \mathbf{R}_h = \mathbf{D} \oplus \sigma_\varepsilon^2\mathbf{I}_{n_h}$.

Under the model (with known parameters), the random components can be predicted, either by application of the Kalman filter, if only current and past observations are available, or by an appropriate smoothing filter, if data for subsequent time periods are known. Estimation of the unknown model parameters is obtained by iterative generalized least squares for the augmented regression prediction and by the method of scoring for

the state-space method. A simulation study and an empirical example, based on Israeli Labor Force data, compared the performances of the proposed methods favorably with those of conventional imputation methods, which do not consider NMAR nonresponse, such as mean imputation (within homogenous groups), nearest neighbor imputation, and simple regression imputation.

5. Effects of informative sample design on longitudinal analysis

Standard analysis of longitudinal survey data often fails to account for the complex nature of the sampling design, such as the use of unequal selection probabilities, clustering, poststratification, and other kinds of weighting used for the treatment of nonresponse. Thus, it does not incorporate all the design variables in the analysis model. This may be due to a variety of reasons: there might be too many of them; they might not be of substantive interest; or their values might be unknown, either because the analyst does not have access to them or because they are latent variables, such as in the case of nonresponse. However, if the sampling design is informative, in the sense that the outcome variable is correlated with design variables not included in the model, even after conditioning on the model covariates, standard estimates of the model parameters can be severely biased, leading possibly to false inference.

Sample survey data may be viewed as the outcome of three processes: the process that generates the values of units in the finite population, often referred to as the superpopulation model; the process of selecting the sample units from the finite population, known as the sample selection mechanism; and the process of response. Analytic inference from repeated survey data refers to inference about the superpopulation model parameters, rather than to inference about finite population parameters. When the sample selection probabilities depend on the values of the model response variable, even after conditioning on auxiliary variables, the sampling mechanism becomes informative and the selection effects need to be accounted for in the inference process. This holds, similarly, for the effects of the response mechanism.

Pfeffermann et al. (1998a) proposed a general method of inference on the population (model) distribution under informative sampling that consists of approximating the parametric distribution of the sample measurements for given population distributions and first-order sample selection probabilities. The (marginal) sample distribution is defined as the conditional distribution, given that unit i is in the sample, that is, as the conditional distribution of the observed data. By application of Bayes' theorem, they obtained the following relationship between the sample distribution, $f_s(y_i|\boldsymbol{\theta})$, and the population distribution, $f_p(y_i|\boldsymbol{\theta})$, as follows:

$$f_s(y_i|\boldsymbol{\theta}) = f_p(y_i|\boldsymbol{\theta}, i \in s) = \frac{E_p(\pi_i|\boldsymbol{\theta}, y_i)}{E_p(\pi_i|\boldsymbol{\theta})} f_p(y_i|\boldsymbol{\theta}), \tag{15}$$

where $\pi_i = \Pr(i \in s)$ is the inclusion probability, which may depend on y_i. Under informative sampling, that is, when $E_p(\pi_i|\boldsymbol{\theta}, y_i) \neq E_p(\pi_i|\boldsymbol{\theta})$, this distribution is different from the corresponding population distribution. For further details see also, Pfeffermann and Sverchkov (Chapter 39).

Eideh and Nathan (2006) extended these results to study the case of longitudinal panel sample observations under informative sampling, by fitting time series models

and, in particular, an autoregressive model of order one, for longitudinal survey data, when the sampling design is informative, as follows:

Under previously defined notation, assume that the observed measurements, y_{it}, follow the first-order AR model; that is

$$y_{it} - \mu = \phi(y_{it-1} - \mu), \quad i = 1, \ldots, N; \quad t = 1, \ldots, T, \tag{16}$$

where the errors $\{\varepsilon_{it}\}$ are normally distributed with zero mean and variance σ^2, and $|\phi| < 1$, and that the errors $\{\boldsymbol{\varepsilon}_{it}\}$ pertaining to the same unit and between units are independent. The sample is assumed to be a panel sample selected at time $t = 1$ and all units remain in the sample till time $t = T$. Then, if the sample design is informative, it is intuitively reasonable to assume that the first-order inclusion probabilities $\pi_i = \Pr(i \in s)$ depend on the population values of the response variable at the first occasion only, y_{i1}. Pfeffermann et al. (1998a) considered two alternative approximation models for the population conditional expectations of the inclusion probabilities:

(a) *Exponential inclusion probability model:* $E_p(\pi_i|y_{i1}) = \exp(a_0 + a_1 y_{i1})$.
Under this model, the sample log likelihood function is given by

$$\begin{aligned} l_e\big(\mu, \varphi, \sigma^2, a_1\big) = {} & -\frac{nT}{2}\log(\sigma^2) + \frac{n}{2}\log(1-\varphi^2) \\ & -\frac{1-\varphi^2}{\sigma^2}\sum_{i\in s}\left(y_{i1} - \mu - \frac{a_1\sigma^2}{1-\varphi^2}\right)^2 \\ & +\sum_{i\in s}\left[-\frac{1}{2\sigma^2}\sum_{i=2}^{T}\{y_{it} - \mu - \varphi(y_{it-1} - \mu)\}^2\right]. \end{aligned} \tag{17}$$

This implies that the sample pdf belongs to the same family as the population pdf but differs only in the mean of y_{i1}, which changes from μ to $\mu + a_1\sigma^2/(1-\varphi^2)$.

(b) *Linear inclusion probability model:* $E_p(\pi_i|y_{i1}) = b_0 + b_1 y_{i1}$.
Under this approximation, assuming that $y_{i1} \geq 0$, the sample log likelihood function is given by

$$\begin{aligned} l_e\big(\mu, \varphi, \sigma^2, b_0, b_1\big) = {} & -n\log(b_0 + b_1\mu) - \frac{nT}{2}\log(\sigma^2) \\ & +\frac{n}{2}\log(1-\varphi^2) - \frac{1-\varphi^2}{\sigma^2}\sum_{i\in s}(y_{i1} - \mu)^2 \\ & +\sum_{i\in s}\left[-\frac{1}{2\sigma^2}\sum_{i=2}^{T}\{y_{it} - \mu - \varphi(y_{it-1} - \mu)\}^2\right]. \end{aligned} \tag{18}$$

The methods of estimation proposed for cross-sectional data, such as the two-step method (Pfeffermann et al., 1998; Pfeffermann and Sverchkov, 1999) and the pseudolikelihood method (Binder, 1996, Skinner, 1989a), can be extended to longitudinal survey data. Eideh and Nathan (2006) proposed a two-step method of estimation and two versions of the pseudolikelihood method for the estimation of the unknown parameters of the above population models. A simulation study shows that the estimators based on the sample distribution differ from those obtained under the assumption that the sample

design is not informative, but that their performances are relatively robust to the choice of model and method of estimation.

Finally, the above results can be extended to incorporate both the effects of informative sample design and those of informative nonresponse. This can be done by assuming a model for the response propensity, such as the logistic model for informative dropout proposed by Diggle and Kenward (1994). By considering both the dropout process and the informative sample design, the joint sample distribution for the incomplete sequence of observations can be obtained and its parameters can be estimated by the methods proposed in Eideh and Nathan (2008).

Sample Surveys: Inference and Analysis, Vol. 29B
ISSN: 0169-7161
© 2009 Published by Elsevier B.V.
DOI: 10.1016/S0169-7161(09)00235-1

Categorical Data Analysis for Simple and Complex Surveys

Avinash C. Singh

1. Introduction

Categorical data analysis (CDA) is a fascinating area of statistics being full of theoretical challenges. The main difference from modeling with continuous response or outcome variables is that CDA requires nonlinear models (such as generalized linear) for the mean function to satisfy range restrictions, for example, the range [0, 1] for a binary variable. This also implies that the variance of the model error depends on the mean and hence on unknown mean parameters. As a result, the usual least squares theory is no longer appropriate for best linear unbiased estimation. One could always use the large sample theory of maximum likelihood (ml) estimation if a suitable parametric model were available. However, practitioners often prefer making light assumptions in modeling for prediction, such as specifying only the first two moments in a semiparametric framework. In fact, count data often exhibit overdispersion (McCullagh and Nelder, 1989, Chapter 4) due to clustering, as in the case of correlated binomial outcomes, which makes it difficult even to specify second moments, let alone choosing an appropriate parametric model. As an alternative to ml-estimation, the large sample quasi-likelihood (ql) estimation theory of Wedderburn (1974) and the more general theory of optimal estimating functions (EFs) of Godambe (1960), and Godambe and Thompson (1989), also referred to as ql-estimation by Godambe and Heyde (1987), can be used, which turns out to be quite powerful as well as practical (see also McCullagh and Nelder, 1989, Chapter 9).

In traditional CDA with contingency tables where the data are in the form of counts cross-classified by levels of factors of interest, log-linear models for studying structural relationship among factors or logit models for studying the effects of factors on another factor treated as response are used. For example, in the Canadian Community Health Survey (CCHS), a survey referred to throughout this chapter, the data can be grouped by age-gender domains corresponding to subpopulations (or domains) of the total Canadian population of individuals (aged 12 and older). A binary response variable (y) of interest might correspond to one of the healthy life style indicators such as smoking ($y = 1$ for regular, $y = 0$ for nonregular—never or sometimes) and basic covariates (among others) collected for each sampled individual k are age (x_{1k}—a categorical variable or factor with five levels, say, 12–19, 19–24, 25–34, 35–54, 55+) and gender (x_{2k}—also a

categorical variable). In the case of smoking outcome variable, another useful covariate (x_{3k}) might be the indicator variable for an individual with asthma. While the log-linear and logit models are at the domain (or aggregate) level in the sense that all the covariates considered are at the domain level (even the individual level covariates such as age and gender are common for all units in the same domain as it is defined by the age-gender factors), unit level models (e.g., logistic) with at least one unit level covariate such as x_{3k} or age defined at a level finer than the age factor level used in the contingency table cross-classification can provide more effective use of the data as well as more efficient estimates; here some covariates in the unit level model could still be at an aggregate level. The method of ql-estimation can be used for all such problems. The main purpose of this chapter is to review the application of ql-estimation to CDA for survey data.

If the sampling design is simple random sampling (with replacement), then we say we have data from a simple survey and all the standard results on CDA based on ql-estimation go through. Even if the design were without replacement, the standard results remain approximately valid because the sample size is typically much smaller than the population size. However, often the design is not simple but complex due to stratification, multistage selection, clustering, and unequal probability sampling used mainly for sample efficiency, although techniques such as stratification and unequal probability sampling are used for estimation efficiency as well. For such designs, if we can incorporate the design variables as covariates in the model, then the design can be ignored in the modeling process (see Pfeffermann, 1993; Pfeffermann and Sverchkov, 2003, and Chapter 39 of this handbook) and therefore can be treated as simple for the analysis. In practice, however, it is not feasible to include all the important design variables in the model for the simple reason of parsimony, lack of data availability, and because the resulting model may not be of interest to the analyst. On the other hand, ignoring the design in the analysis may introduce serious selection bias. A striking example of selection bias induced by the design comes from case-control studies where the cases are oversampled while the controls are sampled at much lower sampling rates from the control population (generally stratified by one or more of the model covariates). Since the sample may look very different from the population, the analysis could be quite misleading, in general, unless the sampling weights are used to reflect varying selection probabilities (See Scott, 2006 and Chapter 38 of this handbook).

For complex surveys, it turns out that a weighted version of the ql-approach (referred to hereafter as wql) can be applied for CDA and this is what is emphasized in this chapter. The present chapter contains a detailed treatment of the wql-approach and thus complements the earlier excellent reviews by Rao and Thomas (1989, 2003). An important point to note for complex surveys is that for estimating finite population quantities (FPQs) such as subpopulation or domain totals, large sample consistent estimators of FPQs are easily obtained using Horvitz–Thompson (H-T) type estimators, usually after some calibration. However, the second moments of these estimators around FPQs are generally not available due to unknown second-order inclusion probabilities. Therefore, nonoptimal ql-estimation with correct specification of only the first moment and a working assumption about the second moment as in the generalized estimating equations framework of Liang and Zeger (1986), and corresponding quasi (q-) score tests (see, e.g., Boos, 1992) form a natural starting point. For complex surveys, we will use the term wql-estimation to refer to the method of survey weighted estimating functions developed by Godambe and Thompson (1986) with desirable theoretical properties; see

also Thompson (1997, Chapter 5) and Chapter 26 of this handbook. In this context, the important initial contributions are due to Fuller (1975) in the case of linear regression models and Binder (1983) in the case of generalized linear models, for developing appropriate FPQ estimation framework, although optimality was not considered. Analogous to Boos (1992) who considered simple surveys, weighted quasi (wq) score tests were developed by Binder and Pathak (1994) for the test inversion method used for interval estimation and by Rao et al. (1998) as a general approach to analysis of survey data.

Before reviewing wql-estimation and wq-score tests, the traditional ml-estimation and ml-score tests for CDA are reviewed in Section 2 to provide the background and to fix the ideas and notation. Here, we consider four broad aspects of CDA, namely, model selection, model diagnostics, inferential testing, and inferential estimation in that order. Although some amount of parameter estimation is needed for inferential testing, inferential estimation is presented after testing to compute point, variance, and interval estimates for model parameters (identified as significant after testing) as well as for other parameters defined as functions of domain means. More emphasis is placed on the quadratic score statistic (Q) and its relation to Pearson's X^2 than on the ml-ratio test statistic (usually denoted by G^2) because of its need later for the ql-approach. We also emphasize on the Neyman or nuisance parameter adjusted score (nscore for short) function in view of its need to deal with nested hypotheses. Since the asymptotic distribution of the nscore statistic is independent of consistently estimated nuisance parameters, it makes it convenient to treat the estimated parameter in the nscore statistic as known. The nscore statistic plays an important role in specifying the Rao–Scott (henceforth, R-S) corrections of the X^2 statistic for survey data. In reviewing ml-based methods for interval estimation, we consider the use of test inversion based on score tests as an alternative to Wald's method and the use of observed information instead of expected information (Efron and Hinkley, 1978) in the distribution of parameter estimates so as to improve finite sample properties. We also consider applications of recently introduced Cholesky residuals (Houseman et al., 2004) for obtaining independent Pearson-type residuals at the cell level and Estrella's (1998) R^2-type measures for assessing model fit in CDA.

Next in Section 3, we review ways of generalizing the ml-based results to results under the ql-framework with only first moment assumptions. Construction of the q-nscore function is reviewed as it is key to obtaining results for the ql-approach when applied to model selection, model diagnostics, inferential testing, and inferential estimation in CDA. In Section 4, we consider complex survey data, the main focus of this chapter and show how wq-score functions (and hence wq-nscore functions) can be obtained as estimates of FPQs defined by population (census) q-score functions corresponding to the model under consideration. Here, the instability in the χ^2 approximation to the asymptotic null distribution of the Q statistic based on the wq-nscore function is discussed. This is due to the instability in the estimated covariance matrix of the wq-nscore statistic in the sense of having high relative variance of each element of the matrix. In such situations, a well-known approach in practice is to use a working but stable covariance matrix which is based on a simplified version (not completely arbitrarily chosen though) of the original complex design, and then correct the asymptotic distribution. This is an intriguing but a rather counter-intuitive idea that has been used successfully in other statistical problems and was used by Rao and Scott (1984) for CDA leading to the

well-known R-S corrections, a review of which is provided in this section. In particular, a proportionality condition relating the working covariance of the vector of wq-score functions to the expected wq-information matrix of parameters is made explicit in order for the wq-score statistic with a working covariance to have the form of a X^2-type statistic. The limiting distribution of this statistic is, however, not χ^2 but only a linear combination of independent χ^2_1-variables. The R-S first- and second-order corrections to the test statistic can be used to obtain approximate χ^2 distributions. Also, whenever the test statistic is based on wq-nscore functions with a working covariance matrix, a simple form for defining the generalized design effects (g-deffs) needed for the R-S corrections is provided in terms of the eigenvalues of the product of the actual covariance and the inverse working covariance matrices of the wq-nscore functions. Moreover, we consider use of g-deffs in smoothing the covariance matrix for standardizing residuals for model diagnostics and for inferential estimation.

In Section 5, we consider unit-level models for CDA. Basically, all the results of aggregate-level modeling presented earlier go through except for residual analysis. In particular, we review the uncertainty in the degrees of freedom associated with the Hosmer–Lemeshow's (1980) chi-square statistic, which uses a X^2-type goodness-of-fit statistic after grouping the individuals based on ranking the predicted means. It turns out that the Hosmer–Lemeshow statistic uses nscore functions but with incorrect (like working) covariance matrix, so its asymptotic null distribution can be obtained as a linear combination of χ^2_1-variables. This implies that the R-S corrected Hosmer–Lemeshow statistic can be approximated by a χ^2 distribution. For survey data, these results carry over by defining suitable FPQs corresponding to groupings based on weighted empirical distribution function of predicted means as in Roberst, Ren, and Rao (2008). Finally, summary and a brief discussion of special topics are presented in Section 6.

2. Likelihood-based methods

For the CCHS example mentioned in the introduction, consider a three-dimensional table of counts $\{z_{aij}\}_{1\le a\le A, 1\le i\le I, 1\le j\le J}$, where z_{aij} is the observed count of individuals in category a of the answer to the smoking question with $A = 2$ categories ($a = 1$ denotes a regular smoker while $a = 2$ a nonregular or nonsmoker), category i of the age variable with $I = 5$ categories, and category j of the gender variable with $J = 2$ categories. Denoting by n the total sample size, we therefore have $\sum_{a=1}^{A}\sum_{i=1}^{I}\sum_{j=1}^{J} z_{aij} = n$. We assume that the sample is simple random which implies that the distribution of the counts $\{z_{aij}\}$ is multinomial. Let y_{aijk} be the random indicator variable that unit k in the sample falls in the category aij and μ_{aij} denote the mean of y_{aijk}, that is, μ_{aij} is the probability that $y_{aijk} = 1$. Thus, $\sum_{k=1}^{n_{ij}} y_{aijk} = n_{ij}\bar{y}_{aij} = z_{aij}$, n_{ij} being the sample size for domain ij. Now, treating the smoking status indicator of an individual k in the age-gender domain defined by ij as the binary response variable and the age and gender variables as explanatory, consider a saturated logit model for the mean μ_{1ij} as follows.

$$\text{Logit Model:}\quad \operatorname{logit}\mu_{1ij} \equiv \log\bigl(\mu_{1ij}/\bigl(1-\mu_{1ij}\bigr)\bigr) = v + v_{1(i)} + v_{2(j)} + v_{12(ij)}, \tag{1}$$

where v is the intercept, $v_{1(i)}$, $v_{2(j)}$ are main or one-factor effects due to age and gender covariates at levels i, j, respectively, and $v_{12(ij)}$ are interactions or two-factor effects. The v-parameters satisfy the constraints $\sum_{i=1}^{I} v_{1(i)} = 0$, $\sum_{j=1}^{J} v_{2(j)} = 0$, $\sum_{i=1}^{I} v_{12(ij)} = 0$, and $\sum_{j=1}^{J} v_{12(ij)} = 0$. Thus, there are I-1 linearly independent v_1-parameters, J-1 independent v_2-parameters, and $(I\text{-}1)(J\text{-}1)$ independent v_{12}-parameters with a total of IJ, the maximum possible including the intercept. In practice, interest lies in finding a parsimonious nonsaturated model such as the one with no interactions, that is, $v_{12(ij)} = 0$ for all (i, j), which implies, in particular, that $\left(\mu_{1i1}(1-\mu_{1i1})^{-1}/\mu_{1i2}(1-\mu_{1i2})^{-1}\right)$, the odds ratio over the two genders is the same for every age category i.

We can also express the model (1) as a nonlinear regression model that may be convenient for understanding. For this purpose, denote the domain ij as d ($d = 1, 2, \ldots, D$, where $D = IJ$), and the lth covariate for the domain d by $A_{x(l),d}$, which is the average of the lth covariate x_l ($l = 1$ to p) over the individuals in domain d with subpopulation size N_d. Here, x_l is 1 or 0 depending on whether the corresponding age-gender covariate category belongs to the domain d or not while for the intercept it is always equal to 1. Thus, for example, the covariate $A_{x(l),d}$ for the parameter $v_{1(i)}$ is zero except when the domain d and the variable x_l do represent the age category i. Now the model (1) can be rewritten for $d = 1$ to D, in the familiar regression form as,

$$\text{logit}\,\mu_{1d} = A'_{x,d}\beta, \tag{2}$$

where β is the vector of v-parameters of dimension p (equals IJ in the saturated case) and $A_{x,d}$ is the corresponding vector of p-covariates.

Next we consider the four main parts of data analysis as mentioned in the introduction. It is assumed that the sample size is large enough for applicability of approximate frequentist methods.

2.1. Model selection under likelihood-based methods

2.1.1. Stepwise covariate selection

This is usually carried out by first splitting the data into two parts of training and validation samples, selecting the model based on the training sample, and confirming the selection based on the validation sample. For this purpose, one can define a baseline (B) model with parameter dimension p_{B} and a full (F) model with parameter dimension p_{F} so that the final reduced model (R) with parameter dimension p_{R} is somewhere in between the two and is obtained by a suitable stepwise selection (backward or forward) procedure. This is typically done under a hierarchy principle of covariate inclusion in the sense that if a higher order interaction is in the model, then all the lower order interactions and main effects are also included in the model. The baseline model basically consists of the overall intercept term and some main effects deemed mandatory based on past experience and subject matter considerations, while the full model consists of all the potential covariates and the important interactions. For our CCHS example, the full model is simply the saturated (S) model (1) with maximum number of parameters being equal to $p_{\mathrm{S}}(=D)$. Now, with categorical covariates, one can define a hierarchy of importance in including covariates corresponding to 1-factor effects, 2-factor effects, and so on. For example, we have one 1-factor effect of gender having two levels, four 1-factor

effects of age having five levels, and four genders by age interactions or 2-factor effects. A useful practical strategy might be a combination of moving forward a block (of factor effects) and then backward within a block as follows. One starts with a baseline model consisting of a block of all 1-factor or main effects. Now within this block, one could perform a backward selection to see which one can be dropped in a stepwise manner based on significance probabilities in the presence of remaining covariates. Alternatively, it may be desirable to retain all of them based on subject-matter considerations. After selecting the first block of covariates, one goes forward to add the second block of 2-factor effects and then goes backward within this block to select significant covariates in a stepwise manner. This process is continued until all the covariates under the full model are tested.

2.1.2. Testing covariate significance

To check significance of any covariate or a group of them in the presence of others, a number of asymptotically equivalent large sample χ^2 tests such as the ml-ratio test G^2, the score test Q_{ml}, Pearson's chi-square test X^2, and Wald's test Q_W (also quadratic but in ml-estimators and not score functions) can be used (see, e.g., Cox and Hinkley, 1974, Chapter 9). Before we define them, we need to compute the ml-estimator $\hat{\beta}^{\text{ml}}_{[p]}$ of $\beta_{[p]}$, where $\beta_{[p]}$ denotes the vector parameter $(\beta_l)_{1\le l\le p}$. In the following, a square-bracketed subscript $[p]$ of a vector β will be used to denote the parameter dimension while the unbracketed subscript l simply denotes the lth element of the vector. The vector $\beta_{[p_1-p_2]}$ will be used to denote the last $[p_1 - p_2]$ elements of $\beta_{[p_1]}$ partitioned as $\beta_{[p_1]} = (\beta'_{[p_2]}, \beta'_{[p_1-p_2]})'$. However, depending on the context, we may not always need this subscript. Now, the likelihood for the counts $\{z_{ad}\}_{1\le a\le 2, 1\le d\le D}$ is multinomial and can be expressed as the product of the marginal likelihood of the domain counts or sample sizes $\{n_d\}_{1\le d\le D}$ and the conditional likelihood of $\{z_{ad}\}_{1\le a\le 2, 1\le d\le D}$ given $\{n_d\}_{1\le d\le D}$, which are sufficient for the nuisance parameters defining the distribution of $\{n_d\}_{1\le d\le D}$. It follows that the conditional likelihood has all the information about the parameters $\beta_{[p]}$of interest and has the form of product binomial. The log-likelihood given $\{n_d\}_{1\le d\le D}$ under model (2) is obtained as

$$\begin{aligned}\log L &= \sum_d \big(z_{1d}\log\mu_{1d} + (n_d - z_{1d})\log(1-\mu_{1d})\big) + \text{const}\\ &= \sum_d \Big[\big(A'_{x,d}\beta\big)\, z_{1d} - n_d \log\Big(1 + e^{A'_{x,d}\beta}\Big)\Big] + \text{const}\end{aligned} \quad . \tag{3}$$

Note that the above likelihood does not change even if the counts were generated by a Poisson sampling scheme because the additional sufficient statistic, namely the total count n, can also be conditioned. There would be no need of conditioning the likelihood if the counts were indeed generated under a product multinomial sampling scheme with domains d as strata and we get the same likelihood. It follows that the p-vector of score functions for the $\beta_{[p]}$ parameters conditional on $\{n_d\}_{1\le d\le D}$ is given by

$$\begin{aligned}\phi_{\beta[p]} \equiv \partial \log L/\partial\beta &= \sum_d A_{x,d}(z_{1d} - m_{1d})\\ &= \sum_d A_{x,d} n_d\big(\overline{y}_{1d} - \mu_{1d}(\beta)\big); \quad m_{1d} = n_d\mu_{1d}(\beta).\end{aligned} \tag{4}$$

An efficient algorithm to estimate $\hat{\beta}^{\mathrm{ml}}_{[p]}$ is the Newton–Raphson iterative procedure. For iteration $(\nu+1)$, we have

$$\begin{aligned}\beta^{(\nu+1)}_{[p]} &= \beta^{(\nu)}_{[p]} + J^{-1}_{\beta[p]} \sum_d A_{x,d}(z_{1d} - n_d\mu_{1d}(\beta))\Big|_{\beta^{(\nu)}_{[p]}} \\ &= J^{-1}_{\beta[p]} \sum_d A_{x,d}\big[z_{1d} - \{n_d\mu_{1d}(\beta) - n_d u_{1d}(\beta)A'_{x,d}\beta\}\big]\Big|_{\beta^{(\nu)}_{[p]}}, \end{aligned} \tag{5}$$

where $J_{\beta[p]}$ is the observed information matrix defined as $\big(-\partial\phi_{\beta[p]}\big/\partial\beta'_{[p]}\big)$, equaling $\sum_d n_d u_{1d}(\beta)A_{x,d}A'_{x,d}$, and $u_{1d}(\beta)$ denotes $\mu_{1d}(\beta)(1-\mu_{1d}(\beta))$. Note that in the above iterative procedure, it is not necessary to choose an initial value $\beta^{(0)}_{[p]}$, which turns out to be particularly convenient in practice. One can directly use the observed proportion $n_d^{-1}z_{1d}$ or $\overline{y}_{1d}$ for $\mu^{(0)}_{1d}$ and hence to obtain $\log\mathrm{it}\ \mu^{(0)}_{1d}$. Alternatively, one can also use instead $(n_d\overline{y}_{1d}+0.5)(n_d+1)^{-1}$ in case $\overline{y}_{1d}$ is 0 or 1 (see McCullagh and Nelder, 1989, p. 117). The relation $\mathrm{logit}\,\mu_{1d} = A'_{x,d}\beta$ makes it possible to bypass $\beta^{(0)}_{[p]}$ for the above iterations by directly finding the initial value of $A'_{x,d}\beta_{[p]}$ for each d without having to define $\beta^{(0)}_{[p]}$ separately. We can now describe various tests of significance for model selection as follows:

G^2 Test: For the logit model (2), suppose we wish to test $H_2 : \beta_{[p_1]} \in \Omega_2$ nested within $H_1 : \beta_{[p_1]} \in \Omega_1$, where Ω_1 is an open subset of R^{p_1}, $\beta_{[p_1]}$ is the vector $(\beta_l)_{1\le l\le p_1}$, and Ω_2 is a subset of Ω_1 such that $\beta_{[p_1]} = \big(\beta'_{[p_2]},\ \beta'_{[p_1-p_2]}=0\big)'$, $1 \le p_2 < p_1$. That is, the model under H_1 is given by

$$\mathrm{logit}\,\mu_{1d}\big(\beta_{[p_1]}\big) = A'_{x[p_1],d}\beta_{[p_1]} = A'_{x[p_2],d}\beta_{[p_2]} + A'_{x[p_1-p_2],d}\beta_{[p_1-p_2]}, \tag{6}$$

and the null hypothesis is simply $H_2 : \beta_{[p_1-p_2]} = 0$. Under suitable regularity conditions, the large sample (in the sense of n_ds being large while D remains bounded) ml-ratio test using (3) is reject H_2 in favor of H_1 if

$$\begin{aligned} G^2(H_2|H_1) &\equiv -2\left\{\log L\left(\hat{\beta}^{\mathrm{ml}}_{[p_2]}\right) - \log L\left(\hat{\beta}^{\mathrm{ml}}_{[p_1]}\right)\right\} \\ &= 2\sum_d\sum_a z_{ad}\log\left\{\hat{m}^{(1)}_{ad}\Big/\hat{m}^{(2)}_{ad}\right\} > \chi^2_{p_1-p_2,\alpha}, \end{aligned} \tag{7}$$

where $\hat{m}^{(1)}_{ad}$, $\hat{m}^{(2)}_{ad}$ are the ml-estimates of the expected count for cell (a,d) under H_1 and H_2 respectively, and $\chi^2_{p_1-p_2,\alpha}$ is the upper α-point of the χ^2 distribution with degrees of freedom equal to the difference in the dimension of β under H_1 and H_2, that is, $p_1 - p_2$ (see, e.g., Fienberg, 1980, Chapter 6). Note that if H_1 is replaced by the saturated model denoted by H_{S}, then $\hat{m}^{(1)}_{ad}$ is simply z_{ad}, and $G^2(H_2|H_{\mathrm{S}})$ simplifies to $2\sum_d\sum_a z_{ad}\log\{z_{ad}\big/\hat{m}^{(2)}_{ad}\}$ taking the well-known form of $2\sum O\log(O/E)$, where O denotes the observed count z_{ad} and E denotes the estimated expected count $\hat{m}^{(2)}_{ad}$. Also, $G^2(H_2|H_1)$ can be expressed as $G^2(H_2|H_{\mathrm{S}}) - G^2(H_1|H_{\mathrm{S}})$. In the case of log-linear models (see below), the expression (7) simplifies further to $2\sum_d\sum_a\hat{m}^{(1)}_{ad}\log\{\hat{m}^{(1)}_{ad}\big/\hat{m}^{(2)}_{ad}\}$ for hierarchical models but not in general (see, e.g., Fienberg, 1980, Chapter 4).

For the CCHS example, if the response variable of smoking status is polytomous with possible values of three categories, say, regular, nonregular, and nonsmoker, then the logit model (1) can be modified to incorporate two logits: one for the odds of regular

smoker against nonsmoker and the other for nonregular smoker against nonsmoker. The regression formulation (2) can be suitably modified and the above description of ml-estimation for the new expanded β-vector and G^2 tests for corresponding hypotheses easily carry over.

If the categorical data are such that there is no distinction between response and explanatory variables as in the case of contingency tables, then one can test for structural relationship via log-linear models (see Bishop et al., 1975, Chapter 3). In particular, for the CCHS example, if the smoking status is not treated as a dependent variable, then a saturated log-linear model for the mean μ_{aij} of the indicator variable y_{aijk} for the kth individual to fall in the cell aij is defined as

$$\text{Log-Linear Model:} \quad \log \mu_{aij} = u + u_{1(a)} + u_{2(i)} + u_{3(j)} + u_{12(ai)} + u_{13(aj)} + u_{23(ij)} + u_{123(aij)}, \tag{8}$$

where the u-parameters satisfy the usual restrictions, that is, the main or one-factor effects $u_{1(a)}$s satisfy $\sum_a u_{1(a)} = 0$, the two-factor effects $u_{12(ai)}$s satisfy $\sum_a u_{12(ai)} = 0$, $\sum_i u_{12(ai)} = 0$, and so on. For the saturated model, the total number of u-parameters is now AIJ, all independent under Poisson sampling but totaling one less (AIJ-1) under multinomial, implying that the intercept u acts like a normalizing constant. Moreover, under product multinomial sampling, there are more restrictions on the u-parameters corresponding to fixed margins. The model (8) can also be expressed as a regression model analogously to (2) by treating the indicator variable y_{aijk} as response, the cell aij as the domain d and defining the corresponding d-level covariates $A_{x,d}$ as coefficients of u-parameters now treated as β-parameters. Under multinomial sampling of size n, the ml-equations for the β-parameters under the log-linear model are given by

$$\sum_{d=1}^{D} A_{x,d}(z_d - m_d) = 0; \quad m_d = n\mu_d, \quad D = AIJ, \tag{9}$$

which can be solved by Newton–Raphson as before; similar to the solution of (4) for logit models. For models having "sufficient" configurations (see Bishop et al., 1975, Chapter 3), a well-known traditional algorithm of iterative proportional fitting can also be used to solve (9). The use of this algorithm produces the estimated expected counts m_d directly, without having to first solve for the β-estimates. In particular, for the hypothesis of no three-factor interactions in model (8), the ml-estimates of $\{m_{aij}\}$ are obtained by solving via iterative proportional fitting the equations (in the traditional notation)

$$z_{ai+} = \hat{m}_{ai+}, \quad z_{a+j} = \hat{m}_{a+j}, \quad \text{and} \quad z_{+ij} = \hat{m}_{+ij}, \tag{10}$$

where z_{ai+}, for example, denotes $\sum_j z_{aij}$. However, the above iterative method is generally not as efficient as Newton–Raphson and also not as applicable. The ml-equations (9) for log-linear models under other sampling schemes such as Poisson sampling do not change by considering the conditional likelihood given the total count n, a sufficient statistic, and also not for product multinomial sampling by forcing appropriate u-parameters in the model corresponding to the margins to be fixed under the sampling

design. The G^2 tests can also be defined for log-linear models in a manner analogous to their definition in the case of logit models. It may be noted that, in practice, logit models are generally preferable in view of their substantive interpretability representing the effect of other categorical variables on a response variable and because of parameter parsimony. Moreover, since more than one log-linear models may give rise to the same logit model (see Fienberg, 1980, Chapter 6) and if the ultimate interest lies in logit models, it may be better to fit them directly rather than obtaining them indirectly through log-linear models. However, log-linear models are useful for general understanding of interactions among various categorical variables.

Going back to the discussion of G^2 tests, we observe that G^2 has the desirable property of being invariant under one-to-one nonlinear parameter transformations as it is not based on the explicit form of ml-estimators. However, it requires parameter estimates under both H_1 and H_2, which may not be desirable in practice.

Score Test Q_{ml}: Following Cox and Hinkley (1974, Chapter 9), the score test statistic for the logit model (2) can be defined as follows. Let $Q(\phi_{\beta[p]})$ denote the quadratic form

$$Q\big(\phi_{\beta[p]}\big) = \phi'_{\beta[p]}\Sigma^{-1}_{\phi[p]}\phi_{\beta[p]}; \quad \Sigma_{\phi[p]} = \text{Cov}\big(\phi_{\beta[p]}\big) \quad (p \times p). \tag{11}$$

For the logit model (2) with $\phi_{\beta[p]}$ defined in (4), $\Sigma_{\phi[p]}$ is given by $A'_x \text{diag}\{n_d u_{1d}(\beta)\}_{1\le d\le D} A_x \left(= \sum_d n_d u_{1d}(\beta_{[p1]}) A_{x,d} A'_{x,d}\right)$, where $A_x = \left(A'_{x,d}\right)_{1\le d\le D}$ is the $D \times p$ matrix of model covariates, and $u_{1d}(\beta)$ is $\mu_{1d}(\beta)(1-\mu_{1d}(\beta))$ as defined earlier in (5). Now the score test rejects H_2 in favor of H_1 if

$$Q_{\text{ml}}(H_2|H_1) = Q(\phi_{\beta[p_1]})\Big|_{\beta_{[p_1]}=\left(\hat{\beta}'^{\text{ml}}_{[p2]},\beta_{[p_1-p_2]}=0\right)'} > \chi^2_{p_1-p_2,\alpha}, \tag{12}$$

where the ml-estimator $\hat{\beta}^{\text{ml}}_{[p_2]}$ needs to be computed only under H_2. The test (12) is also known as the C.R. Rao's score test. If H_1 is the saturated model H_S, then the $D \times D$ matrix $A_{x[p_S]}$ is nonsingular, and Q_{ml} for H_2 is given by

$$\begin{aligned}
Q_{\text{ml}}(H_2|H_S) &= Q(\phi_{\beta[p_S]})\Big|_{\beta_{[p_S]}=\left(\hat{\beta}'^{\text{ml}}_{[p_2]},\beta'_{[p_S-p_2]}=0\right)'} \\
&= \left(A'_{x[p_S]}(z_1-m_1)\right)'\left(A'_{x[p_S]}\text{Diag}\left\{n_d^{-1}m_{1d}(n_d-m_{1d})\right\}A_{x[p_S]}\right)^{-1} \\
&\quad \times \left(A'_{x[p_S]}(z_1-m_1)\right)\Big|_{\beta_{[p_S]}=\left(\hat{\beta}'^{\text{ml}}_{[p_2]},\beta_{[p_S-p_2]}=0\right)'}, \\
&= \sum_{d=1}^{D}\frac{n_d(z_{1d}-m_{1d})^2}{m_{1d}(n_d-m_{1d})}\Big|_{\beta_{[p_S]}=\left(\hat{\beta}'^{\text{ml}}_{[p_2]},\beta_{[p_S-p_2]}=0\right)'} \\
&= \sum_{d=1}^{D}\sum_{a=1}^{2}\frac{(z_{ad}-m_{ad})^2}{m_{ad}}\Big|_{\beta_{[p_S]}=\left(\hat{\beta}'^{\text{ml}}_{[p_2]},\beta_{[p_S-p_2]}=0\right)'}
\end{aligned} \tag{13}$$

which coincides with the usual Pearson's X^2 test statistic of the form $\sum(O-E)^2/E$. It may be of interest to note that with the log-linear model (8) for the table of counts $\{z_{ad}\}$under multinomial sampling, the covariance matrix of the score vector of dimension $p_S = AD$ is singular. In this case, using a g-inverse in defining the quadratic form gives

rise to the usual X^2 as in (13) or alternatively, we can define the score statistic with just $p_S - 1$ score functions by arbitrarily dropping one; (see Cox and Hinkley, 1974, p. 316). An equivalent alternative test (known as $C(\alpha)$, C in honor of Cramér and α for the significance level) using nscore (nuisance parameter adjusted score mentioned earlier in Section 1) function $\phi_{\beta[p_1-p_2|p_2]}$ was developed by Neyman and is given by

$$
\begin{aligned}
Q_{\mathrm{ml}}^{c(\alpha)}(H_2|H_1) &= Q\big(\phi_{\beta[p_1-p_2|p_2]}\big)\Big|_{\beta_{[p_1]}=\left(\hat{\beta}'^{\mathrm{ml}}_{[p_2]},\beta'_{[p_1-p_2]}=0\right)'};\\
\phi_{\beta[p_1-p_2\,|\,p_2]} &= \phi_{\beta[p_1-p_2]} - \Sigma_{\phi[p_1-p_2,p_2]}\Sigma^{-1}_{\phi[p_2]}\phi_{\beta[p_2]},\ \Big(\phi_{\beta[p_1]} = (\phi'_{\beta[p_2]},\phi'_{\beta[p_1-p_2]})'\Big),\\
&\equiv F\phi_{\beta[p_1]},\ \big(F = \big(-\Sigma_{\phi[p_1-p_2,p_2]}\Sigma^{-1}_{\phi[p_2]},\ I_{(p_1-p_2)\times(p_1-p_2)}\big)\big),
\end{aligned}
\tag{14}
$$

where $Q(.)$ and $\Sigma_{\phi[p_1]}$ are defined as in (11). The nscore function $\phi_{\beta[p_1-p_2\,|\,p_2]}$, obtained after adjusting $\phi_{\beta[p_1-p_2]}$, is orthogonal (i.e., uncorrelated) to the score function $\phi_{\beta[p_2]}$ for nuisance parameters because the projection of $\phi_{\beta[p_1-p_2]}$ on the linear space generated by $\phi_{\beta[p_2]}$ (under the covariance norm) is subtracted from it. Using the covariate matrices $A_{x[p_2],d}$, $A_{x[p_1-p_2],d}$ defined in (6), $\Sigma_{\phi[p_1]}$ can be partitioned as

$$
\begin{pmatrix} \Sigma_{\phi[p_2]} & \Sigma_{\phi[p_2,p_1-p_2]} \\ \Sigma_{\phi[p_1-p_2,p_2]} & \Sigma_{\phi[p_1-p_2]} \end{pmatrix}
$$

$$
= \begin{pmatrix} \sum_d n_d u_{1d}(\beta) A_{x[p_2],d} A'_{x[p_2],d} & \sum_d n_d u_{1d}(\beta) A_{x[p_2],d} A'_{x[p_1-p_2],d} \\ \sum_d n_d u_{1d}(\beta) A_{x[p_1-p_2],d} A'_{x[p_2],d} & \sum_d n_d u_{1d}(\beta) A_{x[p_1-p_2],d} A'_{x[p_1-p_2],d} \end{pmatrix},
\tag{15}
$$

where $\Sigma_{\phi[p_2,p_1-p_2]}$, for example, is the covariance between $\phi_{\beta[p_2]}$ and $\phi_{\beta[p_1-p_2]}$.

The covariance $\Sigma_{\phi[p_1-p_2\,|\,p_2]}$ needed for the nscore function $\phi_{\beta[p_1-p_2\,|\,p_2]}$ is then obtained as

$$
\Sigma_{\phi[p_1-p_2\,|\,p_2]} = F\Sigma_{\phi[p_1]}F' = \Sigma_{\phi[p_1-p_2]} - \Sigma_{\phi[p_1-p_2,\,p_2]}\Sigma^{-1}_{\phi\,[p_2]}\Sigma_{\phi[p_2,p_1-p_2]}. \tag{16}
$$

Using a decomposition of $Q(\phi_{\beta[p_1]})$, Neyman's $Q_{\mathrm{ml}}^{c(\alpha)}$of (14) can also be expressed as

$$
Q_{\mathrm{ml}}^{c(\alpha)}(H_2|H_1) = Q\big(\phi_{\beta[p_1]}\big) - Q\left(\phi_{\beta[p_2]}\right)\Big|_{\beta_{[p_1]}=\left(\hat{\beta}'^{\mathrm{ml}}_{[p_2]},\beta'_{[p_1-p_2]}=0\right)'}. \tag{17}
$$

Now with the ml-estimate $\hat{\beta}^{\mathrm{ml}}_{[p_2]}$, the second term with the negative sign in (17) drops out and the form of the Neyman's score statistic coincides with Rao's. However, a practical advantage of the test statistic of (17) is that it is valid for any $\sqrt{n}$-consistent estimate of $\beta_{[p_2]}$ under H_2 and remains applicable when the ml-estimates may be intractable or computationally difficult. Note that with alternative estimators $\hat{\beta}_{[p_2]}$, the second term in (17) with the negative sign does not drop out. In this chapter, the term score statistic will be used to refer to the quadratic form $Q(.)$ regardless of whether the argument in $Q(.)$ is the score function vector $\phi_{\beta[p_1]}$ of dimension p_1 or the nscore function vector $\phi_{\beta[p_1-p_2|p_2]}$ of dimension $p_1 - p_2$, although the former corresponds to Rao's score statistic while the latter to Neyman's statistic.

X^2 *Test*: It follows from (13) that for testing H_2 given $H_1 = H_\mathrm{S}$, the usual Pearson's X^2 test is identical to the score test $Q_\mathrm{ml}(H_2|H_\mathrm{S})$. Now for testing the nested hypothesis H_2 given H_1, the $\mathrm{X}^2(H_2|H_1)$ test is defined by rejecting for large values of

$$\begin{aligned}\mathrm{X}^2(H_2|H_1) &\equiv \mathrm{X}^2(H_2|H_\mathrm{S}) - \mathrm{X}^2(H_1|H_\mathrm{S}) \\ &= \sum_{d=1}^{D}\sum_{a=1}^{2}\frac{\left(z_{ad} - \hat{m}_{ad}^{(2)}\right)^2}{\hat{m}_{ad}^{(2)}} - \sum_{d=1}^{D}\sum_{a=1}^{2}\frac{\left(z_{ad} - \hat{m}_{ad}^{(1)}\right)^2}{\hat{m}_{ad}^{(1)}}\ ,\end{aligned} \tag{18}$$

by referring to the $\chi^2_{p_1-p_2}$ distribution, where $\hat{m}_{ad}^{(2)}$ and $\hat{m}_{ad}^{(1)}$ denote the estimated expected counts under H_2 and H_1, respectively. Note that if $H_1 = H_\mathrm{S}$, then $\hat{m}_{ad}^{(1)} = z_{ad}$ and the second term in $\mathrm{X}^2(H_2|H_1)$ drops out as expected. The form of $\mathrm{X}^2(H_2|H_1)$ defined above is convenient for computation but it requires two ml-estimates, and unfortunately, it is also not necessarily positive unlike the score statistic and the G^2 statistic for nested hypotheses. An alternative and asymptotically equivalent form due to Rao (1973, p. 398), which is positive by construction, is given by

$$\mathrm{X}^2_R(H_2|H_1) = \sum_{d=1}^{D}\sum_{a=1}^{2}\frac{\left(\hat{m}_{ad}^{(1)} - \hat{m}_{ad}^{(2)}\right)^2}{\hat{m}_{ad}^{(2)}}. \tag{19}$$

The two tests $\mathrm{X}^2(H_2|H_1)$ and $Q_\mathrm{ml}^{c(\alpha)}(H_2|H_1)$ of (17) are asymptotically equivalent because

$$\begin{aligned}Q_\mathrm{ml}^{c(\alpha)}(H_2|H_1) \approx\ & Q_\mathrm{ml}(H_2|H_\mathrm{S})\Big|_{\beta_{[p_\mathrm{S}]}=\left(\hat{\beta}'^{\mathrm{ml}}_{[p_2]},\beta'_{[p_\mathrm{S}-p_2]}=0\right)'} \\ &- Q_\mathrm{ml}(H_1|H_\mathrm{S})\Big|_{\beta_{[p_\mathrm{S}]}=\left(\hat{\beta}'^{\mathrm{ml}}_{[p_1]},\beta'_{[p_\mathrm{S}-p_1]}=0\right)'}.\end{aligned} \tag{20}$$

Wald Test $\mathbf{Q_W}$: First observe that the asymptotic normality of $\hat{\beta}^{\mathrm{ml}}_{[p]}$, the solution of the score equation (4), follows from that of $\phi_{\beta[p]}$ because by Taylor expansion,

$$\phi_{\beta[p]} \approx J_{\beta[p]}\left(\hat{\beta}_{[p]} - \beta_{[p]}\right);\ J_{\beta[p]} = -\partial\phi_{\beta[p]}/\partial\beta'_{[p]}, \tag{21}$$

and $J_{\beta[p]}$ is the observed information matrix defined earlier for (5). It follows that the asymptotic distribution of $\hat{\beta}^{\mathrm{ml}}_{[p]}$ is given by

$$\hat{\beta}^{\mathrm{ml}}_{[p]} - \beta_{[p]} \sim_{\mathrm{approx}} N\left(0, \Gamma_{\beta[p]}\right)\Big|_{\beta=\hat{\beta}^{\mathrm{ml}}_{[p]}};\ \Gamma_{\beta[p]} = J^{-1}_{\beta[p]}\Sigma_{\phi[p]}J'^{-1}_{\beta[p]}. \tag{22}$$

The covariance matrix $\Gamma_{\beta[p]}$ in our case reduces to $I^{-1}_{\beta[p]}$, where $I_{\beta[p]} = E(J_{\beta[p]})$, the expected information matrix because $I_{\beta[p]} = J_{\beta[p]}$ for the canonical link function of logit, and the expected information matrix $I_{\beta[p]}$ equals the covariance $\Sigma_{\phi[p]}$ for an optimal EF such as the score function. The Wald statistic $Q_\mathrm{W}(H_2|H_1)$ is defined by a quadratic form in the unrestricted (i.e., under H_1) ml-estimator $\hat{\beta}^{\mathrm{uml}}_{[p_1-p_2]}$ of test parameters $\beta_{[p_1-p_2]}$ and is given by

$$Q_\mathrm{W}(H_2|H_1) = Q\left(\hat{\beta}^{\mathrm{uml}}_{[p_1-p_2]}\right) = \left(\hat{\beta}^{\mathrm{uml}}_{[p_1-p_2]} - 0\right)'\hat{\Gamma}^{-1}_{\beta[p_1-p_2]}\left(\hat{\beta}^{\mathrm{uml}}_{[p_1-p_2]} - 0\right), \tag{23}$$

where the estimated covariance matrix $\hat{\Gamma}_{\beta[p_1-p_2]}$ of the quadratic form is obtained as the lower $(p_1 - p_2) \times (p_1 - p_2)$ principal submatrix of $\Gamma_{\beta[p_1]}$ evaluated at $\beta_{[p_1]} = \hat{\beta}^{\text{ml}}_{[p_1]}$. Observe that $Q_{\text{W}}(H_2|H_1)$ depends on the functional form of the ml-estimator of the test parameter directly.

The Wald test rejects for large values of Q_{W} by referring to the $\chi^2_{p_1-p_2}$ distribution. We remark that although the Wald test is in common use, it does not share the desirable property of invariance to nonlinear parameter transformations with the other tests. Moreover, the finite sample behavior of the χ^2 approximation is invariably not very stable due to a lack of representation in general of $\hat{\beta}^{\text{ml}}_{[p_1]}$ as a linear function of independent terms unlike the case with $\phi_{\beta[p_1]}$ or the log likelihood function, which helps in faster convergence to normality.

2.1.3. Asymptotic functional linear regression approach

We conclude this subsection by describing a simplified linear model approach for large n_ds, which provides the basis for one of the early strategies for analyzing data from complex surveys (see Section 4.1). By a central limit theorem, conditional on n_ds,

$$\overline{y}_{1d} \sim_{\text{approx}} N\big(\mu_{1d}, n_d^{-1}\overline{y}_{1d}\left(1 - \overline{y}_{1d}\right)\big), \tag{24}$$

and so by the delta method, we have

$$\operatorname{logit}\overline{y}_{1d} \sim_{\text{approx}} N\Big(\operatorname{logit}\mu_{1d}, \left(n_d\overline{y}_{1d}\left(1 - \overline{y}_{1d}\right)\right)^{-1}\Big). \tag{25}$$

Now, we can write an approximate functional linear regression model (Bishop et al., 1975, p. 353; Grizzle et al., 1969) for $\operatorname{logit}\overline{y}_{1d}$ as

$$\operatorname{logit}\overline{y}_{1d} \approx A'_{x,d}\beta + \delta_d, \quad \delta_d \sim N\Big(0, \left(n_d\overline{y}_{1d}\left(1 - \overline{y}_{1d}\right)\right)^{-1}\Big). \tag{26}$$

It follows that the standard weighted least squares estimation method for linear models can now be used. Although this approach is attractive because of its simplicity, the transformation bias induced by linearization may seriously affect the finite sample behavior.

2.2. Model diagnostics under likelihood-based methods

We consider two types of diagnostics: informal (graphical plots) and formal (measures of model fit and their significance via testing).

2.2.1. Informal diagnostics (residual and quantile-quantile plots)

The standardized residuals based on the difference between observed and expected counts (like Pearson residuals under the selected model) provide an effective means of model checking. With aggregate level modeling, it is natural to look at domain level residuals, $r_{1d} = n_d\big(\overline{y}_{1d} - \mu_{1d}\big(\hat{\beta}^{\text{ml}}_{[p_{\text{R}}]}\big)\big)$ $\big(\overline{y}_{1d} = n_d^{-1}z_{1d}\big)$, for the final reduced model H_{R} because Pearson's χ^2 involves these residuals. The correlations among residuals $\{r_{1d}\}_{1\le d\le D}$ are due to two possible sources: one is the assumed correlation structure among model errors $(y_{1dk} - \mu_{1d}(\beta))$ at the unit level, and the other source is the use of model parameter estimates $\hat{\beta}^{\text{ml}}_{[p_{\text{R}}]}$, which induces correlations even if model errors are

uncorrelated. In the usual regression diagnostics for noncategorical data, impact of the second source is negligible for large sample sizes. If this were the case for CDA, then with independent model errors, it would have been reasonable to treat the standardized residuals, $\{r^*_{1d}\}$,

$$r^*_{1d} = r_{1d}/\mathrm{se}(r_{1d}); \quad \mathrm{Cov}(r_{1d})_{1\le d\le D} = R(\beta)\,\mathrm{diag}\{n_d u_{1d}(\beta)\}R'(\beta)\big|_{\hat\beta^{\mathrm{ml}}_{[p_R]}}; \quad (27)$$
$$R(\beta) = I_{D\times D} - \mathrm{diag}\{n_d u_{1d}(\beta)\}\,A_x J^{-1}_{\beta[p_R]} A'_x,$$

approximately as iid $N(0, 1)$ under correct model specification, where $se(r_{1d})$s as usual are square roots of the diagonal elements of $\mathrm{Cov}(r_{1d})_{1\le d\le D}$, the Taylor linearization of the residual vector $(r_{1d})_{1\le d\le D}$ is $R(\beta)(n_d(\overline{y}_{1d} - \mu_{1d}(\beta))_{1\le d\le D}$, $J_{\beta[p_R]}$ is as in (22), and A_x or $(A'_{x,d})_{1\le d\le D}$ as in (2) defined under H_R. Now, the residual plot could be checked for possible nonrandom pattern such as the one under dependence or heteroscedasticity. However, for CDA use of the estimate $\hat\beta^{\mathrm{ml}}_{[p_R]}$ in computing the standardized residuals $\{r^*_{1d}\}$ may introduce non-negligible correlations, which make it difficult to check for independence. Such dependence among $\{r^*_{1d}\}$ may be serious even for a large overall sample size n because the residuals are based on grouped or domain level data having large sample sizes n_d while the total number D of groups or domains remains bounded in general.

It follows that it might be better to first transform the residuals so that asymptotically, they behave the same whether or not $\beta_{[p_R]}$ is estimated. We can then transform them further to obtain Cholesky residuals (Houseman et al., 2004), which make their covariance matrix diagonal for ease in interpretation. For this purpose, we first consider the saturated model H_S with number of parameters $p_S(=D)$ by introducing extra parameters $\beta_{[p_S-p_R]}$ of dimension $(p_S - p_R)$ via construction of a $D\times(p_S - p_R)$ matrix $A_{x[p_S-p_R]}$ such that the augmented matrix $\big(A_{x[p_R]}, A_{x[p_S-p_R]}\big) \equiv A_{x[p_S]}$ has full rank p_S. The subscript $x[p_S - p_R]$ denotes new $(p_S - p_R)$ covariates with values given by the rows of $A_{x[p_S-p_R]}$ for each d while the $D\times p_R$ matrix $A_{x[p_R]}$ is similar to the old matrix A_x of (2). A simple way to construct $A_{x[p_S-p_R]}$ is to choose its $(p_S - p_R)$ columns as any subset of the D rows of $\big(I_{D\times D} - A_{x[p_R]}\big(A'_{x[p_R]}A_{x[p_R]}\big)^{-1}A'_{x[p_R]}\big)$ (which are orthogonal to columns of $A_{x[p_R]}$ by being part of a projection matrix on the orthocomplement space) such that $A'_{x[p_S-p_R]}A_{x[p_S-p_R]}$ is nonsingular. Thus, the model for μ_{1d} under H_S is given by (analogous to (6))

$$\mathrm{logit}\,\mu_{1d}\big(\beta_{[p_S]}\big) = A'_{x[p_R],d}\beta_{[p_R]} + A'_{x[p_S-p_R],d}\beta_{[p_S-p_R]} = A'_{x[p_S],d}\beta_{[p_S]}. \quad (28)$$

Next to make the asymptotic dependence of $\big\{r_{1d} = n_d\big(\overline{y}_{1d} - \mu_{1d}\big(\hat\beta^{\mathrm{ml}}_{[p_R]}\big)\big)\big\}_{1\le d\le D}$ on $\hat\beta^{\mathrm{ml}}_{[p_R]}$ negligible, we perform a nonsingular linear transformation using the formula (4) to obtain the score vector $\phi_{\beta[p_S]} = \big(\phi'_{\beta[p_R]}, \phi'_{\beta[p_S-p_R]}\big)'$ evaluated at $\beta_{[p_S]} = \big(\beta'^{\mathrm{ml}}_{[p_R]}, \beta'_{[p_S-p_R]} = 0'\big)'$; here, the first p_R score functions $\phi_{\beta[p_R]}$ are 0 at $\beta_{[p_S]} = \big(\beta'^{\mathrm{ml}}_{[p_R]}, \beta'_{[p_S-p_R]} = 0'\big)'$. The transformed residuals $(\tilde r_{1d})_{p_R+1\le d\le p_S}$, where $\tilde r_1 = \phi_{\beta[p_S-p_R]}$ evaluated at $\beta_{[p_S]} = \big(\beta'^{\mathrm{ml}}_{[p_R]}, \beta'_{[p_S-p_R]} = 0'\big)'$, do not depend on $\hat\beta^{\mathrm{ml}}_{[p_R]}$ asymptotically because at $\beta_{[p_S]} = \big(\beta'^{\mathrm{ml}}_{[p_R]}, \beta'_{[p_S-p_R]} = 0'\big)'$, $\phi_{\beta[p_S-p_R]}$ equals $\phi_{\beta[p_S-p_R|p_R]}$ defined from the formula (14) for nscore functions. Thus, the dimension of the transformed residuals is reduced to $(p_S - p_R)$ due to estimation of the unknown model parameters $\beta_{[p_R]}$. The residuals $\{\tilde r_{1d}\}_{p_R+1\le d\le p_S}$ can be transformed further by a lower triangular matrix–inverse of the

left Cholesky root of the covariance matrix $\Sigma_{\phi[p_S - p_R | p_R]}$ defined similarly to (16). The corresponding standardized residuals $\{\tilde{r}^*_{1d}\}_{p_R+1 \le d \le p_S}$ are approximately iid $N(0, 1)$ and plotted to check for the existence of any nonrandom pattern. Further, a Q-Q (quantile-quantile) plot of quantiles of the empirical residual distribution against those of the standard normal distribution can be used to check for normality.

2.2.2. Informal diagnostics (influential points)
Following Pregibon (1981) and Roberts et al. (1987), influential points could be identified using diagonals of the hat matrix corresponding to the linearized regression problem, analogous to that used in the Newton–Raphson iterations (5); that is, for the adjusted dependent variable $\overline{y}^*_{1d}$ defined below, consider the approximate linear regression model

$$\begin{aligned} \overline{y}^*_{1d} &\equiv \overline{y}_{1d} - [\mu_{1d}(\hat{\beta}_{[p_R]}) - u_{1d}(\hat{\beta}_{[p_R]}) A'_{x,d} \hat{\beta}_{[p_R]}] \\ &\simeq u_{1d}(\hat{\beta}_{[p_R]}) A'_{x,d} \beta + \delta_{1d},\ \delta_{1d} \sim_{\text{approx}} N(0, n_d^{-1} u_{1d}(\hat{\beta}_{[p_R]})). \end{aligned} \tag{29}$$

Now denoting the design matrix of the above linear regression by $X = (X'_d)_{1 \le d \le D}$, where X_d is $u_{1d}(\hat{\beta}_{[p_R]}) A_{x,d}$, and the error covariance matrix $\operatorname{diag}\{n_d^{-1} u_{1d}(\hat{\beta}_{[p_R]})\}$ by Σ_δ, the covariance matrix of the residuals $\{\overline{y}^*_{1d} - X'_d \hat{\beta}_{[p_R]}\}$ is obtained as $\Sigma_\delta^{1/2}(I - H)\Sigma_\delta^{1/2}$, where H is the hat matrix $\Sigma_\delta^{-1/2} X (X' \Sigma_\delta^{-1} X)^{-1} X' \Sigma_\delta^{-1/2}$, whose diagonal elements are between 0 and 1. These diagonal values can be plotted to check for possible high values that would indicate extreme points in the factor space. Next, the influence of these extreme points is checked by estimating the change in model parameters in the absence of each domain corresponding to the point deemed to be extreme.

2.2.3. Formal diagnostics (tests for model adequacy)
We assess model adequacy by computing p-values corresponding to tests of significance-of-fit (sof) and goodness-of-fit (gof). For assessing sof, we check the overall significance of predictors in the final model by computing the test statistic for the baseline model H_B (with $\dim \beta = p_B$) given the final model H_R (with $\dim \beta = p_R$). For instance, the score test from (12) is

$$Q_{\text{ml(sof)}}(H_B | H_R) = Q(\phi_{\beta[p_R]})\Big|_{\beta_{[p_R]} = \left(\hat{\beta}^{\text{ml}}_{[p_B]}, 0\right)'}. \tag{30}$$

The p-value is computed by referring to the upper tail of the $\chi^2_{p_R - p_B}$ distribution. This test is expected to be highly significant because significant predictors were selected in the modeling process. For assessing gof, we check the overall nonsignificance of the omitted predictors in an enlarged full model (such as the saturated model) by computing, for example, the score test statistic

$$Q_{\text{ml(gof)}}(H_R | H_F) = Q\left(\phi_{\beta[p_F]}\right)\Big|_{\beta_{[p_F]} = \left(\hat{\beta}^{\text{ml}}_{[p_R]}, 0\right)'}. \tag{31}$$

The p-value is computed by referring to the upper tail of the $\chi^2_{p_F - p_R}$ distribution. This test is expected to be highly insignificant if the model is adequate.

2.2.4. Formal diagnostics (measures of model fit and their significance levels)
In addition to the above tests for model adequacy, it would be useful to compute the corresponding measures of model fit and their significance levels or p-values for tests

based on them. For this purpose, we follow Estrella's (1998) generalization of R-square to nonlinear models. It is based on likelihood ratio tests for nested models and reduces to the usual R-square for linear models. For nested models such as $H_{\mathrm{B}} \subset H_{\mathrm{R}}$ corresponding to the sof test and the likelihood (3) for product binomial data consisting of AD counts $\{n_d \overline{y}_{ad}\}$ with number of linearly independent terms $\{n_d(\overline{y}_{ad} - \mu_{ad})\}$ being only D when $A = 2$, the Estrella's measure R_e^2 (between 0 and 1) is defined as

$$R_e^2 = 1 - (e_R^*/e_B^*)^{e_B^*}; \quad e_R^* = -2D^{-1} \log L\left(\hat{\beta}^{\mathrm{ml}}_{[p_R]}\right),$$
$$e_B^* = -2D^{-1} \log L\left(\hat{\beta}^{\mathrm{ml}}_{[p_B]}\right), \quad e_B^* > e_R^*. \tag{32}$$

If the term e_B^* is relatively large, then the measure can be simplified to $R_e^{2*} = 1 - \exp(-(e_B^* - e_R^*))$, approximately. It may be instructive to note that R_e^{2*} coincides with the usual expression of R^2 for a normal linear model. More specifically, for the linear model (with only the intercept under H_{B}), $y_i = x_i'\beta + \varepsilon_i$, $\varepsilon_i \sim_{\mathrm{iid}} N(0, \sigma^2)$, $i = 1, \ldots, n$, we have

$$-2n^{-1} \log L\left(\hat{\beta}^{\mathrm{ml}}_{[p]}\right) = 1 + \log(2\pi) + \log \hat{\sigma}^2, \quad \hat{\sigma}^2 = n^{-1} \sum_{i=1}^{n} \left(y_i - x_i' \hat{\beta}^{\mathrm{ml}}_{[p]}\right)^2, \tag{33}$$

which implies that $e_B^* - e_R^* = \log n^{-1} \sum_{i=1}^n (y_i - \overline{y})^2 - \log n^{-1} \sum_{i=1}^n \left(y_i - x_i' \hat{\beta}^{\mathrm{ml}}_{[p]}\right)^2$.

Similarly, for the final reduced model nested within the full model (i.e., $H_{\mathrm{R}} \subset H_{\mathrm{F}}$), corresponding to gof test, a second R_e^2 measure can be defined. For each measure, significance level corresponding to a suitable F-test statistic can be computed. For example, in the case of $H_{\mathrm{B}} \subset H_{\mathrm{R}}$, the test statistic $\left(R_e^2/(p_{\mathrm{R}} - p_{\mathrm{B}})\right)\left((1 - R_e^2)/(D - p_{\mathrm{R}})\right)^{-1}$ is referred to the upper tail of the $F_{p_{\mathrm{R}} - p_{\mathrm{B}}, D - p_{\mathrm{R}}}$ distribution to compute the significance level.

2.3. *Inferential testing under likelihood-based methods*

2.3.1. *Testing significance of regression parameters*

Often in data analysis, the model is selected based on a training sample, and we may wish to confirm significance of factor effects included in the final model through the validation sample. Also, it may be of interest to compute significance of certain factor effects or a group of them given the final model H_{R} since this may not be available as part of the stepwise selection procedure for model covariates. In any case, suppose we wish to test a hypothesis (denoted by H_{T}) nested within the final model H_{R}. For this purpose, we can, for example, use the score test from (12) to define $Q_{\mathrm{ml}}(H_{\mathrm{T}}|H_{\mathrm{R}})$, which has a $\chi^2_{p_{\mathrm{R}} - p_{\mathrm{T}}}$ distribution.

2.3.2. *Testing significance of difference between two domain means*

Suppose we wish to test the significance of the difference between $\mu_d(\beta)$ and $\mu_{d'}(\beta)$ under H_{R} for domains d and d'. For example, in the case of CCHS, the two domains may correspond to two age-gender groups. If we wish to use the score test, we first transform the parameter vector from $\beta_{[p_{\mathrm{R}}]}$ to $\theta_{[p_{\mathrm{R}}]}$ as follows:

$$\theta_1 = \mu_d(\beta) - \mu_{d'}(\beta), \quad \theta_l = \beta_l, \quad l = 2, \ldots, p_{\mathrm{R}}. \tag{34}$$

Now with $\phi_\theta = P'\phi_\beta$, $P = (\partial\beta/\partial\theta')$, and $\Sigma_{\phi(\theta)} = P'\Sigma_{\phi(\beta)}P$, and given H_{R}, the score test for testing $H_{\mathrm{T}} : \theta_1 = 0$ with $\theta_{[p_{\mathrm{R}}-1]}$ unspecified, can be easily obtained from (12). We could also use the Wald test for this problem, which would certainly be simpler in practice. However, as mentioned earlier, the score test would be preferable from the point of view of stability of the χ^2-approximation.

2.4. Inferential estimation under likelihood-based methods

2.4.1. Estimation of model parameters

Given the final model H_{R}, point and variance estimates of the model parameters (β) are given, respectively, by $\hat{\beta}^{\mathrm{ml}}_{[p_{\mathrm{R}}]}$ and $\Gamma_{\beta[p_{\mathrm{R}}]}$ as defined by (22) and evaluated at $\hat{\beta}^{\mathrm{ml}}_{[p_{\mathrm{R}}]}$. For interval estimation of a subset of β-parameters, $\beta_{[p_{\mathrm{R}}-p_{\mathrm{T}}]}$, that are under test while $\beta_{[p_{\mathrm{T}}]}$ being nuisance (as defined by the test model $H_{\mathrm{T}} \subset H_{\mathrm{R}}$), we can use Wald confidence intervals for a scalar parameter or confidence regions for a vector parameter based on the Wald statistic derived from the asymptotic distribution (22) under H_{R}. The use of observed information in the covariance matrix $\Gamma_{\beta[p_{\mathrm{R}}]}$ is preferable in general to the expected information in view of its anticipated improved finite sample performance (see Efron and Hinkley, 1978). Incidentally, the issue of lack of invariance to parameter transformation with the observed information matrix becomes mute when the goal is to estimate parameters and not hypothesis testing.

Improved confidence regions for test parameters $\beta_{[p_{\mathrm{R}}-p_{\mathrm{T}}]}$ can be obtained by the test inversion method based on the score statistic $Q^{c(\alpha)}_{\mathrm{ml}}(H_{\mathrm{T}}|H_{\mathrm{R}})$ defined in (14), where suitable estimates of the nuisance parameters $\beta_{[p_{\mathrm{T}}]}$ need to be substituted because the idea of test inversion basically consists of finding the region in the space of $\beta_{[p_{\mathrm{R}}-p_{\mathrm{T}}]}$ that fall in the acceptance region rather than setting the test parameters $\beta_{[p_{\mathrm{R}}-p_{\mathrm{T}}]}$ to zero under the null hypothesis. Here, use of a profile ml-estimator of $\beta_{[p_{\mathrm{T}}]}$ as a function of $\beta_{[p_{\mathrm{R}}-p_{\mathrm{T}}]}$ is desirable (see Godambe, 1991). However, in practice, we can use for convenience appropriate p_{T}-elements of the unrestricted ml-estimator $\hat{\beta}^{\mathrm{ml}}_{[p_{\mathrm{R}}]}$ as mentioned in Thompson (1997, Chapter 4, p. 138).

2.4.2. Estimation of domain means

For estimating a domain mean $\mu_{1d}(\beta)$ under H_{R}, the point estimator is easily obtained as inv logit$\big(A'_{x,d}\hat{\beta}^{\mathrm{ml}}_{[p_{\mathrm{R}}]}\big)$. The variance estimator is obtained by using Taylor linearization of the inverse logit function, and the asymptotic variance of $A'_{x,d}\hat{\beta}^{\mathrm{ml}}_{[p_{\mathrm{R}}]}$ is given by $A'_{x,d}\Gamma_{\beta[p_{\mathrm{R}}]}A_{x,d}$ $\big($and evaluated at $\hat{\beta}^{\mathrm{ml}}_{[p_{\mathrm{R}}]}\big)$, where $\Gamma_{\beta[p_{\mathrm{R}}]}$ defined in (22) reduces to $I^{-1}_{\beta[p_{\mathrm{R}}]}$. More specifically, the variance estimator of inv logit$\big(A'_{x,d}\hat{\beta}^{\mathrm{ml}}_{[p_{\mathrm{R}}]}\big)$ is given by $u^2_{1d}(\beta_{[p_{\mathrm{R}}]})A'_{x,d}I^{-1}_{\beta[p_{\mathrm{R}}]}A_{x,d}$ and evaluated at $\hat{\beta}^{\mathrm{ml}}_{[p_{\mathrm{R}}]}$, where $u_{1d}(\beta)$ is defined as in (5). For interval estimation, one can use the logit-Wald method where a symmetric normality based interval is first obtained from the asymptotic distribution

$$A'_{x,d}\Big(\hat{\beta}^{\mathrm{ml}}_{[p_{\mathrm{R}}]} - \beta_{[p_{\mathrm{R}}]}\Big) \sim_{\mathrm{approx}} N\big(0,\ A'_{x,d}\Gamma_{\beta[p_{\mathrm{R}}]}A_{x,d}\big)\Big|_{\beta=\hat{\beta}^{\mathrm{ml}}_{[p_{\mathrm{R}}]}}, \tag{35}$$

and then the asymmetric interval is obtained by the inverse logit transformation, see Newcombe (2001) for some comments on the logit-Wald method. As an alternative to avoid possible instability of the Wald method in the above normal approximation, we can use test inversion based on the score statistic $Q_{\mathrm{ml}}(H_{\mathrm{T}}|H_{\mathrm{R}})$, where now H_{T} represents

the transformed test parameter θ_{p_R} (defined by $\mu_{1d}(\beta)$) while the nuisance parameters are defined simply as $\theta_{[p_R-1]} = \beta_{[p_R-1]}$, similar to (34).

2.4.3. Estimation of contrasts and odds ratio

For estimating a contrast $\mu_{d_1}(\beta) - \mu_{d_2}(\beta)$ and the odds ratio $\mu_{d_1}(\beta)\big(1-\mu_{d_1}(\beta)\big)^{-1}\big/ \mu_{d_2}(\beta)\big(1-\mu_{d_2}(\beta)\big)^{-1}$, the point and variance estimates can be obtained along the lines for $\mu_{d_1}(\beta)$, while for interval estimation, we can either use Wald (logit-Wald for odds ratio) or test inversion based on the score test.

3. Quasi-likelihood methods

The semiparametric framework of ql-estimation is conducive for analysis with complex survey data (to be discussed in the next section) as it does not require specification beyond the first two moments (for optimal estimation) and only first moments (for nonoptimal estimation) because in survey sampling, even the second moment of an estimate of the finite population total is generally not available due to too many unknown parameters consisting of second-order inclusion probabilities. In fact, even for the case of infinite populations, it may be difficult to specify the second moment such as in the case of correlated binomial observations when the data is at the cluster level (see, e.g., McCullagh and Nelder, 1989, Chapters 4 and 9). For the CCHS example, individuals from the same cluster (e.g., neighborhood) may have correlated responses that will lead to the correlated Binomial case. It may be remarked that in the interest of parsimony, the analyst often prefers a simple model, and in the process, he may omit from the model some known (and not to mention the unknown) covariates as well as some random cluster effects. This implies that the covariance matrix (conditional on the covariates in the model) may not be correctly specified. In practice, this can be corrected approximately by introducing an overdispersion parameter as a multiplicative factor in the error covariance structure (see, e.g., McCullagh and Nelder, 1989, Chapters 4 and 9). Correction for overdispersion is especially important for categorical data because overdispersion cannot be easily subsumed within the error variance that is functionally dependent on the mean.

Unlike the previous subsection, the unit level data, $\{y_{adk}\}_{1\le k\le n_d, 1\le d\le D, 1\le a\le A}$, are not independent due to intracluster correlations, and without making any further assumptions about the intracluster dependence, it is not feasible to specify the joint likelihood. In this section, although the model is at the aggregate level, we prefer to work with the unit level data to be able to incorporate later on the sampling weights at the unit level in q-score functions when dealing with survey data. For the CCHS example considered in the previous section, the unit-level covariates x_{dk} (indicating age or gender factors) take the common value of 1 if unit k lies in the domain d and if the age or gender indicated by the x-variable matches with that defined by d, and it takes the value of 0 otherwise. Thus, $x_{dk} = A_{x,d}$ for all k in d, and the logit model (2) can be reformulated at the unit level without complete specification of the covariance structure as follows.

$$\begin{aligned} &\textit{Obs. Eqn.}\ y_{1dk} = \mu_{1dk} + \varepsilon_{1dk}, \quad y_{1dk} \sim \text{Ber}(\mu_{1dk}), \\ &\qquad\qquad\qquad\qquad\quad y_{1dk}\text{'s may be cluster-correlated} \\ &\textit{Link Eqn.}\ \text{logit}\,\mu_{1dk} \equiv \log\big\{(1-\mu_{1dk})^{-1}\mu_{1dk}\big\} = A'_{x,d}\beta. \end{aligned} \tag{36}$$

The above model is simply a unit-level version of the aggregate level model because $\mu_{1dk} = \mu_{1d}$, that is, the mean is common for all k in the domain d subpopulation since $x_{dk} = A_{x,d}$. In the following, we consider as before the four parts of data analysis under ql-estimation that will be of two types, nonoptimal if the covariance structure is not completely specified, and optimal if it is.

3.1. Model selection under quasi-likelihood

3.1.1. Quasi-score functions

All the results from Subsection 2.1 basically go through when the score function $\phi_{\beta[p]}$ is replaced by the q-score function $\phi_{\mathrm{ql}(\beta[p])}$ defined below. Following Liang and Zeger (1986) on generalized estimating equations under the working independence assumption of y_{1dk}s, the ql-estimator $\hat{\beta}^{\mathrm{ql}}_{[p]}$ for the model (36) is obtained as a solution of the q-score function $\phi_{\mathrm{ql}(\beta[p])}$ set to zero. The vector of q-score functions is given by

$$\phi_{\mathrm{ql}(\beta[p])} \equiv \sum_{d=1}^{D}\sum_{k=1}^{n_d} A_{x,d}(y_{1dk} - \mu_{1dk}(\beta)) = \sum_{d} A_{x,d} n_d\big(\overline{y}_{1d} - \mu_d(\beta)\big), \quad (37)$$

which turns out to be identical to the score function (4). Note that the above q-score function is not optimal in the sense of Godambe and Thompson (1989) because of the use of the working covariance matrix, see Godambe and Kale (1991), Singh and Rao (1997), and Chapter 26 of this handbook for simple reviews of optimal EFs. For (37) to be optimal, it should take the form of $E(-\partial\psi/\partial\beta')'\Sigma_{\psi}^{-1}\psi$, where $\psi = n_d(\overline{y}_{1d} - \mu_{1d}(\beta))_{1\le d\le D}$, $E(-\partial\psi/\partial\beta') = A_x$, which is a $(D \times p)$ matrix, and Σ_{ψ} is the covariance matrix of elementary d-level EFs ψ, which, in fact, is not specified under (36). Although the term ql-estimator was originally introduced by Wedderburn (1974) and later by Godambe and Heyde (1987) for general optimal EFs, here we will use the qualifier "optimal" for ql-estimation (oql, for short) when we want to emphasize the optimality, else we will simply use the term ql-estimation in a broad sense.

The optimality of EFs refers to the property that in the class of all EFs defined by linear combinations of elementary EFs ψ, the asymptotic covariance $\tilde{\Gamma}_{\mathrm{ql}(\beta[p])}$ of the resulting estimator $\big(\tilde{\Gamma}_{\mathrm{ql}(\beta[p])}$ having a sandwich form like $\hat{\Gamma}_{\mathrm{ql}(\beta[p])}$of (39) given below except that $J_{\mathrm{ql}(\beta[p])}$ is replaced by $I_{\mathrm{ql}(\beta[p])}\big)$ is minimized for the optimal EF under the partial order of non-negative definite matrices (see, Godambe and Heyde (1987) and McCullagh and Nelder, 1989, Section 9.5). The optimal EF also has an important finite sample property in terms of maximizing the information in EFs as defined by Godambe (1960) and Godambe and Thompson (1989). To compute $\hat{\Gamma}_{\mathrm{ql}(\beta[p])}$, a robust consistent variance estimator of $\phi_{\mathrm{ql}(\beta[p])}$ is first obtained using the commonly used technique of independent clusters (like the with replacement assumption of primary sampling units (PSUs) in survey sampling), see Liang and Zeger (1986) in the context of longitudinal data and Bieler and Williams (1995) in the context of cluster-correlated data. Suppose that the number of clusters in each stratum h (such as a subprovincial area $1 \le h \le L$) is c_h with cluster sample sizes n_{hr} for the rth cluster. Then, $\phi_{\mathrm{ql}(\beta[p])}$ is a sum of L independent strata totals $\phi_{\mathrm{ql}(\beta[p],h)}$, where $\phi_{\mathrm{ql}(\beta[p],h)}$ itself is a sum of c_h independent cluster totals $\phi_{\mathrm{ql}(\beta[p],hr)}$. Assuming that the total number of clusters $c\big(=\sum_n c_h\big)$ is large (the cluster sample sizes n_{hr} are typically small but the number of sampled individuals $n_h\big(=\sum_{r=1}^{c_h} n_{hr}\big)$ for each stratum h may be large), a consistent estimate of the covariance

$\Sigma_{\mathrm{ql}(\phi[p])}$ can be obtained as

$$\hat{\Sigma}_{\mathrm{ql}(\phi[p])} = \sum_{h=1}^{L} \frac{c_h}{c_{h^{-1}}} \sum_{r=1}^{c_h} \left(\phi_{\mathrm{ql}(\beta[p],hr)} - \overline{\phi}_{\mathrm{ql}(\beta[p],h)}\right)\left(\phi_{\mathrm{ql}(\beta[p],hr)} - \overline{\phi}_{\mathrm{ql}(\beta[p],h)}\right)' \Big|_{\hat{\beta}^{\mathrm{ql}}_{[p]}};$$

$$\overline{\phi}_{\mathrm{ql}(\beta[p],h)} = c_h^{-1} \sum_{r=1}^{c_h} \phi_{\mathrm{ql}(\beta[p],hr)}, \quad \phi_{\mathrm{ql}(\beta[p],hr)} = \sum_{d=1}^{D} \sum_{k=1}^{n_{hr}} A_{x,d}(y_{1dk} - \mu_{1d}(\beta)) 1_{dk(hr)}, \tag{38}$$

where $1_{dk(hr)}$ is the indicator function taking the value of 1 if the unit k of domain d is in cluster r of stratum h, and 0 otherwise. It follows from (37) that the covariance matrix $\Gamma_{\mathrm{ql}(\beta[p])}$ of $\hat{\beta}^{\mathrm{ql}}_{[p]}$ can be estimated via Taylor expansion in a sandwich form, analogous to (22) as

$$\hat{\Gamma}_{\mathrm{ql}(\beta[p])} = J^{-1}_{\mathrm{ql}(\beta[p])} \hat{\Sigma}_{\mathrm{ql}(\phi[p])} J'^{-1}_{\mathrm{ql}(\beta[p])} \Big|_{\hat{\beta}^{\mathrm{ql}}_{[p]}}; \quad J_{\mathrm{ql}(\beta[p])} = -\partial \phi_{\mathrm{ql}(\beta[p])} / \partial \beta'_{[p]}. \tag{39}$$

The matrix $J_{\mathrm{ql}(\beta[p])}$ for the q-score function (37) is identical to that for the score function (4) and is given in (5). It can be termed as the observed q-information matrix (see, e.g., Fahrmeier and Tutz, 2001, p. 441). However, due to nonoptimality of the q-score $\phi_{\mathrm{ql}(\beta[p])}$, there is no information unbiasedness, that is, $\Sigma_{\mathrm{ql}(\phi[p])}$ is not equal to the expected q-information matrix $I_{\mathrm{ql}(\beta[p])} (= E(J_{\mathrm{ql}(\beta[p])}))$, unlike the case of $\phi_{\beta[p]}$ of Section 2. In fact, the true covariance matrix $\Sigma_{\mathrm{ql}(\phi[p])}$ is not even specified.

3.1.2. Optimal quasi-score functions

Here, we need to assume a more restrictive sampling scheme to specify second moments. Suppose the clusters r do not cut across domains d so that domains can be treated as strata by conditioning on n_ds. Now, as in the product binomial case of Section 2, if for the model (36), we also assume independence of cluster totals within and between domains and that for each domain d, the functional form of the variance of the cluster totals $n_{dr}\overline{y}_{1dr} (r = 1, \ldots, c_d)$ as $\sigma_0^2 u_{1d}(\beta)$, where n_{dr} is the number of observations in domain d from cluster r, and σ_0^2 is the overdispersion parameter (McCullagh and Nelder, 1989, Chapter 4). Then the oq-score function (oq for optimal quasi) based on the elementary EFs $n_{dr}(\overline{y}_{1dr} - \mu_{1d}(\beta))$ (it turns out that it is sufficient to work with cluster level totals within each domain) is given by

$$\phi_{\mathrm{oql}(\beta[p])} = \sigma_0^{-2} \sum_d A_{x,d} n_d \left(\overline{y}_{1d} - \mu_{1d}(\beta)\right), \tag{40}$$

which is identical to the q-score $\phi_{\mathrm{ql}(\beta[p])}$ of (37), except for the multiplicative factor due to the overdispersion parameter. Its covariance, however, can now be estimated more efficiently than in the nonoptimal case and is given by

$$\hat{\Sigma}_{\mathrm{oql}(\phi[p])} = \hat{\sigma}_0^{-2} \sum_d n_d u_{1d}(\beta) A_{x,d} A'_{x,d} \Big|_{\hat{\beta}^{\mathrm{oql}}_{[p]}};$$

$$\hat{\sigma}_0^2 = \frac{1}{c - p} \sum_{d=1}^{D} \sum_{r=1}^{c_d} \frac{n_{dr}\left(\overline{y}_{1dr} - \mu_{1d}(\beta)\right)^2}{u_{1d}(\beta)} \Big|_{\hat{\beta}^{\mathrm{oql}}_{[p]}}, \tag{41}$$

where, as before, c is the total number of clusters and the estimator $\hat{\beta}^{\text{oql}}_{[p]}$ is defined as the solution of the optimal EF $\phi_{\text{oql}(\beta[p])}$ set to zero, which is solved by the Newton–Raphson iterative procedure as in (5). The above method of moments estimator of the overdispersion parameter σ_0^2 follows from McCullagh and Nelder (1989, Chapter 4). The approximate covariance of the optimal estimator $\hat{\beta}^{\text{oql}}_{[p]}$ is now given by

$$\Gamma_{\text{oql}(\beta[p])} = J^{-1}_{\text{oql}(\beta[p])}\Sigma_{\text{oql}(\phi[p])}J'^{-1}_{\text{oql}(\beta[q])}; \quad J_{\text{oql}(\beta[p])} = -\partial\phi_{\text{oql}(\beta[p])}/\partial\beta'_{[p]}, \tag{42}$$

which reduces to inverse of the expected information matrix, $I^{-1}_{\text{oql}(\beta[p])}$, because of $J_{\text{oql}(\beta[p])}$ being equal to $I_{\text{oql}(\beta[p])}$ for the logit model and information unbiasedness of the optimal q-score function. Note that the matrix $J_{\text{oql}(\beta[p])}$ is simply σ_0^{-2} times $J_{\text{ql}(\beta[p])}$ of (5).

A useful simplified alternative to (40) and (41), although not equivalent, was suggested by Rao and Scott (1992) in the context of clustered binary data. Here, the domain total z_{1d} and the domain size n_d are adjusted by a domain-specific factor (like the design or dispersion effect of $\bar{y}_{1d}$) so that the adjusted total $\tilde{z}_{1d}$ behaves approximately (for large n_d) like a Binomial variable with mean $\mu_{1d}(\beta)$ and the index parameter $\tilde{n}_d$, the adjusted domain size. This may be quite appealing in practice as a multipurpose technique. However, the estimated covariance matrix (41) of the oql method is likely to be more stable as it combines information over all the domains. A similar simplified approach for clustered Poisson data was also suggested by Rao and Scott (1999).

3.1.3. Tests for nested hypotheses

For testing the nested model $H_2 \subset H_1$, we can easily define a q-score test statistic $Q_{\text{ql}}(H_2|H_1)$ based on q-nscore function similar to (14) along the lines of Boos (1992). More specifically, we reject H_2 in favor of H_1 for large values of $Q_{\text{ql}}(H_2|H_1)$ by referring to the upper tail of the $\chi^2_{p_1-p_2}$ distribution, where

$$\begin{aligned} Q_{\text{ql}}(H_2|H_1) &= Q(\phi_{\text{ql}(\beta[p_1-p_2|p_2])})\Big|_{\beta_{[p_1]}=\left(\hat{\beta}'^{\text{ql}}_{[p_2]},\hat{\beta}'^{\text{ql}}_{[p_1-p_2]}=0\right)'}; \\ \phi_{\text{ql}(\beta[p_1-p_2\,|\,p_2])} &= \phi_{\text{ql}(\beta[p_1-p_2])} - I_{\text{ql}(\beta[p_1-p_2,p_2])}I^{-1}_{\text{ql}(\beta[p_2])}\phi_{\text{ql}(\beta[p_2])} \\ &\equiv F_{\text{ql}}\phi_{\text{ql}(\beta[p_1])}(F_{\text{ql}} = (-I_{\text{ql}(\beta[p_1-p_2,p_2])}I^{-1}_{\text{ql}(\beta[p_2])}, I_{(p_1-p_2)\times(p_1-p_2)})), \end{aligned} \tag{43}$$

and where the expected q-information matrix $I_{\text{ql}(\beta[p])} = E(J_{\text{ql}(\beta[p])})$ is partitioned as

$$I_{\text{ql}(\beta[p_1])} = \begin{pmatrix} I_{\text{ql}(\beta[p_2])} & I_{\text{ql}(\beta[p_2,p_1-p_2])} \\ I_{\text{ql}(\beta[p_1-p_2,p_2])} & I_{\text{ql}(\beta[p_1-p_2])} \end{pmatrix}. \tag{44}$$

In (44), $J_{\text{ql}(\beta[p])}$ is defined by (39), and the off-diagonal term $I_{\text{ql}(\beta[p_2,p_1-p_2])}$, for example, is $E(J_{\text{ql}(\beta[p_2,p_1-p_2])})$, $J_{\text{ql}(\beta[p_2,p_1-p_2])}$ being $-\partial\phi_{\text{ql}(\beta[p_2])}/\partial\beta'_{[p_1-p_2]}$. Moreover, the covariance of $\phi_{\text{ql}(\beta[p_1-p_2|p_2])}$ is given by $\Sigma_{\text{ql}(\phi[p_1-p_2|p_2])} = F_{\text{ql}}\Sigma_{\text{ql}(\phi[p_1])}F'_{\text{ql}}$. Note that, as with the nscore function in (14), the q-nscore function $\phi_{\text{ql}(\beta[p_1-p_2|p_2])}$ in (43) evaluated at any $\sqrt{n}$-consistent estimator $\hat{\beta}_{[p_2]}$ behaves asymptotically as if $\beta_{[p_2]}$ is known. It may be remarked that with q-score tests, it is not possible in general to write an asymptotically equivalent version as a difference of two Q_{ql}-tests, that is, $Q_{\text{ql}}(H_2|H_S) - Q_{\text{ql}}(H_1|H_S)$, analogous to (20), unless $\Sigma_{\text{ql}(\beta[p])}$ is proportional to $I_{\text{ql}(\beta[p])}$. However, with the optimal q-score function (40) under second-moment assumptions, this is possible due to

information unbiasedness (i.e., $I_{\mathrm{ql}(\beta[p])} = \Sigma_{\mathrm{ql}(\phi[p])}$) because a decomposition like (17) is then feasible. The Wald test can, of course, be defined as in (23) based on the asymptotic normality result,

$$\hat{\beta}^{\mathrm{ql}}_{[p_1]} - \beta_{[p_1]} \sim_{\mathrm{approx}} N(0, \Gamma_{\mathrm{ql}(\beta[p_1])})\Big|_{\hat{\beta}^{\mathrm{ql}}_{[p_1]}}, \tag{45}$$

where $\Gamma_{\mathrm{ql}(\beta[p_1])}$ is defined similar to (39). In practice, it may be of interest to include other domain or cell-level covariates not defined by the cross-classifying variables used for the contingency table (such as the average number of hospital admissions for domain d with asthma as the main diagnosis for the CCHS example, which can be obtained from administrative sources) for improved prediction. The regression-type formulation (2) for logit models easily accommodates such covariates unlike the traditional formulation (1).

3.2. Model diagnostics under quasi-likelihood

For informal diagnostics based on residuals under the final reduced model H_{R}, we can use the asymptotic normality of Cholesky residuals obtained from $r_{1d} = n_d\big(\overline{y}_{1d} - \mu_{1d}\big(\hat{\beta}^{\mathrm{ql}}_{[p_{\mathrm{R}}]}\big)\big)$ as in Section 2.2. Other informal diagnostics based on Q-Q plots and influential points also carry over. In terms of formal diagnostics for tests of model adequacy, q-score tests for both sof ($H_{\mathrm{B}} \subset H_{\mathrm{R}}$) and gof ($H_{\mathrm{R}} \subset H_{\mathrm{F}}$) can be easily constructed analogous to (30) and (31). For Estrella's R^2 measures of model fit, however, we need the likelihood. For this purpose, consider the approximate Gaussian likelihood of the vector of summary statistics $\hat{\beta}^{\mathrm{ql}}_{[p_{\mathrm{S}}]}$ under the saturated model (these summary statistics could be deemed as quasi-sufficient under the quasi-likelihood analogous to the asymptotic sufficiency of ml-estimators under likelihood) which is given by

$$\hat{\beta}^{\mathrm{ql}}_{[p_{\mathrm{S}}]} - \beta_{[p_{\mathrm{S}}]} \sim_{\mathrm{approx}} N\big(0,\ \Gamma_{\mathrm{ql}(\beta[p_{\mathrm{S}}])}\big)\Big|_{\beta=\hat{\beta}^{\mathrm{ql}}_{[p_{\mathrm{S}}]}}. \tag{46}$$

The R^{2*}_e measure of (32) for the baseline model nested within the final reduced model (i.e., $H_{\mathrm{B}} \subset H_{\mathrm{R}}$) from the normal likelihood based on summary statistics $\hat{\beta}^{\mathrm{ql}}_{[p_{\mathrm{S}}]}$ is given by

$$1 - \exp\Bigg\{-D^{-1}\Bigg(Q\Big(\hat{\beta}^{\mathrm{ql}}_{[p_{\mathrm{S}}]} - \beta_{[p_{\mathrm{S}}]}\Big)\Big|_{\beta_{[p_{\mathrm{S}}]}=\left(\hat{\beta}'^{\mathrm{ql}}_{[p_{\mathrm{B}}]},\beta_{[p_{\mathrm{S}}-p_{\mathrm{B}}]}=0\right)'} - Q\Big(\hat{\beta}^{\mathrm{ql}}_{[p_{\mathrm{S}}]} - \beta_{[p_{\mathrm{S}}]}\Big)\Big|_{\beta_{[p_{\mathrm{S}}]}=\left(\hat{\beta}'^{\mathrm{ql}}_{[p_{\mathrm{R}}]},\beta_{[p_{\mathrm{S}}-p_{\mathrm{R}}]}=0\right)'}\Bigg)\Bigg\}, \tag{47}$$

where $p_{\mathrm{S}} = D$, and for convenience, the common covariance matrix $\Gamma_{\mathrm{ql}(\beta[p_{\mathrm{S}}])}$ (evaluated at $\hat{\beta}^{\mathrm{ql}}_{[p_{\mathrm{S}}]}$) is used for the two quadratic forms in (47), which explains why only the kernel of the Gaussian log-likelihood shows up and not the other terms. For large n_ds, the exponent term in (47) can also be expressed equivalently as the usual Wald statistic for nested models, $Q_{\mathrm{W}}(H_{\mathrm{B}}|H_{\mathrm{R}})$ from (23) based on $\hat{\beta}^{\mathrm{ql}}_{[p_{\mathrm{R}}]}$ under H_{R}, or a more stable version for finite samples given by the q-score statistic $Q_{\mathrm{ql}}(H_{\mathrm{B}}|H_{\mathrm{R}})$ as defined in (43). Similarly, a second R^{2*}_e measure for the final reduced model nested within the full model, that is, $H_{\mathrm{R}} \subset H_{\mathrm{F}}$, can also be defined.

3.3. Inferential testing and estimation under quasi-likelihood

As was described in Section 2.3 under likelihood-based methods, all the tests of significance of a single model parameter or groups of them, and tests for significance of

difference between domain means can be carried out when the score test is replaced by the q-score test under the present quasi-likelihood approach.

For estimating model parameters $\beta_{[\mathrm{R}]}$ under H_{R}, the main difference from the likelihood-based methods is that the covariance of the point estimator $\hat{\beta}^{\mathrm{ql}}_{[p_{\mathrm{R}}]}$ has the sandwich form (39) unless the q-score function is optimal. For interval estimation via test inversion based on the q-score statistic for a test model H_{T} nested within H_{R}, we consider the q-nscore function as in (43) with the obvious interpretation of $\beta_{[p_{\mathrm{R}}-p_{\mathrm{T}}]}$ as test parameters and $\beta_{[p_{\mathrm{T}}]}$ as nuisance parameters. For point, variance, and interval estimation of domain means and their contrasts, results from Subsection 2.4 carry over with natural modifications for ql-estimation.

4. Weighted quasi-likelihood methods

So far we considered CDA for simple surveys. The example of CCHS is actually a complex survey with subprovincial areas as strata and multistage cluster sampling of households within strata followed by selection of one individual per household. As mentioned in the introduction, it is important to take the design into account to avoid selection bias in model parameter estimates. It is known that large sample consistent estimators of FPQs such as subpopulation totals can be obtained from H-T estimators or calibrated versions of design-weighted sample sums. To define suitable FPQs for estimating model parameters from complex survey data, we consider two phases of randomization: first, the ξ-randomization for the generation of the finite population from a conceptual infinite superpopulation (the population model) and second, the π-randomization for the generation of the sample from the finite population. It is assumed that the sample size n is much smaller than the population size N. Consider the following aggregate-level model for the binary data $\{y_{1dk}\}_{1\le k\le n_d,\ 1\le d\le D}$ expressed at the unit level to incorporate unit-level sampling weights in the weighted quasi- or wq-score functions defined below.

Phase I (ξ–randomization for the finite population U of size N):

$$y_{1dk} = \mu_{1dk} + \varepsilon_{1dk}, \quad \operatorname{logit}\mu_{1dk} = A'_{x,d}\beta, \quad y_{1dk} \sim_{\mathrm{ind}} \mathrm{Ber}(\mu_{1dk}), \\ 1 \le k \le N_d, \quad 1 \le d \le D, \tag{48}$$

Phase II (π–randomization for the sample s of size n): An arbitrary complex design with first order inclusion probabilities π_{dk} ($=\Pr(k \in s | k \in U_d)$, known atleast for the sampled units), and positive but possibly unkown second order inclusion probabilities, $\pi_{dk,d'l}(=\Pr(k, l \in s | k \in U_d, l \in U_{d'}))$, where U_d, $U_{d'}$ denote respectively domains d and d'.

For the above model, define the q-score function (or census EF, the subscript U below denotes the finite universe) at phase I and wq-score function (or sample EF) at phase II as an H-T estimator of the FPQ defined by the census EF as follows:

$$\text{Census EF: } \phi_{\mathrm{ql}(\beta[p],U)} = \sum_{d=1}^{D}\sum_{k=1}^{N_d} A_{x,d}(y_{1dk} - \mu_{1d}(\beta)) \\ = \sum_{d=1}^{D} A_{x,d} N_d (A_{y_1,d} - \mu_{1d}(\beta)), \tag{49a}$$

$$\text{Sample EF: } \phi_{\text{wq}(\beta[p])} = \sum_{d=1}^{D}\sum_{k=1}^{n_d} w_{dk} A_{x,d}(y_{1dk} - \mu_{1d}(\beta))$$
$$= \sum_{d=1}^{D} A_{x,d}\hat{N}_d(\overline{y}_{1dw} - \mu_{1d}(\beta)), \tag{49b}$$

where N_d is the domain population size and $A_{y_1,d}$ is the domain average ($\overline{Y}_{1d}$ being the traditional notation) analogous to $A_{x,d}$ of Section 2. The sampling weight for the kth sampling unit in domain d is w_{dk} defined as the inverse of the sample inclusion probability π_{dk} or a calibrated version of it like the poststratified weight (see Kott, Chapter 25). The estimator $\hat{N}_d$ is simply the estimated domain population count obtained as $\sum_{k=1}^{n_d} w_{dk}$ and $\overline{y}_{1dw} = \hat{N}_d^{-1}\sum_{k=1}^{n_d} y_{1dk} w_{dk}$. It is assumed that for large n, the domain sample size $n_d > 0$ for all d with high probability.

The q-score function (49a) is an optimal EF (defined in Section 3.1) under ξ-randomization, but the wq-score function (49b) is not an optimal EF under the compound $\pi\xi$-randomization because it only uses correct mean specification of $\hat{N}_d(\overline{y}_{1dw} - \mu_{1d}(\beta))$, in that it is zero in expectation under $\pi\xi$-randomization. However, it uses the identity matrix as the working covariance of the vector $(\hat{N}_d(\overline{y}_{1dw} - \mu_{1d}(\beta)))_{1\le d\le D}$ and not the correct covariance which, in fact, is not available. Nevertheless, the wq-score function has theoretically desirable properties as shown by Godambe and Thompson (1986). Note that one may also use the pseudo ml approach (see, e.g., Binder, 1983; Skinner, 1989, p. 80) as an alternative to the wq-score function. The use of the latter approach generally gives identical results depending on the specification of the finite-population (or the census) likelihood. In the pseudo ml approach, the census log-likelihood is first estimated using sampling weights, and then the estimated log-likelihood (termed pseudo as it is not the likelihood) is maximized to obtain pseudo ml estimators. On the other hand, the wql-approach starts with census EF (under only first two moment assumptions), which is then estimated using sampling weights to obtain sample EF. It turns out that all the results of Section 3 essentially carry over to complex surveys except for some important nuances as described below under each part of data analysis. The review is based on the works of Binder and Pathak (1994) and Rao et al. (1998).

4.1. Model selection under weighted quasi-likelihood

4.1.1. wq-score tests

For the wq-score function given by (49b), the covariance matrix $\Sigma_{\text{wq}(\phi[p])}$ under the compound $\pi\xi$-distribution can be expressed as a sum of two parts, (here, $V_{(.)}$ denotes the variance–covariance operator under the distribution denoted by the subscript)

$$\Sigma_{\text{wq}(\phi[p])} = E_\xi V_\pi(\phi_{\text{wq}(\beta[p])}) + V_\xi E_\pi(\phi_{\text{wq}(\beta[p])})$$
$$= E_\xi V_\pi(\phi_{\text{wq}(\beta[p])}) + V_\xi(\phi_{q(\beta[p],U)}), \tag{50}$$

where the second part is of much smaller order than the first part under the usual condition that the sampling fraction at the unit level is small although the sample size is assumed to be large; in the case of cluster sampling, the condition of a small sampling fraction at the cluster level is needed (Pfeffermann, 1993). So, under regularity conditions, a

consistent estimate of $\Sigma_{\mathrm{wq}(\phi[p])}$ is given by a consistent estimate of the design-based covariance $V_\pi(\phi_{\mathrm{wq}(\beta[p])})$, to be denoted by $\hat{\Sigma}_{\mathrm{wq}(\phi[p])}$. Now, under the usual assumption of with replacement sampling of PSUs within each stratum such as in stratified multistage unequal probability cluster sampling, a consistent estimator, $\hat{\Sigma}_{\mathrm{wq}(\phi[p])}$, of the covariance of the wq-score function (see, e.g., Wolter, 2007, Chapter 2, p. 47) is given below, which is similar to (38) for the q-score function except that clusters are replaced by PSUs selected in the design stratum h and q-scores by wq-scores, all evaluated at $\hat{\beta}^{\mathrm{wql}}_{[p]}$, the solution of the wq-score function. We have

$$\hat{\Sigma}_{\mathrm{wq}(\phi[p])} = \sum_{h=1}^{L} \frac{c_h}{c_h - 1} \sum_{r=1}^{c_h} \left(\phi_{\mathrm{wq}(\beta[p],hr)} - \overline{\phi}_{\mathrm{wq}(\beta[p],h)}\right) \times \left(\phi_{\mathrm{wq}(\beta[p],hr)} - \overline{\phi}_{\mathrm{wq}(\beta[p],h)}\right)' \Big|_{\hat{\beta}^{\mathrm{wql}}_{[p]}};$$

$$\overline{\phi}_{\mathrm{wq}(\beta[p],h)} = c_h^{-1} \sum_{r=1}^{c_h} \phi_{\mathrm{wq}(\beta[p],hr)},$$

$$\phi_{\mathrm{wq}([p],hr)} = \sum_{d=1}^{D} \sum_{k=1}^{n_{hr}} A_{x,d}(y_{1dk} - \mu_{1d}(\beta_{[p]}))w_{dk}1_{dk(hr)}, \tag{51}$$

where the wq-score equation (49b) is solved by the Newton–Raphson iterative method to obtain $\hat{\beta}^{\mathrm{wql}}_{[p]}$ analogous to (5). Note that by using the Taylor expansion under regularity conditions,

$$\phi_{\mathrm{wq}(\beta[p])} \approx J_{\mathrm{wq}(\beta[p])}\left(\hat{\beta}^{\mathrm{wql}}_{[p]} - \beta_{[p]}\right), \quad \text{where } J_{\mathrm{wq}(\beta[p])} = -\partial\phi_{\mathrm{wq}(\beta[p])}/\partial\beta'_{[p]}, \tag{52}$$

where $J_{\mathrm{wq}(\beta[p])}$ is the observed wq-information matrix. For the wq-score function (49b), $J_{\mathrm{wq}(\beta[p])}$ equals $\sum_d \hat{N}_d u_{1d}(\beta)A_{x,d}A'_{x,d}$, which is essentially the same as the one in (5) except for the introduction of sampling weights, that is, n_d is replaced by $\hat{N}_d$. It follows that the asymptotic distribution of $\hat{\beta}^{\mathrm{wql}}_{[p]}$ is given by

$$\hat{\beta}^{\mathrm{wql}}_{[p]} - \beta_{[p]} \sim_{\mathrm{approx}} N\left(0, \hat{\Gamma}_{\mathrm{wq}(\beta[p])}\right); \quad \hat{\Gamma}_{\mathrm{wq}(\beta[p])} = J^{-1}_{\mathrm{wq}(\beta[p])}\hat{\Sigma}_{\mathrm{wq}(\phi[p])} \times J'^{-1}_{\mathrm{wq}(\beta[p])}\Big|_{\beta=\hat{\beta}^{\mathrm{wql}}_{[p]}}. \tag{53}$$

4.1.2. Instability of the estimated covariance matrix $\hat{\Sigma}_{\mathrm{wq}(\phi[p])}$

With the above results, it would seem that all the results on ql-estimation from Section 3 should carry over to wql-estimation. However, the covariance matrix $\hat{\Sigma}_{\mathrm{wq}(\phi[p])}$ is typically unstable with a high condition number (defined by the square root of the ratio of maximum and minimum eigenvalues) due to high variability in eigenvalues, but it is the small eigenvalues that cause extreme values of the test statistic. This instability has a serious effect on the matrix inversion required for the test statistic and renders the test liberal by inflating the Type I error rate. The problem can be explained by noting that for complex surveys, the degrees of freedom for the estimator $\hat{\Sigma}_{\mathrm{wq}(\phi[p])}$ in (51) is often not large as it is equal to the total number of psus (c) minus the total number of strata (L) among those having nonempty intersection with the subpopulation or

domain corresponding to the wq-score function. In practice, typically L is large for good representation of the sample and c_h per stratum is small for sample efficiency. For testing nested hypotheses $H_2 \subset H_1$, the instability of $\hat{\Sigma}_{\mathrm{wq}(\phi[p])}$ affects in general the stability of the covariance $\Sigma_{\mathrm{wq}(\phi[p_1-p_2\,|\,p_2])}(= F_{\mathrm{wq}}\Sigma_{\mathrm{wq}(\phi[p_1])}F'_{\mathrm{wq}})$ of the wq-nscore function appearing in the wq-score statistic $Q_{\mathrm{wq}}(H_2|H_1)$, which is defined analogously to (43) as

$$
\begin{aligned}
Q_{\mathrm{wq}}(H_2|H_1) &= Q\big(\phi_{\mathrm{wq}(\beta[p_1-p_2|p_2])}\big)\Big|_{\beta_{[p_1]}=\left(\hat{\beta}'^{\mathrm{wql}}_{[p_2]},\beta'_{[p_1-p_2]}=0'\right)'};\\
\phi_{\mathrm{wq}(\beta[p_1-p_2|p_2])} &= \phi_{\mathrm{wq}(\beta[p_1-p_2])} - I_{\mathrm{wq}(\beta[p_1-p_2|p_2])}I^{-1}_{\mathrm{wq}(\beta[p_2])}\phi_{\mathrm{wq}(\beta[p_2])}\\
&\equiv F_{\mathrm{wq}}\phi_{\mathrm{wq}(\beta[p_1])}\\
\big(F_{\mathrm{wq}} &= \big(-I_{\mathrm{wq}(\beta[p_1-p_2|p_2])}I^{-1}_{\mathrm{wq}(\beta[p_2])},\ I_{(p_1-p_2)\times(p_1-p_2)}\big)\big),
\end{aligned}
\tag{54}
$$

where the expected wq-information matrix $I_{\mathrm{wq}(\beta[p])}(= E(J_{\mathrm{wq}(\beta[p])}))$ is defined below in (55). The possible instability of $\hat{\Sigma}_{\mathrm{wq}(\phi[p_1-p_2|p_2])}$ results in unstable finite sample behavior of $Q_{\mathrm{wq}}(H_2|H_1)$, such as inflated Type I error rate and reduced power suitably adjusted for size due to inflated Type I error rate (Thomas and Rao, 1987). Note that the instability in $\hat{\Sigma}_{\mathrm{wq}(\phi[p_1-p_2\,|\,p_2])}$ results in both relatively large and small eigenvalues, but the test behavior is affected more by small ones because they cause more cases of rejection of the null hypothesis than acceptance caused by large eigenvalues. To overcome this problem, a way out is to use a suitable working covariance $\Sigma^*_{\mathrm{wq}(\phi[p_1-p_2|p_2])}$ obtained by modifying the estimated covariance matrix under a simplified design such that it is stable regardless of c-L being large or not and then correct for bias in the null distribution of the corresponding test statistic $Q^*_{\mathrm{wq}}(H_2|H_1)$ as suggested by Rao and Scott (1984) and shown below.

Note that here the use of the working covariance is motivated not because of unavailability of the actual covariance matrix as in ql-estimation but because of instability of the estimated covariance matrix. However, the original motivation of Rao and Scott (1984) in using a working covariance was to be able to use χ^2-type tests for secondary analysis from published tables when the microlevel information required for computing design-based covariance matrix was not available. The qualifier "suitable" for the working covariance matrix is used in the sense that, although it is computed under a simplified design, some important design features are preserved through the use of sampling weights in parameter estimation. For example, for the wq-score function (49b), under the working assumption of stratified simple random sampling with domains as strata, $\hat{N}_d$ equals N_d, the design-based variance is $\sum_d A_{x,d}A'_{x,d}N_d^2A_{y_1,d}(1-A_{y_1,d})/n_d$ ignoring the finite-population correction $(N_d-1)^{-1}(N_d-n_d)$. The unknown $A_{y_1,d}$ can be consistently estimated by $\overline{y}_{1dw}$ or $\mu_{1d}(\hat{\beta}^{\mathrm{wql}})$ and N_d by $\hat{N}_d$ if it were unknown. For $\Sigma^*_{\mathrm{wq}(\phi[p_1])}$, we use $\mu_{1d}(\hat{\beta}^{\mathrm{wql}})$ and not $\overline{y}_{1dw}$, and replace n_d by $\tilde{n}_d = n(N_d/N)$ to satisfy a proportionality condition explained below. A useful alternative motivation for the above choice of $\Sigma^*_{\mathrm{wq}(\phi[p_1])}$ is as follows. Observe that the covariance matrix of the census EF vector is equal to the corresponding expected information matrix $\sum_d N_d u_{1d}(\beta)A_{x,d}A'_{x,d}$ because of its optimality. Therefore, the anticipated covariance matrix (with respect to the joint $\pi\xi$-randomization) of the sample EF vector under the simplified assumption of simple random sampling is given by N/n times the covariance matrix of the census EF-vector which, indeed, coincides with $\Sigma^*_{\mathrm{wq}(\phi[p_1])}$.

4.1.3. *Null distribution of $Q^*_{\mathrm{wq}}(H_2|H_1)$ as a linear combination of independent χ^2_1-variables*

For testing $H_2 \subset H_1$, consider the wq-score statistic $Q^*_{\mathrm{wq}}(H_2|H_1)$ based on the wq-nscore function $\phi_{\mathrm{wq}(\beta[p_1-p_2|p_2])}$ and its working covariance $\Sigma^*_{\mathrm{wq}(\phi[p_1-p_2|p_2])}$. Observe that if the working covariance matrix $\Sigma^*_{\mathrm{wq}(\phi[p_1])}$ for the full p_1-vector of wq-score functions $\phi_{\mathrm{wq}(\beta[p_1])}$ is chosen such that it is proportional to the expected wq-information matrix $I_{\mathrm{wq}(\beta[p_1])}$ evaluated at $\beta_{[p_1]} = \big(\hat{\beta}^{'\mathrm{wql}}_{[p_2]}, \beta'_{[p_1-p_2]} = 0\big)'$, which is given by

$$
\begin{aligned}
I_{\mathrm{wq}(\beta[p_1])} &= E(J_{\mathrm{wq}(\beta[p_1])}) \\
&= E\left(\sum_d \hat{N}_d u_{1d}(\beta) A_{x,d} A'_{x,d}\right) = \sum_d N_d u_{1d}(\beta) A_{x,d} A'_{x,d},
\end{aligned}
\tag{55}
$$

then the wq-nscore function of (54) will not change if $I_{\mathrm{wq}(\beta[p])}$ is replaced by $\Sigma^*_{\mathrm{wq}(\phi[p])}$ because $I_{\mathrm{wq}(\beta[p_1-p_2\,|\,p_2])} I^{-1}_{\mathrm{wq}(\beta\,[p_2])}$ equals $\Sigma^*_{\mathrm{wq}(\phi[p_1-p_2\,|\,p_2])} \Sigma^{*-1}_{\mathrm{wq}(\phi\,[p_2])}$. Thus, the wq-nscore function is similar to the nscore function of (14) with Σ replaced by Σ^* and therefore can also be obtained from $\phi_{\mathrm{wq}(\beta[p_1-p_2])}$ by subtracting its projection on $\phi_{\mathrm{wq}(\beta[p_2])}$ under a working covariance norm. The above proportionality condition, implicit in Rao and Scott (1984), is sufficient to express $Q^*_{\mathrm{wq}}(H_2|H_1)$ as a difference of two X^2-type statistics. It follows that the decomposition like (17) holds for $Q^*_{\mathrm{wq}}(H_2|H_1)$, and therefore, along the lines of the argument used in (20) where the stronger sufficient condition of information unbiasedness ($I_{\beta[p_1]} = \Sigma_{\phi[p_1]}$) is satisfied, we can express the test statistic as

$$
Q^*_{\mathrm{wq}}(H_2|H_1) = X^{2*}(H_2|H_{\mathrm{S}}) - X^{2*}(H_1|H_{\mathrm{S}}),
\tag{56}
$$

where $X^{2*}(H_2|H_{\mathrm{S}})$, for example, is $\sum_d \tilde{n}_d u^{-1}_{1d}(\beta)(\overline{y}_{1dw} - \mu_{1d}(\beta))^2$ evaluated at $\beta_{[p_2]} = \hat{\beta}^{\mathrm{wql}}_{[p_2]}$ and $\beta_{[p_{\mathrm{S}}-p_2]} = 0$. Note that it may be tempting to use n_d in the above X^{2*} statistic instead of $\tilde{n}_d$ but that would not satisfy the required proportionality condition.

It follows from Rao and Scott (1984) that the asymptotic null distribution of $Q^*_{\mathrm{wq}}(H_2|H_1)$ is not $\chi^2_{p_1-p_2}$ but a linear combination $\sum_{i=1}^{p_1-p_2} \delta_i \chi^2_{1i}$ of independent χ^2_1 variables. This result requires that the test statistic be based on nscore functions defined in (54) so that even after substitution of consistent estimates for the unknown nuisance parameters $\beta_{[p_2]}$, they can be treated as known asymptotically. The non-negative coefficients δ_is, known as g-deffs (generalized design effects; Skinner, 1989, p. 43), are the eigenvalues of the design effect matrix $\Delta_{\mathrm{wq}(\phi[p_1-p_2|p_2])}$ defined as the product $\Sigma^{*-1}_{\mathrm{wq}(\phi[p_1-p_2|p_2])} \hat{\Sigma}_{\mathrm{wq}(\phi[p_1-p_2|p_2])}$ associated with the wq-nscore function. The working covariance matrix $\Sigma^*_{\mathrm{wq}(\phi[p_1-p_2|p_2])}$ for the wq-nscore function can be easily computed from (16) after replacing Σ by Σ^*, while $\hat{\Sigma}_{\mathrm{wq}(\phi[p_1-p_2|p_2])}$ is computed as $F^* \hat{\Sigma}_{\mathrm{wq}(\phi[p_1])} F^{*\prime}$, where F^* is defined as F_{wq} of (54) with $I_{\mathrm{wq}(\beta[.])}$ replaced by $\Sigma^*_{\mathrm{wq}(\phi[.])}$ (F^* and F_{wq} being identical under the proportionality condition). The matrix $\Delta_{\mathrm{wq}(\phi[p_1-p_2|p_2])}$ is identical to the design effect matrix of Rao and Thomas (1989, eqn. 4.36) defined for nested hypotheses under log-linear models. Note that for testing nested hypotheses under log linear models, the expression for the design effect matrix given in Rao and Thomas (1989, eqn. 4.36) may seem somewhat different, but is indeed identical to the matrix $\Delta_{\mathrm{wq}(\phi[p_1-p_2|p_2])}$ presented here.

4.1.4. Rao–Scott first- and second-order corrections

In practical applications of the above test, it is convenient to make simple χ^2 approximations to the linear combination by correcting the test statistic for bias by suitable scaling. A first-order correction to $Q^*_{\mathrm{wq}}(H_2|H_1)$ is obtained by dividing it by $\overline{\delta}$, the average of $(p_1 - p_2)$ eigenvalues δ_is, and then treating $\overline{\delta}^{-1} Q^*_{\mathrm{wq}}(H_2|H_1)$ as a $\chi^2_{p_1-p_2}$ variable under the null hypothesis. A more accurate second-order correction to $Q^*_{\mathrm{wq}}(H_2|H_1)$ is obtained by dividing it by $\overline{\delta}(1+a^2)$, where a is the coefficient of variation of the δ_is $(\mathrm{CV}(\delta) = \left|\sum_i \delta_i\right|^{-1}\sqrt{(p_1 - p_2)\sum_i (\delta_i - \overline{\delta})^2})$ and then treating $Q^*_{\mathrm{wq}}(H_2|H_1)/\overline{\delta}(1+a^2)$ as a χ^2 variable with degrees of freedom equal to $(p_1 - p_2)(1+a^2)^{-1}$ under the null hypothesis. If the coefficient of variation of δ_is is small, then the simple first-order correction would be adequate. The second-order correction gives a better control on size of the test as it is more conservative than the first-order correction. The first-order correction is motivated from the desire to match the first moment of the reference distribution with that of the $\chi^2_{p_1-p_2}$ distribution, while the second-order correction, also known as Satterthwaite's approximation, attempts to match the first two moments. Notice that under the null hypothesis, we have

$$\begin{aligned} E\Big[Q^*_{\mathrm{wq}}(H_2|H_1)/\overline{\delta}\left(1+a^2\right)\Big] &\simeq (p_1 - p_2)/(1+a^2), \\ \mathrm{Var}\Big[Q^*_{\mathrm{wq}}(H_2|H_1)/\overline{\delta}\left(1+a^2\right)\Big] &\simeq 2(p_1 - p_2)/(1+a^2). \end{aligned} \tag{57}$$

Also note that both corrections can be obtained directly from $\Delta_{\phi[p_1-p_2|p_2]}$without having to compute the eigenvalues (Skinner, 1989, Chapter 2, p. 44) since

$$\begin{aligned} (p_1 - p_2)\,\overline{\delta} &= \mathrm{tr}\big(\Delta_{\mathrm{wq}(\phi[p_1-p_2|p_2])}\big)\left(= \sum_{i=1}^{p_1-p_2} \delta_i\right), \\ (p_1 - p_2)\overline{\delta}^2\left(1+a^2\right) &= \mathrm{tr}\Big(\Delta^2_{\mathrm{wq}(\phi[p_1-p_2|p_2])}\Big)\left(= \sum_{i=1}^{p_1-p_2} \delta_i^2\right), \end{aligned} \tag{58}$$

using the spectral decomposition of $\Delta_{\phi[p_1-p_2|p_2]}$. For illustrative applications of R-S corrections to the Canada Health Survey, see Hidiroglou and Rao (1987). In practice, for the first-order R-S correction, $\mathrm{tr}(\Delta_{\mathrm{wq}(\phi[p_1-p_2|p_2])})$ can also be more conveniently computed as $\mathrm{tr}\big(\Sigma^{*-1}_{\mathrm{wq}(\phi[p_1])}\hat{\Sigma}_{\mathrm{wq}(\phi[p_1])}\big) - \mathrm{tr}\big(\Sigma^{*-1}_{\mathrm{wq}(\phi[p_2])}\hat{\Sigma}_{\mathrm{wq}(\phi[p_2])}\big)$ (compare with eqn. 4.31 of Rao and Thomas, 1989) because $\mathrm{tr}\big(\Sigma^{*-1}_{\mathrm{wq}(\phi[p_1])}\hat{\Sigma}_{\mathrm{wq}(\phi[p_1])}\big)$ equals $\mathrm{tr}\big[\big(C^*\Sigma^*_{\mathrm{wq}(\phi[p_1])}C^{*\prime}\big)^{-1}\big(C^*\hat{\Sigma}_{\mathrm{wq}(\phi[p_1])}C^{*\prime}\big)\big]$, where C^* is a nonsingular matrix defined as $(G', F^{*\prime})'$, $G = \big(I_{p_2\times p_2}, O_{p_2\times(p_1-p_2)}\big)$, F^* already defined earlier in this section, and the fact that $\mathrm{tr}\big[\big(C^*\Sigma^*_{\mathrm{wq}(\phi[p_1])}C^{*\prime}\big)^{-1}\big(C^*\hat{\Sigma}_{\mathrm{wq}(\phi[p_1])}C^{*\prime}\big)\big]$ is sum of $\mathrm{tr}\big(\Sigma^{*-1}_{\mathrm{wq}(\phi[p_2])}\hat{\Sigma}_{\mathrm{wq}(\phi[p_2])}\big)$ and $\mathrm{tr}(\Delta_{\mathrm{wq}(\phi[p_1-p_2|p_2])})$ due to a block diagonal structure of $C^*\Sigma^*_{\mathrm{wq}(\phi[p_1])}C^{*\prime}$.

4.1.5. F-based versions

It was mentioned earlier that the behavior of the wq-score statistic is unstable in the sense of being liberal or having inflated size because the degrees of freedom (taken as $f = c - L$) for estimating the covariance $\Sigma_{\mathrm{wq}(\phi[p_1-p_2|p_2])}$ is usually not large. To improve stability, a conservative F-version of the $Q_{\mathrm{wq}}(H_2|H_1)$ statistic of (54) can be

used, which treats $[(f-\nu+1)/\nu](Q/f)$ as an F-variable with ν and $(f-\nu+1)$ degrees of freedom when Q is asymptotically $\chi^2_\nu(\nu = p_1 - p_2)$ (see Rao et al., 1998; Korn and Graubard, 1990). It is motivated from heuristic arguments based on Hotelling's $T^2(\sim F_{p,n-p})$ for testing a p-dimensional mean of a multivariate normal with a random sample of size n, where $(n-1)$ corresponds to f as the degrees of freedom in estimating the covariance matrix and p corresponds to ν. However, for R-S corrected $Q^*_{\mathrm{wq}}(H_2|H_1)$ statistics, somewhat different scaling adjustments are required for conservative F-versions (see Thomas and Rao, 1987 for a heuristic motivation). For the first-order correction, $\overline{\delta}^{-1}Q^*_{\mathrm{wq}}/\nu$ is treated as $F_{\nu,f\nu}$, while for the second-order correction, $\overline{\delta}^{-1}Q^*_{\mathrm{wq}}/\nu$ is treated as $F_{\nu(1+a^2)^{-1},f\nu(1+a^2)^{-1}}$ since $\big(\overline{\delta}(1+a^2)\big)^{-1}Q^*_{\mathrm{wq}}\big/\big(\nu(1+a^2)^{-1}\big)$ reduces to $\overline{\delta}^{-1}Q^*_{\mathrm{wq}}/\nu$.

4.1.6. Alternative tests

Here, we consider only methods due to Fay (1985) and Singh (1985, see also Kumar and Singh, 1987). Some important early studies among others are due to Fellegi (1980), Nathan and Holt (1980) and Holt et al. (1980). Fay proposed an innovative jackknifed adjustment to the usual Pearson's X^2 and likelihood ratio statistic G^2 for complex samples whenever a replication method such as the jackknife provides a consistent estimate of the covariance of the domain level estimates; see Rao and Thomas (2003) for a good summary of Fay's test. The basic idea underlying Fay's method is to develop a correction $\overline{K}$ (based on the variability of Q^*_{wq} over jackknife replicates) as an alternative to $\overline{\delta}$ so that the test statistic $\overline{K}^{-1}Q^*_{\mathrm{wq}}$ has a better control on size than the R-S first-order corrected statistic $\overline{\delta}^{-1}Q^*_{\mathrm{wq}}$ when the coefficient of variation of δ_is is not small. Thus, Fay's test statistic provides an alternative to R-S second-order correction, but its asymptotic null distribution is given by a function of weighted linear combinations of independent χ^2_1 variables, weights being functions of δ_is or g-deffs used for R-S corrections. Empirically, it was found that the distribution of the Fay's statistic can be well approximated by $\sqrt{2}\big\{\sqrt{\chi^2_\nu}-\sqrt{\nu}\big\}$ obtained under the working condition that δ_is are same where ν is the degrees of freedom under the null hypothesis.

The method proposed by Singh is based on the idea of collapsing the full p_{S}-vector of wq-score functions $\phi_{\mathrm{wq}(\beta[p_{\mathrm{S}}])}$ by means of a $T \times p_{\mathrm{S}}$ transformation matrix comprising T principal components of $\hat{\Sigma}_{\mathrm{wq}(\phi[p_{\mathrm{S}}])}$ corresponding to the T largest eigenvalues, $T < p_{\mathrm{S}}$. The T principal components are chosen such that the proportion of total variance retained after dropping small eigenvalues is at least $1-\varepsilon$ for a prespecified small positive constant ε. The wq-score statistic is constructed analogous to (54) but with the transformed vector $\phi^{(T)}_{\mathrm{wq}(\beta[p_{\mathrm{S}}])}$ of reduced dimension T. The resulting test (denoted by $Q^{(T)}$) attempts to avoid the problem of instability caused by small eigenvalues of the estimated covariance $\hat{\Sigma}_{\mathrm{wq}(\phi[p_{\mathrm{S}}])}$. In the simulation study of Thomas et al. (1996) for testing independence in two-way tables from cluster samples, Fay's test performs well with respect to size and power when the number of clusters is greater than 30 but is overly liberal for small number of clusters. The F-version of $Q^{(T)}$ (although in the study the quadratic statistic based on the Wald test for the log-linear model was used instead of the wq-score statistic) with $\varepsilon = 0.05$ performed well for small number of clusters in terms of controlling Type I error but had reduced power. However, in comparison

to Fay's Jackknifed χ^2, Singh's $Q^{(T)}$, and other related tests, the F-versions of Rao–Scott's second-order corrected χ^2 test provided a reasonable control on Type I error and adequate power over a wide range of situations.

4.1.7. Asymptotic functional linear regression approach

Some earlier attempts (see, e.g., Koch et al., 1975) on CDA from complex surveys are based on the asymptotic functional linear regression approach of Grizzle et al. (1969) for simple random samples. Here, the link function-based nonlinear transformation of the domain-level sample weighted averages $(\overline{y}_{1dw})_{1\leq d\leq D}$ is assumed to follow approximately a linear model along the lines of (25) for simple surveys, and then the standard analysis based on weighted least squares is used. This approach is generally not satisfactory because of the poor Gaussian approximation of the nonlinearly transformed domain-level data.

4.2. Model diagnostics under weighted quasi-likelihood

4.2.1. Covariance smoothing

Unlike testing in model selection where the working covariance $\Sigma^*_{\mathrm{wq}(\phi[p])}$ can be used, we do need a stable version of the correct covariance $\hat{\Sigma}_{\mathrm{wq}(\phi[p])}$ for other parts of the data analysis such as standardization in residual diagnostics, measures of model fit, and variance and interval estimation. For this purpose, by analogy with R-S corrections, we observe that if the coefficient of variation of the g-deffs is small, then the R-S first-order correction is adequate, that is, the average g-deff times the working covariance of the wq-score statistic provides a smoothed (and hence stable) version of the correct covariance. It follows that for the full vector of wq-score functions $\phi_{\mathrm{wq}(\beta[p_S])}$, a smoothed version $\overline{\Sigma}_{\mathrm{wq}(\phi[p_S])}$ of $\hat{\Sigma}_{\mathrm{wq}(\phi[p_S])}$ is given by $\overline{\lambda}\,\Sigma^*_{\mathrm{wq}(\phi[p_S])}$, where λ_is are simply the eigenvalues of $\Sigma^{*-1}_{\mathrm{wq}(\phi[p_S])}\hat{\Sigma}_{\mathrm{wq}(\phi[p_S])}$ because $\Sigma^*_{\mathrm{wq}(\phi[p_S])}$ is assumed to be stable. This smoothing would be reasonable if the coefficient of variation of the λ_is is small as in R-S first-order corrections. We remark that for smoothing, we could also use alternative choices of a stable working covariance matrix without requiring the proportionality condition. For instance, one could use the high entropy variance to approximate the variance of H-T estimators (Brewer and Donadio, 2003; Deville, 1999), which does not require second-order inclusion probabilities (π_{kl}). The term "high entropy" here refers to sampling designs for which π_{kl} is close to $\pi_k\pi_l$ for $k \neq l$, as in the case of simple random sampling without replacement and randomized systematic probability proportional to size sampling.

The idea of smoothing based on R-S first-order correction was generalized by Singh et al. (2005), see also You (2008) for a further application, using the simultaneous decomposition (Rao, 1973, p. 41),

$$\hat{\Sigma}_{\mathrm{wq}(\phi[p_S])} = \sum_{i=1}^{p_S} \tilde{\lambda}_i (\tilde{M}P_i)(\tilde{M}P_i)', \; \tilde{\Sigma}_{\mathrm{wq}(\phi[p_S])} = \sum_{i=1}^{p_S} (\tilde{M}P_i)(\tilde{M}P_i)', \tag{59}$$

where $\tilde{\Sigma}_{\mathrm{wq}(\phi[p_S])}$ is a suitable choice of a working covariance matrix not necessarily equal to $\Sigma^*_{\mathrm{wq}(\phi[p_S])}$, $\tilde{M}$ (a lower triangular matrix) is the left Cholesky root of $\tilde{\Sigma}_{\mathrm{wq}(\phi[p_S])}$, that is,

$\tilde{\Sigma}_{\mathrm{wq}(\phi[p_S])} = \tilde{M}\tilde{M}'$, $\tilde{\lambda}_i$s are eigenvalues of $\tilde{M}^{-1}\hat{\Sigma}_{\mathrm{wq}(\phi[p_S])}(\tilde{M}')^{-1}$ (same as the eigenvalues of $\tilde{\Sigma}^{-1}_{\mathrm{wq}(\phi[p_S])}\hat{\Sigma}_{\mathrm{wq}(\phi[p_S])}$), and P_is are the corresponding eigenvectors. Now, since the working covariance $\tilde{\Sigma}_{\mathrm{wq}(\phi[p_S])}$ is assumed to be stable, the instability in $\hat{\Sigma}_{\mathrm{wq}(\phi[p_S])}$ can be alleviated by reducing variability in $\tilde{\lambda}_i$s. This is done by dividing the eigenvalues into K homogeneous subgroups or classes of sizes $\{m_c : 1 \le c \le K\}$ using clustering algorithms, if necessary, and then replacing $\tilde{\lambda}_i$s by the average $\bar{\lambda}_c$ of the subgroup it belongs to obtain an improved version of $\overline{\Sigma}_{\mathrm{wq}(\phi[p_S])}$. Thus, the desired smoothed covariance is given by

$$\overline{\Sigma}_{\mathrm{wq}(\beta[p_S])} = \sum_{c=1}^{K} \bar{\lambda}_c \sum_{i=1}^{m_c} (\tilde{M}P_i)(\tilde{M}P_i)'. \tag{60}$$

In the simulation study of Singh et al. (2005) in the context of small-area estimation, it was found that the small domain or area estimates with the above smoothed error covariance of the input vector of direct domain or area-level estimates performed very well compared with the unsmoothed case even for very small sample sizes with respect to mean square error and coverage properties.

Now using the smoothed covariance $\overline{\Sigma}_{\mathrm{wq}(\phi[p_S])}$, the Cholesky residuals from $r_d = \hat{N}_d\big(\bar{y}_{1dw} - \mu_{1d}\big(\hat{\beta}^{\mathrm{wql}}_{[p_R]}\big)\big)$ under the final model H_R can be obtained similarly to Subsection 3.2. Other informal diagnostics such as Q-Q plots and detection of influential points can also be performed as before. For tests of sof and gof under formal diagnostics, we can use the wq-score tests Q^*_{wq} involving the working covariance along with R-S corrections. For R^2-type measures of model fit, we can invoke the asymptotic normality of $\hat{\beta}^{\mathrm{wql}}_{[p_S]}$ under the saturated model and then define these measures based on wq-score test statistics as in Subsection 3.2 with the covariance $\hat{\Sigma}_{\mathrm{ql}(\phi[p_S])}$ replaced by $\overline{\Sigma}_{\mathrm{wq}(\phi[p_S])}$.

4.3. Inferential testing and estimation under weighted quasi-likelihood

Similar to Subsection 2.3, all the tests of significance of single-model parameters or groups of them, and tests of significance for difference between domain means can be carried out by the wq-score test involving working covariances followed by R-S corrections. For estimation, the results here are also similar to those in Subsection 2.3. The main difference is that the point estimator $\hat{\beta}^{\mathrm{wql}}_{[p_R]}$ is now the wql-estimator, which is asymptotically normally distributed with a sandwich covariance given by (53), except for the use of the smoothed covariance $\overline{\Sigma}_{\mathrm{wq}(\phi[p_S])}$. For interval estimation via test inversion, we follow essentially the same steps as in Subsection 2.4, employing the smoothed covariance. Also for point, variance, and interval estimation of domain means and their contrasts, all the results go through with natural modifications for wql-estimation.

We end this subsection by providing an illustrative application of the wql-approach to the two traditional problems of testing a simple gof of weighted-cell proportions to a completely specified distribution and testing complete independence in a three-dimensional contingency table of weighted counts. It would be helpful to first consider score and q-score tests for simple surveys to highlight the modifications needed for complex surveys. Consider a regression model formulation in terms of β-parameters of the log-linear model (8) with u-parameters under multinomial sampling as mentioned in

Section 2. The first testing problem of interest is $H_{\mathrm{T1}} : \beta = \beta^{(0)}$ with no nuisance parameters because the distribution under test is completely specified, and there is a 1-1 correspondence between the true-cell proportions μ_ds ($D = AIJ$) and u-or $\beta_{[D]}$-parameters under the saturated model. For example, for the two-dimensional table (Bishop et al., 1975, p. 24), $u = (IJ)^{-1} l_{++}$, $u_{1(i)} = J^{-1} l_{i+} - (IJ)^{-1} l_{++}$, and so on, where $l_{ij} = \log \mu_{ij}$. Now, the score vector $\phi_{\beta[D-1]}$ is $A'_{x[D-1]} \mathrm{diag}\{m_d\} \mathrm{Cov}^{-1}(z_d)(z_d - m_d)_{1 \le d \le D-1}$, where z_d is the observed count for the dth cell, $m_d = n\mu_d(\beta)$ is the expected count, and $\mathrm{Cov}(z_d)_{1 \le d \le D-1}$ is the multinomial covariance matrix $n\big(\mathrm{diag}\{\mu_d(1-\mu_d)\}_{1 \le d \le D} - \mu\mu'\big)$ without the last row and column because the elements of the full D-vector $(z - m)$ are linearly dependent due to the constraint $\sum_{d=1}^{D}(z_d - m_d) = 0$ under multinomial sampling. Equivalently, the score vector $\phi_{\beta[D-1]}$ with the additional EFs $\sum_{d=1}^{D}(z_d - m_d)$ can be simplified after a nonsingular linear transformation to obtain the D-dimensional score vector $\phi_{\beta[D]}$ given by $A'_{x[D]}(z_d - m_d)_{1 \le d \le D}$, which happens to coincide with score vector under Poisson sampling where the total sample size n is not fixed. It follows from (13) that the score statistic $Q_{\mathrm{ml}}(H_{T1}|H_{\mathrm{S}})$ is the usual Pearson's statistic $\sum_{d=1}^{D}(m_d(\beta^{(0)}))^{-1}(z_d - m_d(\beta^{(0)}))^2$, which rejects $H_{\mathrm{T1}} : \beta = \beta^{(0)}$ by referring to the upper tail of χ^2_{D-1} distribution.

For the q-score test for simple surveys, we allow for unspecified intracluster correlations. With only first-moment assumptions and a working Poisson or multinomial covariance matrix, the q-score vector $\phi_{\mathrm{ql}(\beta[D])}$ is identical to the score vector $\phi_{\beta[D]}$, but with a different covariance matrix $\hat{\Sigma}_{\mathrm{ql}(\phi[D])}$ similar to the one given by (38). Here, $\hat{\Sigma}_{\mathrm{ql}(\phi[D])}$ is singular for fixed sample size because $1'_{D\times 1}\big(A'_{x[D]}\big)^{-1}\phi_{\mathrm{ql}[\beta[D]]} = 0$, $A_{x[D]}$ being nonsingular. So using a g-inverse of $\hat{\Sigma}_{\mathrm{ql}(\phi[D])}$, the q-score statistic $Q_{\mathrm{ql}}(H_{\mathrm{T1}}|H_{\mathrm{S}})$, as given in (43), can be defined although it is much simpler because there is no nuisance parameter adjustment needed here. To avoid singularity of $\hat{\Sigma}_{\mathrm{ql}(\phi[D])}$, we can alternatively define the q-score test with just $(D-1)$ q-score functions by arbitrarily dropping one.

For complex surveys, the census EF-vector $\phi_{\mathrm{ql}(\beta[D],U)}$ under the assumption of independent Bernoulli observations y_{dk}s (here, the cell or the domain d is defined by (a, i, j) instead of just (i, j) for the logit model), analogous to (49a), is $\sum_{d=1}^{D} A_{x,d} N(A_{y,d} - \mu_d(\beta_{[D]}))$, and the corresponding sample EF-vector $\phi_{\mathrm{wq}(\beta[D])}$, analogous to (49b), is $\sum_{d=1}^{D} A_{x,d}\hat{N}(\bar{y}_{dw} - \mu_d(\beta_{[D]}))$. The estimated covariance matrix $\hat{\Sigma}_{\mathrm{wq}(\phi[D])}$, obtained as in (51), is singular as in the case of q-score functions, and so is the working covariance matrix $A'_{x[D]}N^2 n^{-1}(\mathrm{diag}\{\mu_d(1-\mu_d)\} - \mu\mu')A_{x[D]}$ motivated from multinomial counts. However, we can use instead Poisson-motivated working covariance matrix $\Sigma^*_{\mathrm{wq}(\phi[D])}$ defined as $A'_{x[D]}N^2 n^{-1}\mathrm{Diag}\{\mu_d\}_{1 \le d \le D} A_{x[D]}$ to make it nonsingular or we can use the wq-score vector $\phi_{\mathrm{wq}(\beta[D-1])}$ after dropping one EF. Now, it follows from (56) that the wq-score test $Q^*_{\mathrm{wq}}(H_{\mathrm{T1}}|H_{\mathrm{S}})$ has a Pearson's form $X^{2*}(H_{\mathrm{T1}}|H_{\mathrm{S}})$, which has approximately a χ^2 distribution after R-S corrections. We note an important observation by Rao and Scott (1984) that under certain conditions, only a set of design effects (such as cell deffs for the problem considered here) are needed (and not the full specification of the design-based covariance matrix $\hat{\Sigma}_{\mathrm{wq}(\phi[D])}$) for computing the sum of g-deffs required for the R-S first-order correction. This is useful when analyzing published tables from survey data where the full design information is usually not available. For example, as in the study by Rao and Thomas (1989, eqn. 4.15), the

first-order correction factor $\overline{\delta}$ for $Q^*_{\mathrm{wq}}(H_{\mathrm{T1}}|H_{\mathrm{S}})$ is given by

$$\begin{aligned}(D-1)\overline{\delta} &= \mathrm{tr}\Big(\Sigma^{*-1}_{\mathrm{wq}(\phi[D])}\hat{\Sigma}_{\mathrm{wq}(\phi[D])}\Big) = \mathrm{tr}\Big(\Sigma^{*-1}_{\mathrm{wq}(\psi[D])}\hat{\Sigma}_{\mathrm{wq}(\psi[D])}\Big)\\ &= \sum_d \hat{V}_\pi\left(\hat{N}\overline{y}_{dw}\right)\left(N^2n^{-1}\mu_d\big(\beta^{(0)}\big)\right)^{-1}\\ &= \sum_{d=1}^{D}(\mathrm{deff}_d)\overline{y}_{dw}(1-\overline{y}_{dw})/\mu_d(\beta^{(0)}),\end{aligned} \tag{61}$$

where $\psi_{\mathrm{wq}[\beta[D]]} = \big(A'_{x[D]}\big)^{-1}\phi_{\mathrm{wq}[\beta[D]]} = \hat{N}(\overline{y}_{dw} - \mu_d(\beta_{[D]}))_{1\le d\le D}$, a nonsingular transformation of $\phi_{\mathrm{wq}[\beta[D]]}$ and deff_d is the design effect of $\hat{N}\overline{y}_{dw}$ defined as the dth diagonal element of $\hat{\Sigma}_{\mathrm{wq}(\psi[D])}$ divided by $N^2n^{-1}\overline{y}_{dw}(1-\overline{y}_{dw})$.

For the second problem of testing complete independence in a three-dimensional table of counts under a log-linear model, the null hypothesis using the traditional notation is defined as $H_{\mathrm{T2}} : u_{12(ai)} = u_{13(aj)} = u_{23(ij)} = u_{123(aij)} = 0$ for all (a, i, j). The ml-equations for the model parameters (or the nuisance parameters for H_{T2}) under multinomial or Poisson sampling, similar to (10), are given by $z_{a++} = \hat{m}_{a++}$, $z_{+i+} = \hat{m}_{+i+}$, and $z_{++j} = \hat{m}_{++j}$, which have a closed form solution for $\hat{m}_{aij}$ as $z_{a++}z_{+i+}z_{++j}/n^2$. The corresponding estimates of p_{T2} u-parameters ($p_{\mathrm{T2}} = 1+(A-1)+(I-1)+(J-1)$), although not needed for testing purposes, can be obtained from the estimates $\hat{m}_{aij}$. Using an equivalent regression model formulation for H_{T2}, the score functions $\phi_{\beta[p_{\mathrm{T2}}]}$ have the standard form, that is, $A'_{x[p_{\mathrm{T2}}]}(z_d - m_d)_{1\le d\le D}$, which are set to zero to obtain model parameter estimates denoted by $\hat{\beta}^{\mathrm{ml}}_{p_{\mathrm{T2}}}$. Next, analogous to (13), $Q_{\mathrm{ml}}(H_{\mathrm{T2}}|H_{\mathrm{S}})$ is given by $\sum_{d=1}^{D}\big(m_d\big(\hat{\beta}^{\mathrm{ml}}_{[p_{\mathrm{T2}}]}\big)\big)^{-1}\big(z_d - m_d\big(\hat{\beta}^{\mathrm{ml}}_{[p_{\mathrm{T2}}]}\big)\big)^2$ with its asymptotic null distribution $\chi^2_{D-p_{\mathrm{T2}}}$, $D-p_{\mathrm{T2}}$ equals $(A-1)(I-1)(J-1)$. Now, for the q-score test, the EFs $\phi_{\mathrm{ql}(\beta[p_{\mathrm{T2}}])}$ do not change under the working multinomial or Poisson assumption, but $\hat{\Sigma}_{\mathrm{ql}(\phi[p_{\mathrm{T2}}])}$ based on (38) does. We can, therefore, compute the statistic $Q_{\mathrm{ql}}(H_{\mathrm{T2}}|H_{\mathrm{S}})$ as in (43).

For the wq-score test of complete independence, the EFs $\phi_{\mathrm{wq}(\beta[p_{\mathrm{T2}}])}$ are given by $A'_{x[p_{\mathrm{T2}}]}\hat{N}(\overline{y}_{dw} - \mu_d(\beta_{[p_{\mathrm{T2}}]}))_{1\le d\le D}$ and need to be adjusted for nuisance parameters to obtain the wq-nscore function $\phi_{\mathrm{wq}(\beta[p_{\mathrm{S}}-p_{\mathrm{T2}}|p_{\mathrm{T2}}])}$ as in (54). Now to compute the test statistic $Q^*_{\mathrm{wq}}(H_{\mathrm{T2}}|H_{\mathrm{S}})$ as $X^{2*}(H_{\mathrm{T2}}|H_{\mathrm{S}})$ (analogous to (56) but the term to be subtracted becomes zero because H_1 coincides with H_{S}), we can choose the working covariance matrix $\Sigma^*_{\mathrm{wq}(\phi[p_{\mathrm{S}}])}$ as $A'_{x[p_{\mathrm{S}}]}N^2n^{-1}\mathrm{Diag}\{\mu_d\}_{1\le d\le D}A_{x[p_{\mathrm{S}}]}$, which is clearly proportional to the wq-information matrix $I_{\mathrm{wq}(\beta[p_{\mathrm{S}}])}$ given by $A'_{x[p_{\mathrm{S}}]}N\mathrm{diag}\{\mu_d\}_{1\le d\le D}A_{x[p_{\mathrm{S}}]}$. Note that $\Sigma^*_{\mathrm{wq}(\phi[p_{\mathrm{S}}])}$ for $Q^*_{\mathrm{wq}}(H_{\mathrm{T2}}|H_{\mathrm{S}})$ is identical to the one chosen for $Q^*_{\mathrm{wq}}(H_{\mathrm{T1}}|H_{\mathrm{S}})$ because the enlarged models for both problems are same and given by H_{S}. For R-S corrections, eigenvalues of $\Sigma^{*-1}_{\mathrm{wq}(\phi[p_{\mathrm{S}}-p_{\mathrm{T2}}|p_{\mathrm{T2}}])}\hat{\Sigma}_{\mathrm{wq}(\phi[p_{\mathrm{S}}-p_{\mathrm{T2}}|p_{\mathrm{T2}}])}$ are computed. For the first-order correction, it follows from Rao and Scott (1984) that $\overline{\delta}$ can be more conveniently computed as (see also Rao and Thomas, 1989, eqn. 4.28) follows:

$$\begin{aligned}(D-p_{\mathrm{T2}})\overline{\delta} &= \mathrm{tr}\Big(\Sigma^{*-1}_{\mathrm{wq}(\phi[p_{\mathrm{S}}])}\hat{\Sigma}_{\mathrm{wq}(\phi[p_{\mathrm{S}}])}\Big) - \mathrm{tr}\Big(\Sigma^{*-1}_{\mathrm{wq}(\phi[p_{\mathrm{T2}}])}\hat{\Sigma}_{\mathrm{wq}(\phi[p_{\mathrm{T2}}])}\Big)\\ &= \mathrm{tr}\Big(\Sigma^{*-1}_{\mathrm{wq}(\psi[p_{\mathrm{S}}])}\hat{\Sigma}_{\mathrm{wq}(\psi[p_{\mathrm{S}}])}\Big) - \mathrm{tr}\Big(\Sigma^{*-1}_{\mathrm{wq}(\psi[p_{\mathrm{T2}}])}\hat{\Sigma}_{\mathrm{wq}(\psi[p_{\mathrm{T2}}])}\Big)\end{aligned}$$

$$= \sum_{d=1}^{D} (\text{deff}_d)\overline{y}_{dw}(1-\overline{y}_{dw})\mu_d\left(\hat{\beta}^{\text{wql}}_{[p_{T2}]}\right)^{-1} \tag{62}$$

$$- \left[\sum_{d_1=1}^{D_1} (\text{deff}_{d_1})(1-\overline{y}_{d_1w}) + \sum_{d_2=1}^{D_2} (\text{deff}_{d_2})(1-\overline{y}_{d_2w}) + \sum_{d_3=1}^{D_3} (\text{deff}_{d_3})(1-\overline{y}_{d_3w})\right],$$

where the generic notations for cells of the three one-dimensional margins are d_1, d_2, and d_3 along with D_1, D_2, and D_3 for the corresponding total number of cells. The D (or p_S) EFs $\psi_{\text{wq}(\beta[D])}$ are as in (61) except that $\beta_{[D]}$ is now $\beta_{[p_{T2}]}$, the p_{T2} EFs $\psi_{\text{wq}(\beta[p_{T2}])}$ are a simplified version after a nonsingular transformation of the original EFs $\phi_{\text{wq}(\beta[p_{T2}])}$ (for obtaining $\hat{\beta}^{\text{ml}}_{p_{T2}}$) and are defined by differences between observed and expected counts for one-dimensional and lower order margins; that is, for $1 \le d_1 \le D_1 - 1$, $1 \le d_2 \le D_2 - 1$, $1 \le d_3 \le D_3 - 1$, $\hat{N}(\overline{y}_{d_1w} - \overline{\mu}_{d_{1w}}(\beta_{[p_{T2}]}))$, $\hat{N}(\overline{y}_{d_2w} - \overline{\mu}_{d_2w}(\beta_{[p_{T2}]}))$, $\hat{N}(\overline{y}_{d_3w} - \overline{\mu}_{d_3w}(\beta_{[p_{T2}]}))$, and $\hat{N}(\overline{y}_w - \overline{\mu}_w(\beta_{[p_{T2}]}))$, where $p_{T2} = 1 + (D_1 - 1) + (D_2 - 1) + (D_3 - 1)$; one EF is dropped for each one-dimensional margin in the interest of linear independence. Here, $\overline{y}_{d_1w}$ and $\overline{\mu}_{d_1w}$, for example, are the weighted averages of y_{dk} and μ_{dk} over all units k belonging to domains d within the cell d_1 of the first one-dimensional margin. The deff_d is as in (61), while deff_{d_1}, for example, is the design effect of $\hat{N}\overline{y}_{d_1w}$ and defined as the d_1^{st} diagonal element of $\hat{\Sigma}_{\text{wq}(\psi[D_1])}$ divided by $N^2n^{-1}\overline{y}_{d_1w}(1-\overline{y}_{d_1w})$, where $\psi_{\text{wq}(\beta[D_1])}$ is $\hat{N}(\overline{y}_{d_1w} - \overline{\mu}_{d_{1w}}(\beta_{[p_{T2}]}))_{1\le d_1\le D_1}$. Note that in (62), $(\text{deff}_{d_1})(1-\overline{y}_{d_1w})$ is identical to $(\text{deff}_{d_1})\overline{y}_{d_1w}(1-\overline{y}_{d_1w})\overline{\mu}_d\left(\hat{\beta}^{\text{wql}}_{[p_{T2}]}\right)^{-1}$ because all the one-dimensional margins are satisfied by the wql-estimator under H_{T2}.

5. Unit-level models

As mentioned in the introduction, it is better in the interest of estimation efficiency to use unit-level models if some or all the covariates are available at the unit level; that is, they may not take common values for all the units in the same domain d defined by the cross-classification. For example, for analyzing data from CCHS, the age covariate category at the unit level used for predicting the smoking prevalence may be finer than the broad five categories used for the count table or may even be treated as a continuous variable. In such cases, the data can not be summarized in terms of domain-level totals as was the case in previous sections for aggregate-level models because of certain covariates not taking common values for all the units in the domain. Here, the data analysis problem is still categorical because the response or the outcome variable of smoking status is categorical. Note that although we only considered aggregate-level models so far, we did express in Section 3 the aggregate-level model as a unit-level model to incorporate later on in Section 4 the unit level sampling design weights.

We note that the results of Section 3 on ql-estimation are sufficiently general in that they remain applicable to unit-level modeling. However, the main difference lies in the model diagnostics aspect. This occurs because the individual-level residuals are not

particularly useful for binary response variables or polytomous indicators. To overcome this problem, an important idea of grouping for unit-level models was proposed by Hosmer and Lemeshow (1980) following Truett et al. (1967) and a generalization by Horton et al. (1999). Here, the data are first divided into G groups s_g of size n_g, $1 \le g \le G$ using percentiles of the empirical distribution function of the estimated predictive mean for each unit. Next, as in Horton et al. (1999), group indicators can be used to enlarge the set of covariates, and the resulting q-score functions corresponding to new regression parameters in the enlarged model turn out to be differences between observed and estimated expected counts for each group. These group-level residuals, like Pearson residuals, are approximately normal (assuming large n_g for all g) under the model and can be checked for any departures from iid $N(0, 1)$ after suitable standardization.

For complex surveys, unit-level residuals are even more problematic due to potential selection bias. For residual plots, the presence of sampling weights with unit-level residuals can be represented as bubbles (see Korn and Graubard, 1999, p. 80) to signify that the weighted residuals are needed for consistent estimation of FPQs such as residual totals for groups (which have mean zero under the model), when the finite population is divided into Hosmer–Lemeshow-type groups (Note that the weighted unit-level residual by itself is not meaningful as an estimate of an FPQ.). This is the basic idea underlying the weighted quasi-likelihood approach for unit-level models but before that, results for likelihood-based methods and quasi-likelihood are reviewed. In the following, we restrict, for convenience, our attention to only logistic models for the unit level (which is a natural analogue of logit models for the aggregate level) although all the results after suitable modifications remain applicable to other generalized linear models such as probit and complementary log-log for the Binomial distribution, and log-linear for Poisson and Multinomial distributions (see McCullagh and Nelder, 1989, Chapter 2).

5.1. Likelihood-based methods for unit-level models

Consider the unit-level logistic model for a simple random sample. We have, for each sampled unit k in domain d,

$$\begin{aligned} &\text{Obs. Eqn.} \quad y_{1dk} = \mu_{1dk} + \varepsilon_{1dk},\ y_{1dk} \sim^{\text{iid}} \text{Ber}(\mu_{1dk}) 1 \le k \le n_d,\ 1 \le d \le D, \\ &\text{Link Eqn.} \quad \text{logit}\mu_{1dk} \equiv \log\{(1-\mu_{1dk})^{-1}\mu_{1dk}\} = x'_{dk}\gamma = \sum_{l=1}^{p} x_{l,dk}\gamma_l, \end{aligned} \tag{63}$$

where the covariates $x_{l,dk}$s are unit level for each unit k in domain d. Then, by conditioning on the domain sample sizes $\{n_d\}$, the log-likelihood (3) for the aggregate-level model is modified to the unit level as

$$\log L = \sum_{d=1}^{D}\sum_{k=1}^{n_d}\left[(x'_{dk}\gamma)y_{1dk} - \log(1 + e^{x'_{dk}\gamma})\right] + \text{const.} \tag{64}$$

The score function, denoted by $\varphi_{\gamma[p]}$ (open phi notation to contrast with closed phi used for aggregate-level models in Section 2), for the regression parameters $\gamma_{[p]}$ is

given by

$$\varphi_{\gamma[p]} \equiv \partial \log L/\partial \gamma' = \sum_d \sum_k x_{dk}(y_{1dk} - \mu_{1dk}(\gamma)). \tag{65}$$

Under usual regularity conditions, the ml-estimator $\hat{\gamma}^{\mathrm{ml}}_{[p]}$, which solves $\varphi_{\gamma[p]} = 0$ has the asymptotic normal distribution

$$\hat{\gamma}^{\mathrm{ml}}_{[p]} - \gamma_{[p]} \sim_{\mathrm{approx}} N\big(0,\ \Gamma_{\gamma[p]}\big)\big|_{\gamma=\hat{\gamma}^{\mathrm{ml}}_{[p]}}; \quad \Gamma_{\gamma[p]} = J^{-1}_{\gamma[p]} \Sigma_{\varphi[p]} J'^{-1}_{\gamma[p]}, \tag{66}$$

where $J_{\gamma[p]}$ is the observed information matrix analogous to the definition in (22) and given by $\sum_d \sum_k u_{1dk}(\gamma) x_{dk} x'_{dk}$, while $\Sigma_{\varphi[p]}$ is the covariance matrix of the score vector $\varphi_{\gamma[p]}$, which turns out to be identical to $J_{\gamma[p]}$. For our model, since $\varphi_{\gamma[p]}$ is an optimal EF in the sense defined in Section 3.1, we have $\Sigma_{\varphi[p]}$ equal to the expected information $I_{\gamma[p]}$ which, for the canonical link, is identical to $J_{\gamma[p]}$. Moreover, for computing $\hat{\gamma}^{\mathrm{ml}}_{[p]}$, an analog of the Newton–Raphson algorithm (5) for the unit-level model can be used. We now consider the four aspects of data analysis.

5.1.1. Model selection

All the results for the aggregate-level model of Subsection (2.1) carry through based on the new score function $\varphi_{\gamma[p]}$. The quadratic function of the score statistic will now be denoted as $Q(\varphi_{\gamma[p]})$. However, for testing $H_2 \subset H_1$, we cannot express the score test statistic approximately as a difference of X^2-type statistics as in the case of aggregate-level modeling (see (20)) because in the case of unit level models, the score function is not a linear function of domain level differences between observed and expected counts. Also, for the same reason, the simplified asymptotic functional regression approach of Grizzle et al. (1969) is not directly applicable to the unit-level case.

5.1.2. Model diagnostics

We first define data groups as suggested by Hosmer and Lemeshow (1980) based on ranking of the estimated proportions $\mu_{dk}\big(\hat{\gamma}^{\mathrm{ml}}_{[p]}\big)$ and then using the percentiles as group boundaries. In practice, typically, the number of groups chosen is 10 corresponding to deciles. In general, with G groups, as mentioned earlier, the group residuals can also be obtained as new EFs for the enlarged model obtained from (63) by adding indicator covariates for the new groups. The G new EFs are

$$\phi_{\gamma(g)} = n_g(\overline{y}_{1g} - \overline{\mu}_{1g}(\gamma)), \quad 1 \le g \le G, \tag{67}$$

where n_g denotes the gth group sample size, and $\overline{y}_{1g}$ and $\overline{\mu}_{1g}$ denote the group-level averages, $\overline{y}_{1g} = n_g^{-1} \sum_{d=1}^{D} \sum_{k=1}^{n_d} y_{1dk(g)} 1_{dk(g)}$, where $1_{dk(g)}$ indicates the membership of the (dk)th unit in the gth group and $\overline{\mu}_{1g}$ is similarly defined. It will be assumed, and this is generally the case, that all the new EFs except for one (typically due to the presence of an intercept term in the model (63)) are linearly independent of the original EFs $\varphi_{\gamma[p]}$ of (65). Note that the EFs (67) are obtained from EFs (65) by simply replacing the vector x_{dk} by $(1_{dk(g)})_{1\le g\le G}$. We denote the enlarged vector of EFs by $\varphi_{\gamma[p^+]}$ of dimension $p^+ = p + G - 1$, where one of the new EFs, say $\varphi_{\gamma(G)}$, is dropped. Thus,

$$\varphi_{\gamma[p^+]} = \big(\varphi'_{\gamma[p]}, \varphi'_{\gamma[p^+-p]}\big)', \varphi_{\gamma[p^+-p]} = (\varphi_{\gamma(g)})_{1\le g\le G-1}. \tag{68}$$

The enlarged vector $\varphi_{\gamma[p^+]}$ is the score vector for the following model, which is analogous to the saturated model (28) with p_S replaced by p^+ and obvious modifications for the unit-level model and is given by

$$\text{logit}\,\mu_{1dk}(\gamma_{[p^+]}) = x'_{dk[p]}\gamma_{[p]} + x'_{dk[p^+-p]}\gamma_{[p^+-p]} = x'_{dk[p^+]}\gamma_{[p^+]}, \tag{69}$$

where the new $(p^+ - p)$ $(= G - 1)$ covariates $x_{dk[p^+-p]}$ correspond to the indicators $1_{dk(g)}$, $1 \le g \le G-1$. Here we assume without loss of generality that the covariate matrix $X_{n\times p^+}$ for the model (69) is of full rank p^+. It is important to note that although the group boundaries are random, they can be treated as fixed for the asymptotic normal distribution of $n_g(\bar{y}_g - \overline{\mu}_g(\gamma))$ for large n_gs, see Moore and Spruill (1975) for the case of continuous distributions and Kulperger and Singh (1982) for a modification needed for discrete distributions.

Now, as in the Subsection 2.2, for the final selected model H_R, we first obtain using the Hosmer–Lemeshow grouping idea, the $(p^+ - p_R)$-vector of nuisance parameter adjusted residuals $\tilde{r}_1 (= \varphi_{\gamma[p^+-p_R|p_R]})$ evaluated at $\hat{\gamma}^{ml}_{[p_R]}$. This makes use of the information matrix $I_{\gamma[p^+]} (= \Sigma_{\varphi[p^+]})$. We then obtain Cholesky residuals $(\tilde{r}^*_{1g})_{1\le g\le G-1}$ using the covariance $\Sigma_{\varphi[p^+]}$, which is defined as the matrix $\Sigma_{\varphi[p]}$ of (66). Note that in this process, the first p_R EFs $\varphi_{\gamma[p_R]}$ are sacrificed for estimating $\gamma_{[p_R]}$. The Q-Q plots can also be checked with G-1 approximately normal residuals $(\tilde{r}^*_{1g})_{1\le g\le G-1}$. Regarding detecting influential points, the hat matrix from the linearized regression with the adjusted dependent variable in the case of unit level can be used as was done in (29). Also, the R^2_e measures for sof and gof of the model H_R can be computed easily using the likelihood (64).

For testing sof of H_B given H_R, we can compute the nscore test $Q^{c(\alpha)}_{ml}(H_B|H_R)$ as $Q\big(\varphi_{\gamma[p_R-p_B|p_B]}\big)$ evaluated at $\gamma_{[p_R]} = \big(\hat{\gamma}^{ml}_{[p_B]}, \gamma'_{[p_R-p_B]} = 0'\big)'$, similar to what was done in (14), but with the unit-level nscore function $\varphi_{\gamma[p_R-p_B|p_B]}$ adjusted for the nuisance parameters $\gamma_{[p_B]}$, that is, $\varphi_{\gamma[p_R-p_B]} - \Sigma_{\varphi[p_R-p_B,\,p_B]}\Sigma^{-1}_{\varphi[p_B]}\varphi_{\gamma[p_B]}$, where $\Sigma_{\varphi[p_B]}$ and $\Sigma_{\varphi[p_R-p_B,p_B]}$ are obtained by partitioning $\Sigma_{\varphi[p_R]}$ as in (15). Also, for testing gof of H_R given H_F where the full model has $p_F (= p^+_R)$ γ-parameters under the enlarged model (69), a similar nscore test $Q^{c(\alpha)}_{ml}(H_R|H_F)$ can be constructed having the asymptotic null distribution $\chi^2_{p^+_R-p_R}$ with $p^+_R - p_R = G - 1$. It may seem somewhat surprising at first that with G groups, there is no loss of degrees of freedom despite the use of estimated parameters except for one due to the obvious constraint that every individual always belongs to one of the G groups, that is, $\sum_{g=1}^{G} 1_{dk(g)} = 1$. However, there is in fact a loss in degrees of freedom (from $p_R + G - 1$ to $G - 1$) because there is extra information in the original unit-level score functions $\varphi_{\gamma[p_R]}$ that is consumed in estimating the nuisance parameters $\gamma_{[p_R]}$.

5.1.3. Hosmer-Lemeshow test for model fit

An alternative test for H_R given H_F, to be denoted by X^2_{HL}, was proposed by Hosmer and Lemeshow (1980) which, based on an empirical study, exhibits a further loss in degrees of freedom under the null hypothesis (i.e., degrees of freedom being less than $G - 1$) along with an uncertainty in the asymptotic null distribution. It turns out that X^2_{HL} is asymptotically less powerful than the above nscore test $Q^{c(\alpha)}_{ml}(H_R|H_F)$ with the null distribution having full degrees of freedom $G - 1$. We remark that although

X^2_{HL} is also based on the nscore statistic $\varphi_{\gamma[p^+_R - p_R \mid p_R]}$, it uses the incorrect covariance matrix $\Sigma_{\varphi[p^+_R - p_R]}$ instead of the correct matrix $\Sigma_{\varphi[p^+_R - p_R \mid p_R]}$, all evaluated at $\hat{\gamma}^{ml}_{[p_R]}$. Now, since the distribution of counts $\{n_g\overline{y}_{1g}, n_g\overline{y}_{2g}\}_{1\le g\le G}$ under simple designs is a product of binomial distributions conditional on n_gs (sample sizes in the groups) for a total sample of size n, the X^2_{HL} statistic resembles the Pearson's X^2 because $\varphi_{\gamma[p^+_R - p_R \mid p_R]}$ at $\hat{\gamma}^{ml}_{[p_R]}$ is simply of the form $(O - E)$ corresponding to $\varphi_{\gamma(g)}$ of (67), as the second term in $\varphi_{\gamma[p^+_R - p_R \mid p_R]}$, reflecting adjustments for nuisance parameters $\gamma_{[p_R]}$, becomes 0 at $\hat{\gamma}^{ml}_{[p_R]}$, and $\Sigma_{\varphi[p^+_R - p_R]}$ conditional on the n_gs is basically a covariance matrix of independent binomial variables as in the X^2 test of (13). The X^2_{HL} test statistic is given by

$$X^2_{HL}(H_R|H_F) = \sum_{g=1}^{G} \frac{(n_g(\overline{y}_{1g} - \overline{\mu}_{1g}(\gamma))^2)}{n_g\overline{\mu}_{1g}(\gamma)(1 - \overline{\mu}_{1g}(\gamma))}\Big|_{\hat{\gamma}^{ml}_{[p_R]}}. \tag{70}$$

Note that in the quasi-likelihood framework, use of $\Sigma_{\varphi[p^+_R - p_R]}$ instead of the actual $\Sigma_{\varphi[p^+_R - p_R \mid p_R]}$ can be viewed as employing a working covariance, and by analogy with results leading to Rao–Scott Corrections in Subsection 4.1, the asymptotic null distribution of X^2_{HL} is in fact a linear combination $\sum_{i=1}^{G-1} \zeta_i X^2_{1,i}$ of independent χ^2_1 variables with coefficients ζ_is between 0 and 1. The coefficients ζ_is are easily seen to be the eigenvalues of the matrix $\Sigma^{-1}_{\varphi[p^+_R - p_R]}\Sigma_{\varphi[p^+_R - p_R \mid p_R]}$, which are between 0 and 1 because $\Sigma_{\varphi[p^+_R - p_R]} - \Sigma_{\varphi[p^+_R - p_R \mid p_R]}$ is positive definite. Thus, the null distribution of $\overline{\zeta}^{-1}X^2_{HL}$ can be treated as approximately χ^2_{G-1} by analogy with the R-S first-order correction, where $\overline{\zeta}$ is the average of ζ_is. Note also that the X^2-type form of X^2_{HL} is inherently different from the χ^2 statistics of (18) because of the lack of a meaningful definition of the saturated model under unit-level modeling.

To further understand the above behavior of the Hosmer–Lemeshow statistic, it is helpful to relate it to the earlier result of Chernoff and Lehmann (1954) in the context of the traditional Pearson's X^2 test of gof to a given distribution after transforming to the multinomial distribution by grouping the unit-level data. They showed that with $K + 1$ groups and p unknown model parameters such that $K + 1 > p$, the asymptotic null distribution is no longer χ^2_{K-p} if the more efficient ungrouped data ml-estimators of the model parameters are used instead of the grouped data ml-estimators. More specifically, they showed that in this case, the Pearson's X^2 statistic behaves like $\chi^2_{K-p} + \sum_{i=1}^{p} \omega_i \chi^2_{1,i}$ with independent χ^2_1 variables and weights ω_i satisfying $0 \le \omega_i \le 1$. Equivalently, the asymptotic null distribution can be expressed as that of $\sum_{i=1}^{K} \delta_i \chi^2_{1,i}$, where δ_is are suitable eigenvalues similar to those in the problem requiring R-S corrections. Later, Rao and Robson (1974) showed that the partial loss degrees of freedom (due to ω_is being between 0 and 1) in the Pearson's X^2 can be recovered by adding a suitable quadratic term to it; see Singh (1987) for review and optimality of the Rao–Robson statistic as well as a generalization to the case when ml-estimators are not easily available. These results on traditional gof tests can be easily reconciled with the nscore statistic $Q^{c(\alpha)}_{ml}(H_R|H_F)$ of this section, which uses the nuisance parameter-adjusted covariance and thus exhibits no loss of degrees of freedom.

Finally, for inferential testing and estimation, all the results of Subsections 2.3 and 2.4 carry over to the case of unit-level models.

5.2. *Quasi-likelihood methods for unit-level models*

Here, the model (63) with the unit-level covariates is generalized to allow for unspecified intracluster correlation, similar to (36) for the aggregate-level covariates. As in Section 3, neither the likelihood nor even the covariance structure of the unit-level data due to possible intracluster correlation is fully specified (see, e.g., McCullagh and Nelder, 1989, Chapter 9). Under the working independence assumption as in Subsection 3.1, the q-score functions $\varphi_{\mathrm{ql}(\gamma[p])}$ are given by (65) and all the previous results with natural adjustments for the q-score (such as the use of the information matrix and not the covariance matrix in (43) for nuisance parameter adjustment) carry over. In particular, for *model selection*, the q-score test $Q_{\mathrm{ql}}(H_2|H_1)$, analogous to (43), can be constructed. For *model diagnostics*, data are grouped as in the previous section to construct Cholesky residuals, Q-Q plots, gof tests, and R_e^2 measures of model fit. Notice that in the absence of the likelihood of the original data, we can use (as in Subsection 3.2) the approximate Gaussian likelihood of the p_{R}^+-vector of quasi-sufficient statistics $\hat{\gamma}^{\mathrm{ql}}_{[p_{\mathrm{R}}^+]}$ under H_{S} with $p_{\mathrm{S}} = p_{\mathrm{R}}^+$ to construct the R_e^2 measures. For model fit, q-nscore tests can be constructed. Note that if one were to use $\mathrm{X}^2_{\mathrm{HL}}(H_{\mathrm{R}}|H_{\mathrm{F}})$ of (70) for testing gof, coefficients in the linear combination of χ^2_1-variables for the asymptotic null distribution are the eigenvalues of $\Sigma^{-1}_{\mathrm{ql}(\varphi[p_{\mathrm{R}}^+ - p_{\mathrm{R}}])}\Sigma_{\mathrm{ql}(\varphi[p_{\mathrm{R}}^+ - p_{\mathrm{R}}\,|\,p_{\mathrm{R}}])}$, which, unlike in Section 5.1, need not be between 0 and 1 unless $\Sigma_{\mathrm{ql}(\varphi[p_{\mathrm{R}}^+ - p_{\mathrm{R}}])} - \Sigma_{\mathrm{ql}(\varphi[p_{\mathrm{R}}^+ - p_{\mathrm{R}}\,|\,p_{\mathrm{R}}])}$ is positive definite; here, a consistent estimate of the covariance matrix $\Sigma_{\mathrm{ql}(\varphi[p_{\mathrm{R}}^+])}$, similar to (38), is used for computing purposes. The test statistic $\mathrm{X}^2_{\mathrm{HL}}(H_{\mathrm{R}}|H_{\mathrm{F}})$ is a special case of the Horton et al. (1999) statistic for longitudinal data.

5.3. *Weighted quasi-likelihood methods for unit-level models*

Here, the model is similar to (48) except that the aggregate-level covariates are replaced by unit-level covariates and the regression parameters β by γ. Now as in Section 4, the wq-score function $\varphi_{\mathrm{wq}(\gamma[p])}$ and the corresponding asymptotically normal estimators $\hat{\gamma}^{\mathrm{wql}}_{[p_{\mathrm{R}}^+]}$ with a sandwich form of the covariance are obtained as in (53). The methods for *model selection* can be applied as before except for the natural modification to R-S corrections to suit unit-level models because of the absence of Pearson's X^2-type statistics (see Rao and Thomas, 2003). Note that here, as in Section 4, the unweighted q-score function $\varphi_{\mathrm{ql}(\gamma[p])}$ is replaced by the weighted q-score function $\varphi_{\mathrm{wq}(\gamma[p])}(= \hat{N}_g(\overline{y}_{1gw} - \overline{\mu}_{1gw}))$. For *model diagnostics*, we create Hosmer–Lemeshow groups as before but with quantiles based on the weighted empirical distribution after ordering the data by means of the predictive means (see Roberts et al., 2008) and Cholesky residuals using the smoothed covariance as in Subsection 4.2. The R_e^2 measures likewise use the smoothed version of the covariance of $\hat{\gamma}^{\mathrm{wql}}_{[p_{\mathrm{R}}^+]}$. For the gof test of H_{R} against the enlarged model H_{F} (where $p_{\mathrm{F}} = p_{\mathrm{R}}^+$ as in (69)), in particular, the q-score test statistic with the working covariance matrix under a simple design followed by R-S corrections would be preferable to the q-score test statistic with the actual (but possibly unstable) estimated covariance used by Graubard et al. (1997). Finally, all the other methods under inferential testing and estimation continue to hold with appropriate modifications to suit sampling weights and unit–level modeling as encountered earlier.

6. Conclusions

The emphasis in this chapter is on the large sample quasi-likelihood (under only first moment assumption) approach to CDA as it is amenable to data from complex surveys. There are limitations with what is feasible with complex survey data, mainly due to unavailability of second-order inclusion probabilities in general, which are required for variance–covariance estimation. Therefore, we emphasized employing only correct specification of first-order moment conditions and considered the use of a working covariance matrix to define EFs, along with a nonparametric robust estimate of the covariance matrix of them.

We present a unified and practical framework that covers past and recent solutions to different problems arising from aggregate and unit-level models for simple and complex sampling designs. For each case, four main aspects of data analysis are discussed, namely, model selection, model diagnostics, inferential testing, and inferential estimation. We show for each aspect how the standard methods can be modified to suit survey data, using tools based on quasi-likelihood estimation and quasi-score tests. Potential applications of logit (logistic in the case of unit level) models to the binary data from the CCHS are used to motivate various analysis methods.

In the case of unit-level models, the commonly used technique of grouping data suggested by Hosmer–Lemeshow is considered to define gof tests based on discrepancies between observed and expected counts and also for residual diagnostics. The idea of grouping for defining residuals not only fits in well with categorical data but is also natural for survey data with sampling weights because weighted unit-level residuals are not really meaningful as estimates of residual totals for subpopulations in the design-based context. Since model diagnostics are an important part of any data analysis, we propose the use of Cholesky residuals for obtaining uncorrelated residuals for diagnostics (which are especially important for categorical data because of non-negligible correlation among group data residuals induced by estimated model parameters), the use of generalized design effects to smooth the covariance matrix under complex designs for standardizing residuals, and the use of a generalized R-square measure developed by Estrella (1998) for model fit.

For point estimation, robust variance estimates of quasi-scores are proposed to counter possible misspecification of second moments under the model. The method of test inversion based on q-nscore (Neyman or nuisance parameter adjusted quasi-score) test statistics is considered for interval estimation as an alternative to the commonly used Wald statistics for improved stability.

In the process of reviewing existing methods, some suggestions toward their enhancement are also made. In particular, it is observed that the asymptotic null distribution of the well-known Hosmer–Lemeshow statistic for gof tests with unit-level models for binary data is in fact a linear combination of independent χ^2_1-variables with coefficients between 0 and 1. It is also observed that the partial loss of degrees of freedom in the above test due to coefficients being less than 1 can be recovered by using appropriate quasi-score tests but then it no longer has the form of Pearson's X^2.

There are important topics that are not a traditional part of CDA that we could not cover in the interest of limiting the scope of this review. For instance, mixed categorical models arise in situations when there is a large number of domains with domain-specific factor effects being modeled by random effects, and the interest lies in marginal

means and variance components (see, e.g., Jiang and Zhang, 2001; Singh and Wu, 1998; Sutradhar and Rao, 2003). Models for longitudinal categorical data with time-varying covariates but time-independent parameters come up with repeated observations or surveys over time (see, e.g., Horton et al., 1999; Liang and Zeger, 1986; Roberts et al., 2008; Singh and Sutradhar, 1989, Nathan in Chapter 34) of this handbook. Finally, we also mention models for combined longitudinal and cross-sectional categorical data with time-dependent parameters (see, e.g., Singh and Roberts, 1992 and the book by Fahrmeier and Tutz, 2001).

We conclude this by briefly discussing the special topics of sparse tables, gross flows, missing data, and ordered categorical data, which are a traditional part of CDA but are not considered in this chapter. The problem of sparse tables arises if cell counts are small. For example, in CCHS, for the outcome variable of regular drinking, the observed counts for the younger age groups such as 12–19 is very small as expected. Some cell counts may also be zero, which may affect the existence of ml-estimators; see, Fienberg (1980, Appendix IV) and Haberman (1974) for conditions for the existence of ml-estimators. It turns out that the parameter estimation may not be affected as much by the presence of empty cells or small counts as the sampling distribution of gof test statistic, which may be more seriously affected. The empirical results of Koehler and Larnz (1986) show that for gof tests, the Pearson's X^2 statistic tends to perform better than the likelihood ratio statistic G^2, which turns out to be too liberal for the 5% level, that is, rejects the null hypothesis more often. They also considered an interesting normal approximation (instead of chi-square) to the distribution of the G^2 test statistic when the number of cells increases at the same rate as the number of observations.

The problem of estimation of gross flows arises in longitudinal surveys when one is interested in estimating the transition rates between categories from one time point to the other in the presence of classification errors. Often the transition rates of interest are very small, for example, the proportion of individuals in a domain changing status from regular smoker to nonsmoker in consecutive cycles of CCHS or the proportion of individuals in a domain changing employment status from one month to the next in the monthly Canadian Labour Force Survey. Even with a very small probability of misclassification in response and a negligible bias in the estimated margins (i.e., the separate state proportions), there could be considerable bias in the estimated gross flows. Using validation data such as reinterview data, several authors proposed adjustments to estimated gross flows under fairly weak set of assumptions except for the assumption of independent classification errors from one time to other, (see, e.g., the papers by Abowd and Zellner, 1985; Poterba and Summers, 1986; Chua and Fuller, 1987; Singh and Rao, 1995). When validation data are not available, an interesting alternative method was suggested by Pfeffermann et al. (1998) who used auxiliary variables at the unit level to model true state transition probabilities and classification error probabilities relating the observed state with the true state, thus avoiding the assumption of independent classification errors.

It is assumed implicitly throughout this chapter that the sample data is complete. However, in practice, this is seldom the case. In the case of unit nonresponse, the sampling weights can be adjusted under a nonresponse model and then the covariance matrix of the EF-vector can be adjusted for nonresponse, similar to the problem of estimating population totals adjusted for nonresponse (see Chapter 8 of this handbook). For item

nonresponse, on the other hand, the data could be completed using an appropriate imputation method (see Chapter 10), but then the problem of adjusting the covariance matrix of the EF-vector based on imputed values is more difficult, see, for example, the multiple imputation methodology of Rubin (1996), some alternative methods more suitable for survey data considered by Rao and Fay in their discussions, and a more recent method using fractional imputation due to Kim and Fuller (2004); see also the discussion in Chapter 10 of this handbook. A general alternative without making any adjustments for nonresponse is to perform the analysis in the presence of missing data under a suitable response model (see the book of Little and Rubin, 2002), although with large-scale multipurpose survey data, it is preferable in practice, for the sake of convenience, to be able to use analysis methods for complete data after being adjusted for unit and item nonresponse.

The last special topic we mention is that of ordinal categorical data commonly arising in sociological studies, where the values of the outcome variable are ordered categories. Here, cumulative link models (such as cumulative logit) are often considered (see, e.g., Agresti, 2002). If some of the covariates are also ordinal, then by assigning suitable scores to covariate categories (actually what is needed essentially is the distance between ordered categories), more degrees of freedom in parameter estimation could be secured by taking advantage of the parsimony of parameters due to the extra information provided by the category scores (see, Fienberg, 1980, Chapter 4).

Acknowledgments

The author would like to thank Jon Rao for helpful discussions, Danny Pfeffermann for his detailed review with many constructive suggestions, and the referees for their useful comments. This work was supported, in part, by a grant from the Natural Sciences and Engineering Research Council of Canada held at Carleton University, Ottawa under an adjunct research professorship.

Sample Surveys: Inference and Analysis, Vol. 29B
ISSN: 0169-7161
© 2009 Published by Elsevier B.V.
DOI: 10.1016/S0169-7161(09)00236-3

Inference on Distribution Functions and Quantiles

Alan H. Dorfman

1. Introduction

The focus of this chapter is on the estimation of the finite population distribution function, on the basis of a sample taken. The subject is important: the distribution function is a basic statistic underlying many others (Serfling, 1980, Section 2.1); for purposes of assessing and comparing finite populations it can be more revealing than means and totals (Sedransk and Sedransk, 1979). Intimately tied to distribution functions are quantiles, and we shall be concerned with their estimation as well, which are typically estimated by the inversion of estimates of the distribution function. *Note*: the distribution function is also known as the cumulative distribution or the cumulative distribution function. For shorthand purposes, we shall adopt *cdf*, lest anyone arriving in the middle of the chapter think we are talking about degrees of freedom.

In some important respects, research on the estimation of finite population cdfs from survey samples is relatively young, and there is a good deal which is uncertain. We will suggest some avenues for further investigation explicitly in the last section and implicitly in the course of the chapter with the phrase "it is not clear that…", "it would be of interest to see…", or similar expressions.

1.1. Definitions

The distribution function, also known as the empirical distribution function, of a variable y for a finite population U having N units is defined by

$$F_N(t) = N^{-1} \sum_{i=1}^{N} I\,(y_i \leq t)\,, \quad -\infty < t < \infty,$$

where $I(u)$ is the indicator function (or "truth function") defined by $I(u) = 1$, if u is true, $I(u) = 0$, if not. In words, for a given value of t, $F_N(t)$ is the proportion of y in the population not exceeding t. An alternate symbolism sometimes employed in the cdf literature is $F_N(t) = N^{-1} \sum_{i=1}^{N} H(t - y_i)$, where $H(u)$ is the Heaviside function, given by $H(u) = 1, u \geq 0$ and $H(u) = 0, u < 0$, and sometimes represented using Δ instead of H. In this chapter we will stick to the indicator function notation. It will occasionally be convenient to use a shorthand $z_i = I(y_i \leq t)$.

From the definition immediately follows the basic properties: (a) $F_N(t)$ is monotonic nondecreasing in t, (b) $F_N(t)$ is a step function with step size N^{-1}, and (c) $0 \leq F_N(t) \leq 1$.

If the y_i's are regarded as realizations of independent random variables Y_i each generated from a probability distribution having distribution function $F(t)$, then $F_N(t)$ is an unbiased and strongly consistent estimator for $F(t)$ (Serfling, 1980, pp. 56–57). For $0 \leq p \leq 1$, the pth quantile is defined by $Q(p) = \min\{t : F(t) \geq p\}$. In the case where F is monotonic strictly increasing, $Q(p) = F^{-1}(p)$, in the ordinary sense of function inversion. We shall use this notation even when we do not have strict monotonicity, when we wish to identify the distribution function from which the quantile is derived.

Estimation of the distribution function leads to estimates of quantiles, $Q(p)$ by $Q_N(p) = \min\{t : F_N(t) \geq p\}$. In this chapter, we shall be concerned with the transition from sample to finite population, and ignore the further question of estimating the underlying generating distribution function. It can be safely assumed, however, that a good job of estimating $F_N(t)$ typically implies a good job of estimating $F(t)$, and vice versa. From this point on, for notational convenience, we will write $F(t) \equiv F_N(t)$, dropping the subscript N, with, we hope, no confusion.

Estimating the finite population distribution function $F(t)$ is in some respects easier and in others more difficult than estimating a population total or mean. On the one hand, for fixed t, $F(t)$ is simply a mean of 0's and 1's. As such it should show no more complication than estimating a mean, and sometimes can be eased using a transformed version of y (see Remark 1.1). On the other hand, (a) we typically want to estimate $F(t)$ for more than one value of t and these estimates need to be coordinated, especially if we have a further interest in estimating quantiles, and (b) where y is related to an auxiliary variable x, it becomes a question how to use this information, since now we are concerned with $z \equiv I(y \leq t)$ and not y itself, which is what usually gets modeled on x.

REMARK 1.1. *If $u = g(y)$ is a strictly monotone increasing function of y, then one readily sees that estimating the cdf for u is equivalent to estimating the cdf for y: $F_Y(t) = P(Y \leq t) = P(g(Y) \leq g(t)) = F_U(g(t))$. In some cases, it may help to use the relation to auxiliary information x of a transformed version of y rather than y itself. Compare for example Section 3.4.*

1.2. The design-based and model-based perspectives

In this chapter, we will be referring to "design-based" and "model-based" estimators, alluding to two different perspectives on sampling inference. Most basic sampling texts adhere to the design-based (or "probability sampling") approach, for example, Cochran (1977), Särndal et al. (1992). Valliant et al. (2000) describe the model-based (or "prediction") approach. See also especially Chapters 1 and 23 of this volume.

In brief, the design-based approach bases inference on the probability distribution generated by the activity of the survey sampler; the model-based approach, on a hypothesized relation of the data in the sample to that in the population.

The design-based approach treats the finite population $U = \{(x_i, y_i), i = 1, 2, \ldots, N\}$ as a fixed large set of parameters, where y is the variable of interest and x is the auxiliary variable, embodying extra information possibly relevant to y. The

random variable on which inference rests is the indicator variable I_i, $i \in U$, which is 1 if unit i is chosen for the sample s, and 0 if i is in the "remainder" $r = U - s$. The "inclusion probability" $\pi_i = \text{prob}\,(I_i = 1)$ plays a fundamental role in design-based estimators, with the inverse of π_i the default "weight" placed on the ith sample point. Likewise the "second-order inclusion probabilities" $\pi_{ij} = \text{prob}\left(I_i = 1 \;\&\; I_j = 1\right)$ play a role in variance calculations and in some estimators of the distribution function (see Section 3.2).

The model-based approach regards the population values of y as realizations of random variables that are generated according to some probability model - the "working model." Estimators are grounded in the model, and the π_i (and even the random selection of s) are not typically regarded as necessary ingredients for inference. At the same time, efforts are made to protect inference against model misspecification (see Chapter 23). This turns out to be intrinsically more difficult in the case of estimating distribution functions than for totals and means (see Section 3.1).

Thus there is more emphasis in the case of distribution functions on model diagnostics (Section 3.5), on estimators that seek to automatically adjust for model failure (Section 3.6.2), and on weakening the model (Section 3.7).

Model-assisted estimation is a form of design-based estimation that incorporates both the inclusion probabilities π_i and a model, attempting "to provide valid conditional inferences under the assumed model and [to protect] against model misspecification in the sense of providing valid design-based inferences irrespective of the population y-values" (Rao, 1994). There is often a strong resemblance of these estimators to the class of model-based estimators that seek to automatically adjust for model failure (Section 3.6.2).

Most simulation studies, even those studying the behavior of model-based estimators, are essentially constructed from a design-based perspective, looking at biases and mean square error over repeated sampling from a single population. They may therefore be regarded as slightly favoring design-based estimators.

1.3. Desirable properties of estimators of the distribution function

Many authors, notably Nascimento Silva and Skinner (1995), have advocated that a sample based cdf estimator $\hat{F}(t)$ should possess some or all of the following properties:

(1) $\hat{F}(t)$ is itself a distribution function, satisfying

   (a) the *boundary condition*, $0 \le \hat{F}(t) \le 1$, and
   (b) the *monotonicity* condition $t < t' \Rightarrow \hat{F}(t) \le \hat{F}(t')$.

   Of these, the boundary condition (a) is the more fundamental, because $\hat{F}(t)$ outside these bounds are necessarily off target. To achieve a quantile estimate $\hat{Q}(p)$ by inverting $\hat{F}(t)$ requires (b).

(2) $\hat{F}(t)$ is simple,

   (a) in form, for example, have a single set of weights w_i applied to $z_i \equiv I\,(y_i \le t)$ for all t,
   (b) to calculate, and
   (c) to understand.

(3) $\hat{F}(t)$ is readily invertible to get quantiles. This is perhaps a particular aspect of 2(a).

(4) $\hat{F}(t)$ is automatically constructed in all its details, that is, it does not require choices (of for example, a particular model or bandwidths) or diagnostic acuity on the part of the analyst.

(5) $\hat{F}(t)$ is calibrated with respect to any auxiliary variables x that are correlated with y. Typically, this means that if y is replaced by x, then $\hat{F}(t) = F(t)$.

(6) $\hat{F}(t)$ is efficient: its mean square error with respect to $F(t)$ is smaller than competing estimators – the mean square error is usually calculated across samples from the same population, that is, with respect to the sample design. It makes full use of available relevant auxiliary information.

(7) $\hat{F}(t)$ is unbiased or nearly so: usually taken with respect to the sample design.

(8) $\hat{F}(t)$ is consistent: $\hat{F}(t)$ tends to be closer and closer to the target $F(t)$, as the sample size n gets larger and larger.

(9) $\hat{F}(t)$ is robust: it stands up well against competing estimators under a variety of conditions. If not invariably the most efficient, it is not much less efficient than the best estimator for given circumstances.

(10) $\hat{F}(t)$ has a readily formulated variance, and a readily calculated and simple variance estimator.

These properties are not entirely compatible. For example, aiming at simplicity we may pay a price in efficiency, and vice versa. They are also not equally important; for example, it may suffice that an estimator be efficient and robust, without requiring unbiasedness as well. Also, standard algorithms of isotonic regression, for example, the pool-adjacent-violaters algorithm, can be used to make minor adjustments to a potentially wayward $\hat{F}(t)$ to guarantee Property 1, albeit at some sacrifice with respect to Property 2.

Which properties to most emphasize will depend on circumstances. Large repetitive government surveys will value property 4. A one time survey, with a statistically savvy analyst who has time to explore the data, will perhaps put efficiency and robustness considerations uppermost.

Different priorities (cdf or quantiles), different emphases on the aforementioned properties, the different forms auxiliary information can take, as well as the tension between model-based and design-based sampling ideas, perhaps account for the great variety of cdf estimators that have been put forth in the recent years.

2. Estimating the distribution function with no auxiliary information

We assume a sample s of size n, perhaps selected by probability sampling with inclusion probabilities π_i. Although the inclusion probabilities inevitably rest on some auxiliary information, that information may not be available to the data analyst. In this section, we consider estimation lacking any data on an auxiliary variable x. All we have is the sample y and possibly also the sample design weights $d_i \equiv \pi_i^{-1}$. For the moment we do not worry about second-order inclusion probabilities, which are rarely available, except in simple random or stratified sampling.

2.1. The Hájek estimator

A convenient estimator has the form

$$\hat{F}_w(t) = \sum_{i \in s} w_i I\,(y_i \le t), \tag{1}$$

where the weights w_i satisfy $0 \le w_i \le 1$ and $\sum_{i \in s} w_i = 1$. This estimator in general satisfies Properties 1, 2, and 3 mentioned earlier. Assume the y_i are listed in ascending order. Then the corresponding quantile estimator, readily calculated, is

$$\hat{Q}(p) = \min \left\{ y_i \Big| \sum_{i'=1}^{i} w_{i'} \ge p \right\}.$$

If $w_i = d_i \Big/ \sum_{i' \in s} d_{i'}$, we get the Hájek estimator, the "customary design based estimator"

$$\hat{F}_\pi(t) = \sum_{i \in s} \pi_i^{-1} I\,(y_i \le t) \Big/ \sum_{i \in s} \pi_i^{-1} = \sum_{i \in s} d_i z_i \Big/ \sum_{i \in s} d_i \tag{2}$$

which is design-consistent and approximately design-unbiased. This tends to be the estimator against which other estimators of $F(t)$ are (usually successfully) compared.

In the case of simple random sampling, $\hat{F}_\pi(t)$ reduces to the sample empirical distribution function or "naïve estimator"

$$\hat{F}_n\,(t) = n^{-1} \sum_{i \in s} I\,(y_i \le t) = n^{-1} \sum_{i \in s} z_i. \tag{3}$$

When there are no preconditions on the values of the population y_i, then $\hat{F}_n$ is admissible with respect to four standard loss functions for all sample designs (Cohen and Kuo, 1985a) and is minimax under simple random sampling for one of these loss functions (Cohen and Kuo, 1985b).

Another special case of $\hat{F}_\pi(t)$ is the stratification-based estimator studied by Sedransk and Sedransk (1979).

Design-based variances and variance estimators for $\hat{F}_\pi(t)$ use standard design-based formulas with z_i the variable of interest.

2.2. Alternatives to Hájek

Kuk (1988) compared three estimators, the Hájek as mentioned above, the Horvitz–Thompson estimator $\hat{F}_{\mathrm{HT}}(t) = N^{-1} \sum_{i \in s} d_i I\,(y_i \le t)$, and the "the complementary proportion" $\hat{F}_R(t) = 1 - N^{-1} \sum_{i \in s} d_i I\,(y_i > t)$. Neither $\hat{F}_{\mathrm{HT}}(t)$ nor $\hat{F}_R(t)$ are cdf. He gives theoretical reasons for preferring $\hat{F}_R(t)$ or $\hat{F}_\pi(t)$ to $\hat{F}_{\mathrm{HT}}(t)$. In an empirical study on several populations, $\hat{F}_R(t)$ showed best as basis for construction of medians. He also considered a fourth estimator $\hat{F}_{\hat{\lambda}}(t) = \hat{\lambda} \hat{F}_{\mathrm{HT}}(t) + (1 - \hat{\lambda}) \hat{F}_{\mathrm{R}}(t)$, a weighted average aimed at giving minimal mean square error. The weights are estimates based on the population of x values. Thus properly speaking, this estimator belongs to Section 3 below. In simulations, it is often best, but, surprisingly, sometimes not as good as $\hat{F}_{\mathrm{R}}(t)$, which does not require the auxiliary information.

For the sake of quantile estimation, Hyndman and Fan (1996) (henceforth HF) consider linear interpolated alternatives to the naïve estimator $\hat{F}_n(t)$ that are of the form

$\hat{F}_{\alpha,\beta}(t) = \lambda \tilde{F}(y_k) + (1-\lambda)\tilde{F}(y_{k+1})$, for $y_k \le t \le y_{k+1}$, where $\tilde{F}(y_k) = (k-\alpha)/(n-\alpha-\beta+1)$, and $\lambda = (y_{k+1}-t)/(y_{k+1}-y_k)$, the y_i being taken in an increasing order. This estimator is continuous over the range of sample y values, and gives a readily calculable quantile estimator $\hat{Q}_{\alpha,\beta}(p) = \hat{F}^{-1}_{\alpha,\beta}(p)$ for $\tilde{F}(y_1) \le p \le \tilde{F}(y_n)$.

There is some controversy as to the best values of α and β. The pair $\alpha = 0$, $\beta = 1$ gives the standard naive estimator F_n above, interpolated. Other values yield estimators with the property that $\hat{F}_{\alpha,\beta}(y_i) < 1$ for all sample y_i, which may be desirable, because it is unlikely that in srs the largest point in the sample is the largest possible. HF suggest a list of desirable properties of quantiles and indicate which different values of α and β satisfy them. We shall not go into detail here, except to note that HFs Property P2, p.361, should be corrected to say that the counts of occurrences of the variable that are less than or equal to $\hat{Q}(p)$ is no less than the floor of pn– the "floor" of a quantity being the largest integer less than a quantity. This revised criterion is met if $\alpha \ge 0$ and $\beta \le 1$, which replaces a statement in HF p.363 (Rob Hyndman, personal communication.) It is an interesting question as to what the appropriate generalization of HF is in the case where we seek an alternative to $\hat{F}_w$ in general. Shah and Vaish (2006) give one approach to this question.

Shuster (1973) and Modarres (2002) consider the case where the targeted cdf has a known point of symmetry, so that $F(\theta+t) + F(\theta-t) = 1$, for all t. Transforming so the point of symmetry is at zero, an estimator that puts weight $1/2n$ at $\pm y_i$ is a good deal more efficient than the naïve estimator F_n.

3. Estimating the distribution function with complete auxiliary information

The groundbreaking paper for using population information on an auxiliary variable x is that of Chambers and Dunstan (1986), described below. The papers that followed, which we describe in this section, most immediately and notably the paper by Rao et al. (1990) can be thought of us the responses to the fundamental property of the Chambers Dunstan estimator: when the model it assumes is correct, it tends to be far and away better than other estimators; when the model is incorrect, the estimator has an inevitable bias, and it can do worse than even the naïve estimator.

3.1. The Chambers–Dunstan estimator

Chambers and Dunstan (1986) (henceforth CD) suppose a regression model holds such as

$$Y_i = x_i^T \beta + v_i^{1/2} \varepsilon_i, \tag{4}$$

where the errors $\varepsilon_i \sim G\left(0, \sigma^2\right)$ are independent, having some (typically unknown) distribution function G, with mean 0 and variance σ^2. The auxiliary variables x_i and v_i are assumed known for all units in the population. Then the central idea of CD is to estimate $z_j = I\left(Y_j \le t\right)$, for j in the nonsample r, by an estimate of its expectation under the model. From (4) we get that this expectation can be written

$$E\left(z_j\right) = G\left(\frac{t - x_j^T \beta}{v_j^{1/2}}\right).$$

Estimating this involves two steps using the sample data: (a) get a regression estimate $\hat{\beta}$ of β, and (b) use the (standardized) residuals

$$\hat{\varepsilon}_i = \left(y_i - x_i^T \hat{\beta}\right) \Big/ v_i^{1/2} \tag{5}$$

to estimate $G(u)$ by

$$\hat{G}(u) = n^{-1} \sum_{i \in s} I\left(\hat{\varepsilon}_i \leq u\right). \tag{6}$$

The result is the classic Chambers–Dunstan estimator

$$\hat{F}_{\mathrm{CD}}(t) = N^{-1} \left\{ \sum_{i \in s} I\left(y_i \leq t\right) + \sum_{j \in r} \hat{G}\left(\frac{t - x_j^T \hat{\beta}}{v_j^{1/2}} \right) \right\}. \tag{7}$$

$$= N^{-1} \left\{ \sum_{i} I\left(y_i \leq t\right) + n^{-1} \sum_{i} \sum_{j} I\left(\hat{\varepsilon}_i \leq \frac{t - x_j^T \hat{\beta}}{v_j^{1/2}} \right) \right\} \tag{8}$$

We note:

(1) $\hat{F}_{\mathrm{CD}}$ is a distribution function, fulfilling both parts of Property 1.

(2) It is calibrated: if y_i is any of the components x_{ki} of x_i, then $\hat{\beta}$ is 1 in the kth place, 0 elsewhere, the residuals $\hat{\varepsilon}_i$ are all zero, and the estimator reduces to $F(t)$.

(3) The model that CD focused on was the through the origin model

$$Y_i = \beta x_i + v_i^{1/2} \varepsilon_i. \tag{9}$$

However, they make it clear that their approach suits the more general model (4).

(4) In the case of no auxiliary information, that is, in the case of the simple model $Y_i = \mu + \varepsilon_i$, $\hat{F}_{\mathrm{CD}}$ reduces to the naïve estimator $\hat{F}_n$. However, if the model is expanded slightly to $Y_i = \mu + v_i^{1/2} \varepsilon_i$, the result is not an estimator of the form (1).

(5) If the working model (4) holds, $\hat{F}_{\mathrm{CD}}$ has a strong tendency to be much more efficient than $\hat{F}_\pi$ and other estimators of the cdf. This has been verified in a great variety of simulation studies. However, under unusual circumstances, unlikely, we believe, to be encountered in practice, $\hat{F}_{\mathrm{CD}}$ can have a falling off of efficiency, even if the model is correct (Chambers et al., 1992); see the discussion in Section 3.3. From a practical point of view, this is probably not of much consequence, but from a theoretical viewpoint it is quite intriguing. There is no analogous property among standard model-based estimators of means and totals. This suggests that indirect estimates of means using the cdf are unlikely always to be competitive.

(6) More serious is the robustness question: what happens to $\hat{F}_{\mathrm{CD}}$ if the specific regression model adopted as the "working model" is mistaken? CD show that even if only the variance structure is misconstrued, biases can arise. This is a more serious consideration from both the theoretical and practical standpoints, and is, in large measure, the motivating force behind the development of a large number of alternative estimators.

3.2. The Rao–Kovar–Mantel estimator

An estimator that is design consistent, and which, in effect, compensates for model misspecification, is the difference estimator of Rao et al. (1990) (henceforth RKM), also constructed with reference to some specific linear regression model represented generally by (4)

$$\hat{F}_{\mathrm{RKM}}(t) = N^{-1}\left\{\sum_{i\in s} d_i I\left(y_i \le t\right) + \sum_{k\in U}\hat{G}_{\pi}\left(\frac{t - x_k^T\hat{\beta}_{\pi}}{v_k^{1/2}}\right) - \sum_{i'\in s} d_{i'}\hat{G}_{\pi C}\left(\frac{t - x_{i'}^T\hat{\beta}_{\pi}}{v_{i'}^{1/2}}\right)\right\}, \tag{10}$$

Where $\hat{\beta}_{\pi}$ is the weighted least squares estimate of β using weights $d_i = \pi_i^{-1}$,

$$\hat{G}_{\pi}(u) = \frac{\sum_{i\in s}\pi_i^{-1} I\left(\hat{e}_{\pi i} \le u\right)}{\sum_{i\in s}\pi_i^{-1}}, \quad \text{with } \hat{e}_{\pi i} = \frac{y_i - x_i^T\hat{\beta}_{\pi}}{v_i^{1/2}}, \tag{11}$$

and

$$\hat{G}_{\pi C}\left(u_{i'}\right) = \frac{\sum_{i\in s}\pi_{i'}\pi_{i'i}^{-1} I\left(\hat{e}_{\pi i} \le u_{i'}\right)}{\sum_{i\in s}\pi_{i'}\pi_{i'i}^{-1}}, \quad \text{for } u_{i'} = \frac{t - x_{i'}^T\hat{\beta}_{\pi}}{v_{i'}^{1/2}}. \tag{12}$$

We note:

(1) Because of the differencing, $\hat{F}_{\mathrm{RKM}}$ is not necessarily a distribution function: it can fail both the boundary and monotonic aspects of Property 1. We do not regard this as a serious concern. Algorithms such as the pool-adjacent-violaters exist to make rectifications where needed. This should not be much of a problem.
(2) The estimator is calibrated (fulfills Property 5).
(3) The estimator is complicated. In particular, it incorporates second-order inclusion probabilities, a fact that causes no problems in simple random or stratified sampling, but in general is a nuisance.
(4) The estimator is design and model unbiased, and, if the model is even roughly correct will tend to do better than $\hat{F}_{\pi}$. $\hat{F}_{\mathrm{RKM}}$ and simplified variants of $\hat{F}_{\mathrm{RKM}}$ will do better than $\hat{F}_{\mathrm{CD}}$ under severe misspecification of the model.
(5) Nonetheless, if the working model holds true, then $\hat{F}_{RKM}$ can be considerably less efficient than $\hat{F}_{\mathrm{CD}}$. Often, for a given not-well-modeled population, $\hat{F}_{RKM}$ will be considerably better than $\hat{F}_{\mathrm{CD}}$ for some values of t, and the reverse for other values of t; for example, see Table 2 in RKM.

3.3. Asymptotic variances for CD and RKM estimators

Under the working model, both $\hat{F}_{\mathrm{CD}}$ and $\hat{F}_{\mathrm{RKM}}$ will have negligible bias. CD gave an analytic expression for the variance of $\hat{F}_{\mathrm{CD}}$ under the model (9). Chambers et al. (1992) (henceforth CDH) give expressions for the asymptotic variances of both $\hat{F}_{\mathrm{CD}}$ and $\hat{F}_{\mathrm{RKM}}$.

They assume a simple homoscedastic linear model $Y_i = a + bx_i + \varepsilon_i$, where the errors ε_i are independent and have distribution function G with mean 0 and variance σ^2. The density function $g = G'$ is assumed to exist. Sample and nonsample x are assumed to have a common asymptotic density $d(x)$, with μ and τ^2 the corresponding mean and variance of x.

They define four integrals:

$$I_1 = \int (x-\mu) g\{t-(a+bx)\} d(x)dx$$

$$I_2 = \int\int G\left[\{t-(a+bx)\} \wedge \{t-(a+bx^*)\}\right] d(x)d(x^*)dxdx^*$$

$$\text{(where } a \wedge b = \min(a,b)\text{)} \qquad (13)$$

$$I_3 = \int G\{t-(a+bx)\} d(x)dx$$

$$I_4 = \int \left[G\{t-(a+bx)\} - G\{t-(a+bx)\}^2\right] d(x)dx,$$

and prove

$$\operatorname{var}\left(\hat{F}_{\mathrm{CD}}(t) - F(t)\right) = n^{-1}(1-\pi)\left\{\tau^{-2}\sigma^2 I_1^2 + I_2 - I_3^2\right\} + N^{-1}(1-\pi)I_4 + o\left(n^{-1}\right) \qquad (14)$$

$$\operatorname{var}\left(\hat{F}_{\mathrm{RKM}}(t) - F(t)\right) = n^{-1}(1-\pi)I_4 + N^{-1}(1-\pi)I_4 + o\left(n^{-1}\right), \qquad (15)$$

where $\pi = \lim\left(N^{-1}n\right)$, as $n, N \to \infty$, assumed to exist.

Note that I_1, etc. are more properly $I_1(t)$, etc, functions of t.

Comments

(1) The term $N^{-1}(1-\pi)I_4$, common to the two variances, is the limit of the variance of the unknown, nonsample component $N^{-1}\sum_{j\in r} I(y_i \le t)$ of the target $F(t)$. As in the case of estimation of totals (see Chapter 23), the variance corresponding to the nonsample term tends to be an order of magnitude less than the variance of the sample component.

(2) CDH note that $I_2 \le I_3^2 + I_4$, so that, except for the term with I_1 in it, $\operatorname{var}\left(\hat{F}_{\mathrm{CD}}(t) - F(t)\right) \le \operatorname{var}\left(\hat{F}_{\mathrm{RKM}}(t) - F(t)\right)$. This term arises out of the fact that we only have estimates of the parameters a and b, and leads to the surprising fact that even when the working model is correct, $\hat{F}_{\mathrm{CD}}$ can do relatively badly for some t. CDH constructed a population ("the CDH population") to illustrate by simulation such a situation. The population had a rather unusual configuration of error and x distributions, which made I_1 large for a particular value of t.

3.4. The weighted average estimator

On the basis of the asymptotic variances (14) and (15), Wang and Dorfman (1996) (henceforth WD) construct a weighted average of the CD and RKM estimators: $\hat{F}_{\mathrm{WA}}(t) = w\hat{F}_{\mathrm{CD}} + (1-w)\hat{F}_{\mathrm{RKM}}$. The weight w is a function of $I_1, \ldots, I_4$ above, calculated to

achieve minimal (asymptotic) mean square error of the resulting estimator, and estimable from the available data. Note that w, like the I, depends on t. In simulation studies on both real and artificial populations, $\hat{F}_{\mathrm{WA}}$ performs well compared to $\hat{F}_{\mathrm{CD}}$ and $\hat{F}_{\mathrm{RKM}}$. Table 1 gives results on root mean square error for samples of size 60 from the "Beef Population," abstracted from Tables 3 and 4 in WD.

Two sets of estimators are considered, one based on a linear model of y on x, which fails adequately to capture the curvature in the data, the other modeling $\log(y)$ on $\log(\log(x))$, which fits the data well (see Remark 1.1). The winner within each set is in bold italics. $\hat{F}_{\mathrm{WA}}$ is never worse than second place, and is usually the winner or nose-to-nose with the winner. This suggests it adapts well, even if the working model is incorrect. There is another lesson to be gleaned from these results, noted in the next to last paragraph of the next section.

3.5. The fundamental issue of diagnostics

One of the effects of the dominance of the design-based approach to survey sampling has been the isolation of finite population estimation from mainstream statistics, which is dominated by modeling and data exploration. One striking aspect of this is the unwillingness of survey samplers to pay attention to the modeling process. This has special importance for the estimation of finite population cdf. In the case of estimating totals or means, there are interesting workarounds, combining balanced designs and weighting, that mitigate the effects of model failure (see Chapter 26). This appears to be impossible for the model-based CD estimator of a cdf. Thus the question of model diagnostics, which plays such a prominent role in regression theory and practice and in statistical practice generally, may in the case of estimating cdf be very important.

As noted earlier, the RKM estimator and its variants automatically adjust to model misspecification, and can do a good deal better than CD if the working model is wrong (or very peculiar, as in the CDH population.) On the other hand, if the model is reasonably correct, CD will, as a rule, far outperform RKM.

Dorfman (1993) examines the issue and concludes that under the circumstance, of say time constraints or analyst inexperience, where careful diagnostics and modeling does not occur, the RKM estimator is the safer bet. On the other hand, it pays to do such analysis, and great efficiencies tend to come from using the proper model-based estimator $\hat{F}_{\mathrm{CD}}$.

One can go a bit further. The degree of correctness of the model has an impact on the RKM estimator itself. One sees that in Table 1, where RKM under the transformed model does much better than RKM under the untransformed model; for example, in estimating the median transformed RKM is 64% more efficient than RKM based on the untransformed model. The same holds also for $\hat{F}_{\mathrm{WA}}$. Thus even apart from the issue of which estimator, diagnostics can be important for getting the best out of the estimator of choice.

One open question is to what extent the diagnostic process can be automated, replacing the analysts' observations and decisions by a well designed sequence of testing. It is not clear that this would improve on the results of automatic model adaptation of nonparametric regression, the use of which for estimating cdf is described in Section 3.8.

Table 1
Root mean square error × 10,000, 500 random samples ($n = 60$) from the beef population ($N = 410$), from Wang and Dorfman (1996)

p:	0.10	0.25	0.50	0.75	0.90
Untransformed					
$\hat{F}_{\pi}$	335	529	609	523	383
$\hat{F}_{\mathrm{RKM}}$	313	447	***447***	***370***	322
$\hat{F}_{\mathrm{CD}}$	286	618	800	433	311
$\hat{F}_{\mathrm{WA}}$	***267***	***440***	541	409	***306***
$\log(y+1)$ versus $\log(\log(x))$					
$\hat{F}_{\pi}$	358	505	604	533	371
$\hat{F}_{\mathrm{RKM}}$	303	383	349	330	279
$\hat{F}_{\mathrm{CD}}$	***214***	***293***	***227***	***204***	***177***
$\hat{F}_{\mathrm{WA}}$	222	296	***227***	209	191

3.6. RKM-like and CD-like estimators

There exist several estimators that are variants on the CD and RKM estimators.

3.6.1. CD-like estimators

Instead of estimating $G(u)$ using the sample residuals as in (5), Mak and Kuk (1993) take G to be a normal $(0, \sigma^2)$ distribution, replacing $\hat{G}\big((t - x_j^T\hat{\beta})\big/\big(v_j^{1/2}\big)\big)$ in the CD estimator (7) mentioned earlier, by $\Phi\big((t - x_j^T\hat{\beta})\big/\big(\hat{\sigma}v_j^{1/2}\big)\big)$, where $\hat{\sigma}$ is the estimate of standard deviation from the weighted regression fit. The main advantage of this is relative ease of calculation. It is not clear how well this alternative, $\hat{F}_{\mathrm{CD},\Phi}$, performs against CD calculated as at (7,8).

CD had noted (see Section 3.1 mentioned earlier) that, unlike estimators of mean or total, $\hat{F}_{\mathrm{CD}}$ was sensitive to departures from the working model for the variances of the errors. Lombardía et al. (2005) construct a CD-like estimator based on nonspecific variance structure, replacing the prespecified $v_j = v(x_j)$ in (4) by nonparametric regression estimates of variances at the sample and nonsample points. They take $\hat{v}(x_j) = \sum_{i\in s} w_{ij}(b)r_i^2$, where r_i are the residuals from the fit of sample y on sample x using the working model, and the $w_{ij}(b)$ are nonparametric regression weights, larger for sample units i with x_i closer to x_j, and b is a "bandwidth" which controls how flat the weights are. Nonparametric regression is discussed further in Section 3.7; see also Chapter 27. In a simulation study on several well-behaved populations, using the standard linear model $Y_i = \alpha + \beta x_i + v_i^{1/2}\varepsilon_i$, their estimator $\hat{F}_{\mathrm{CD},\hat{v}}$ does consistently better than a standard CD estimator based on the homoscedastic assumption $v_i = 1$.

Welsh and Ronchetti (1998) robustify the CD estimator against the presence of outliers (outliers in the sense of sharp local deviations of y from the working model) in two ways: (1) they replace the standard least squares estimator of β by a robust estimator (they choose the bi-weight estimator $\hat{\beta}_{\mathrm{R}}$ (for a discussion of outlier-robust estimation see Chapter 11) and (2) they use robust estimation of the distribution function G, replacing the residuals $\hat{\varepsilon}_i$ in (5) by the $c\hat{\sigma}$ scaled Huber ψ-function $\max\{-c\hat{\sigma}, \min(\hat{\varepsilon}_i, c\hat{\sigma})\}$, which puts bounds on how effectively large the contributing residuals can be. Here $\hat{\sigma}$ is a robust estimate of standard deviation equal to 1.4826 times the median absolute deviation of

the $\hat{\varepsilon}_i$ from their median value, and c is a positive tuning constant at the discretion of the user. They suggest varying c with the quantile at which the cdf is being estimated and give some tentative guidelines, recommending smaller values of c for t at the low end of y range and larger at larger. In a simulation study on the Beef Population (which has almost become the standard messy population in the finite population cdf literature), the resulting estimators at the 0.5, 0.75, and 0.90 quantiles improve considerably on the CD estimator with respect to mean square error.

A nonparametric version of the CD estimator is discussed in Section 3.7.3.

3.6.2. RKM-like estimators

Although $\hat{F}_{\mathrm{RKM}}$ originated as a design-based alternative to $\hat{F}_{\mathrm{CD}}$, it can also be viewed from a model-based perspective as an estimator that self corrects for model failure. Dorfman (1993) strips away much of the π structure to get a variant of $\hat{F}_{\mathrm{RKM}}$, which may be termed the residual corrected estimator

$$\begin{aligned}\hat{F}_{\mathrm{rc}}(t) &= N^{-1}\left\{\sum_{i\in s} I\left(y_i \le t\right) + \sum_{j\in r}\hat{G}\left(\frac{t - x_j^T\hat{\beta}}{v_j^{1/2}}\right) + \sum_{i\in s}\left(\pi_i^{-1} - 1\right)R_i\right\},\\ &= \hat{F}_{\mathrm{CD}}(t) + N^{-1}\sum_{i\in s}\left(\pi_i^{-1} - 1\right)R_i \qquad (16)\end{aligned}$$

where the upper level residuals $R_i = I\left(y_i \le t\right) - \hat{G}\left(\frac{t - x_i^T\hat{\beta}}{v_i^{1/2}}\right)$, and G and β are estimated as for the CD estimator. This is of the same form as Eq. (10), but avoids the design-based components (11) and (12). A similar estimator can be found in Godambe (1989). If the factors $\pi_i^{-1} - 1$ are a reasonable reflection of the number of nonsample units that are like the ith sample unit, then the residual adjustment term will give a measure of the difference of what the working model is yielding and what it should yield, blown up to nonsample size, yielding a kind of model-robustness. $\hat{F}_{\mathrm{RKM}}$ basically does the same thing, but $\hat{F}_{\mathrm{rc}}$ has the advantage of avoiding extra π calculations, in particular the second-order probabilities. In practice, $\hat{F}_{\mathrm{RKM}}$ and $\hat{F}_{\mathrm{rc}}$ will behave quite similarly (see Tables 3 and 4 of Dorfman (1993)). Of course, if the model is correct, the third term adds noise, and these estimators will not do as well as $\hat{F}_{\mathrm{CD}}$.

Chambers et al. (1993) (henceforth CDW) take (16) one step further, replacing the "a priori" factors $\pi_i^{-1} - 1$ by weights derived from the relation of units in the particular, chosen sample to those in the nonsample. Suppose we had the upper level residuals R_i for *all* units in the population. Then $\hat{F}^*_{\mathrm{adj}}(t) = \hat{F}_{\mathrm{CD}}(t) + N^{-1}\sum_{j\in r} R_j$ equals the target cdf $F(t)$ by the definition of the R_i. The nonsample R_j can be estimated using nonparametric regression: $\hat{R}_j = \sum_{i\in s} w_{ij}(b)R_i$, where the weights $w_{ij}(b)$ are larger the nearer x_i is to x_j. The bandwidth b measures how close an x_i need be to x_j to give much weight to the ith sample unit. CDW describe one method for selecting b, but bandwidth selection is always a difficult issue. Section 3.7 and also Chapter 27 further discuss nonparametric regression. The resulting estimator takes the form

$$\hat{F}_{\mathrm{CDW}}(t) = \hat{F}_{\mathrm{CD}}(t) + N^{-1}\sum_{i\in s} w_i R_i, \qquad (17)$$

where $w_i = \sum_{j \in r} w_{ij}(b)$ represents the total contribution the ith sample point makes to estimating the different nonsample R_j. This estimator appears to have some asymptotic advantage over $\hat{F}_{\text{RKM}}$ or $\hat{F}_{\text{rc}}$ (see discussion, Section 3.7.3), but it is more complicated to calculate, and a distinct advantage has not yet been shown in simulation studies. Like $\hat{F}_{\text{RKM}}$, it is not necessarily a proper distribution function.

As with the CD estimator, there is a nonparametric regression version of the RKM estimator, given in Section 3.7.3. Additionally, there is the family of calibration estimators, described in Section 3.8, which bear a strong kinship to RKM.

3.7. Nonparametric regression-based estimators

It is an interesting fact that the first use of nonparametric regression in survey sampling (discussed more broadly in Chapter 27) was for the purpose of estimating the distribution function. This is not entirely a historical accident. In estimating totals and means, robustness against model failure can be achieved by adroit sampling with concomitant weighting (see Chapter 23). This is not the case for the CD estimator of the cdf. A natural alternative to the model diagnosis of Section 3.5 is to weaken the model so its failure is unlikely. Nonparametric regression assumes only that the expected value of the target variable given x is a smooth function.

Dorfman and Hall (1993) (henceforth DH) describe three models in which nonparametric regression can play a role:

MODEL 1. *Conditional on x, y obeys a parametric model, such as the linear model (4).*

MODEL 2. *The expectation of y given x is a smooth function of unspecified form:* $y_i = m(x_i) + \varepsilon_i$, *with* ε_i *independent with a common distribution function* G.

MODEL 3. *The expectation of* $z_i \equiv h(y_i) \equiv I(y_i \leq t)$ *given x is a smooth function of unspecified form:* $E(z_i) = H(x_i)$.

It may be worth noting that in case 3, there are as many models implied as values t of interest. They do not consider a 4th case:

MODEL 4. *Conditional on x,* $z_i \equiv h(y_i) \equiv I(y_i \leq t)$ *obeys a parametric model, such as the logistic.* To our knowledge, the possible use of this model for estimating cdf has not been studied.

We have already seen a use of nonparametric regression in Model 1 by CDW (Section 3.6.2) and Lombardía et al. (2005) (Section 3.6.1). The original nonparametric regression estimator, due to Kuo (1988), relies on Model 3. The nonparametric CD and RKM estimators, described later, rely on Model 2.

Consider Model 2. There are many methods for estimating $m(x_i)$, all coming under the heading of nonparametric regression, and typically coming down to an estimate of the form $\hat{m}(x_j) = \sum_{i \in s} w_{ij}(b)\, y_i$. Here x_j may be a sample or nonsample point and the $w_{ij}(b)$ are weights that depend on the type of nonparametric regression and on a tuning

parameter b that needs to be carefully selected. For example, we might take $w_{ij}(b)$ to be kernel weights

$$w_{ij}(b) = \frac{k\left(b^{-1}\left[x_i - x_j\right]\right)}{\sum\limits_{i' \in s} k\left(b^{-1}\left[x_{i'} - x_j\right]\right)}, \tag{18}$$

where $k(\ )$ is typically a symmetric density function with finite support such as the Epanichnikov $k(u) = 0.75\left(1 - u^2\right) I\left(|u| \leq 1\right)$, and b is the bandwidth that controls how flat and extensive the weights will be.

In using nonparametric regression the question arises as to what the role is of the sampling weights. Traditional sampling would still seem to require their presence. The other view, as is emphasized in the CDW study, is that nonparametric weights give an *alternative* to sampling weights for tying the given sample units to the population units they represent. We expect that as a rule adding in the sampling weights will not affect results too much, for better or worse.

3.7.1. The Kuo estimator

Kuo (1988) suggested the following estimator

$$\hat{F}_{\text{Kuo}}(t) = N^{-1}\left\{\sum_{i \in s} I\left(y_i \leq t\right) + \sum_{j \in r} \hat{P}\left(y \leq t | x_j\right)\right\},$$

with $\hat{P}\left(y \leq t | x_j\right) = \sum_{i \in s} w_{ij}(b) I\left(y_i \leq t\right)$. Kuo actually used nearest neighbor weights: the k units with x values nearest x_j have equal weight, the remainder weight *zero*. The implied model is Model 3 and $H(x) = E\left(I(y \leq t)|x\right) = P(y \leq t|x)$. $\hat{F}_{\text{Kuo}}$ is readily seen to be of the form (1), with weights $w_i = \sum_{j \in r} w_{ij}(b)$ that are necessarily positive, and so readily yields quantile estimates. It is not calibrated with respect to x. Dorfman and Hall (1993) (DH) give the asymptotic bias and variance of $\hat{F}_{\text{Kuo}}$; see Section 3.7.3.

3.7.2. Kuo alternatives

Kuk (1993) likewise employs nonparametric regression estimates of the conditional distribution function $P(Y \leq t \,|\, x)$ and takes $\hat{F}_{\text{kuk,np}}(t) = \sum_{k \in U} \hat{P}\left(Y \leq t \,|\, x_k\right)$, with $\hat{P}(Y \leq t \,|\, x) = \sum_s d_i k\left(b^{-1}\left[x - x_i\right]\right) K\left(b^{-1}\left[t - y_i\right]\right) \Big/ \sum_s d_i k\left(b^{-1}\left[x - x_i\right]\right)$, where k is a selected density function and K the corresponding distribution function. This estimator is conceived in the spirit of design-based sampling, and we note three differences from the Kuo: (1) it uses the design weights d_i, (2) it estimates the known sample portion of $F(t)$, and (3) it replaces the indicator variables I by K. For the bandwidth, Kuk uses $b = R_x/n$, where R_x is the population range of the x-values; also, Y is rescaled to match its range to x, in order that b apply to both variables. The estimator is a cdf. In simulations, it outperforms RKM for several modes of sampling on two populations. We would expect it and $\hat{F}_{\text{Kuo}}(t)$ to behave similarly. (*Note*: Kuk uses the logistic distribution for K (and k). It might be safer to use a symmetric distribution, because a nonsymmetric K can shift the estimate of mean entailed by $\hat{F}_{\text{kuk,np}}$ by an additive term dependent on the bandwidth.)

In the case where x is a vector, it may be appropriate to use a smoothing device which appears in Durrant and Skinner (2006), namely to fit the sample y on the sample x and use the resulting fitted values $\hat{y}_k$, $k \in U$, in place of the x to form the weights w_{ij}.

DH considers a design-adjusted version of the Kuo, an analogue for Model 3 of the RKM estimator:

$$\hat{F}_{\text{Kuo,adj}}(t) = N^{-1}\left\{\sum_{i\in s} I\left(y_i \le t\right) + \sum_{j\in r} \hat{P}\left(y \le t|x_j\right) + \sum_{i\in s}\left(\pi_i^{-1} - 1\right)R_i\right\},$$

where $R_i = I\left(y_i \le t\right) - \hat{P}\left(y_i \le t|x_i\right)$. Theorem 3 of DH suggests a subtle difference between Kuo and Kuo-adjusted: when the sample and nonsample x densities are the same their "I_7" term goes out in the Kuo-adjusted. In empirical work on several artificial populations, using nonparametric local linear regression, Johnson et al. (2004) find the design-adjusted Kuo doing better than Kuo. Model 3 aforementioned actually fails only in one of the seven populations they use, the "Jump" population. Thus "twicing" (estimating and adding on an estimate of the error of one's estimate) seems to be helpful even when the assumed model is correct.

Nascimento Silva and Skinner (1995) suggest a poststratified estimator of $F(t)$. The range of x is partitioned into intervals of adjacent x, each interval corresponding to a poststratum g, $g = 1, \ldots, G$. Then $\hat{F}_{\text{ps}}(t) = \sum_{g=1}^{G} \frac{N_g}{N}\hat{F}_{\pi g}(t)$, where $\hat{F}_{\pi g}(t) = \sum_{i\in g\cap s} d_i I\left(y_i \le t\right)\Big/\sum_{i\in g\cap s} d_i$, and N_g is the number of population units in the gth poststratum. The basic model here is the same as for the Kuo and Kuk estimators, except that the underlying continuous $P(Y \le t|x)$ is being estimated by a step function. The choice of length and number of intervals defining the poststrata is analogous to the choice of bandwidth. In simulations, $\hat{F}_{\text{ps}}(t)$ with best stratification did well compared to $\hat{F}_{\text{Kuo}}(t)$ with a fixed b, but it is not known how they would compare under optimal choice of bandwidth for the Kuo. $\hat{F}_{\text{Kuo}}(t)$ can in general be expected to do somewhat better, because the boundaries of the intervals in $\hat{F}_{\text{ps}}(t)$ can prevent points with x near to x_i being used for estimation of $\hat{P}\left(y \le t|x_j\right)$. An exception would be the case where the underlying $P\left(y \le t|x_j\right)$ really is a step function and the selected poststratification matches it. On the other hand, the poststratified estimator has the advantage of simplicity and familiarity to survey samplers. Modarres (2002) suggests a similar idea and offers a modification in the special circumstance that F is known to have a point of symmetry. (See discussion of cdf symmetry in Section 2.2.)

3.7.3. Nonparametric CD and RKM

Based on Model 2, DH derive analogues of the CD and RKM estimators (their $\hat{F}_1$ and $\breve{F}_1$, respectively). Let $\hat{m}\left(x_k\right) = \sum_{i\in s} w_{ik} y_i$ be a nonparametric fit of y on x. They define

$$\hat{F}_{\text{np,CD}} = N^{-1}\left[\sum_{i\in s} I\left(y_i \le t\right) + \sum_j \hat{G}\left(t - \hat{m}\left(x_j\right)\right)\right]$$

$$\hat{F}_{\text{np,RKM}} = N^{-1}\left[\sum_{i\in s} I\left(y_i \le t\right) + \sum_j \hat{G}\left(t - \hat{m}\left(x_j\right)\right) + \sum_{i\in s}\left(\pi_i^{-1} - 1\right)R_{\text{np},i}\right],$$

where $R_{\text{np},i} = I\left(y_i \le t\right) - \hat{G}\left(t - \hat{m}\left(x_i\right)\right)$ with $\hat{G}(u) = n^{-1}\sum_{i\in s} I\left(\hat{\varepsilon}_i \le u\right)$ and $\hat{\varepsilon}_i = y_i - \hat{m}\left(x_i\right)$.

DH analyze eight estimators which, aside from the naïve estimator $\hat{F}_\pi$, belong to two classes: three for which a given chosen parametric (in fact, linear) model plays a role, namely $\hat{F}_{\mathrm{CD}}$, $\hat{F}_{\mathrm{RKM}}$, and $\hat{F}_{\mathrm{CDW}}$, and four nonparametric estimators ("the *np* group") for which it does not: $\hat{F}_{\mathrm{Kuo}}$, $\hat{F}_{\mathrm{Kuo,adj}}$, $\hat{F}_{\mathrm{np,CD}}$, and $\hat{F}_{\mathrm{np,RKM}}$, described earlier.

In the following description of variances and biases, we consider just the case where the sample and nonsample densities for x are the same, as would arise under srs or simple balanced sampling (see Chapter 23). Several formulas then simplify.

In terms of variance, the estimators fall into three categories: (a) estimators that have a variance asymptotically equivalent to that of $\mathrm{var}\left(\hat{F}_\pi\right) \approx n^{-1}(1-\pi)^2 I_4$, with I_4 as in (13) aforementioned (and referred to as I_6 in DH), (b) $\mathrm{var}\left(\hat{F}_{\mathrm{CD}}\right)$ as at (14) abovementioned, and (c) $\mathrm{var}\left(\hat{F}_{\mathrm{np,CD}}\right)$, which is similar in structure to $\mathrm{var}\left(\hat{F}_{\mathrm{CD}}\right)$, except that $\tau^{-2}\sigma^2 I_1^2$ is replaced by $\sigma^2 I_1^{(\mathrm{np})}$, with $I_1^{(\mathrm{np})} = \int \{g\,(t - m\,(x)) - \int g(t - m(x))d_s(x)\}^2 d_s(x)dx$ (notated I_3 in DH), where $d_s(x)$ is the density of x among sampled units. It is not clear that this term makes $\mathrm{var}\left(\hat{F}_{\mathrm{np,CD}}\right)$ more or less vulnerable to peculiar population structure than $\tau^{-2}\sigma^2 I_1^2$ makes $\mathrm{var}\left(\hat{F}_{\mathrm{CD}}\right)$. In any case this result suggests that $\hat{F}_{\mathrm{np,CD}}$ will have efficiency close to that of $\hat{F}_{\mathrm{CD}}$ without the same vulnerability to model failure. Provided the $I_1^{(\mathrm{np})}$ term is not large, we would expect $\hat{F}_{\mathrm{np,CD}}$ to be more efficient than the remaining estimators that DH consider, because of the inequality $I_2 \leq I_3^2 + I_4$ noted earlier.

Bias (compare DH, Table 1): Under suitable bandwidth choice, the *np* group will have bias that is $o\left(n^{-1/2}\right)$, lower than the order of their standard deviation. The naïve estimator $\hat{F}_\pi$ has bias $O\left(n^{-1/2}\right)$, same as the standard deviation. If the working model is correct, then the model dependent estimators $\hat{F}_{\mathrm{CD}}$, $\hat{F}_{\mathrm{RKM}}$, and $\hat{F}_{\mathrm{CDW}}$ all have bias of order $O\left(n^{-1}\right)$. If the model fails, they have bias $O(c)$, $O\left(n^{-1/2}\right)$, and $o\left(n^{-1/2}\right)$, respectively. DH note the relative strength of the CDW estimator compared to RKM. It is not clear from simulations under what circumstances their difference will be important. CDW did not find there to be much difference between them in their empirical study.

DH note that empirical comparison of the np estimators requires a practical means of selecting bandwidths. Lombardía et al. (2004) use a bootstrap methodology to estimate the optimal bandwidth b for $\hat{F}_{\mathrm{np,CD}}$. It would be of interest to apply their approach to estimating bandwidths of other estimators as well, and to compare the resulting estimates among themselves and to the parametric-based estimators.

Model 2 underlying $\hat{F}_{\mathrm{np,CD}}$ and $\hat{F}_{\mathrm{np,RKM}}$ is a homoscedastic model. We may expect that in the nonparametric case, deviations from homoscedasticity will have minimal impact. Nonetheless, this has not been investigated. We have noted the use of np regression by Lombardía et al. (2005) to estimate the variance function for the CD estimator. It would be of interest to carry out something parallel to this for $\hat{F}_{\mathrm{np,CD}}$ and see the effect.

3.8. Calibration estimators

Here we consider estimators that calibrate the estimate of the cdf with respect to the auxiliary x in a variety of ways, with the twin goals of design consistency and some degree of efficiency under a chosen model – protecting against the worst in case of

model failure and improving on the usual design-based estimator in the case of model correctness. Despite the name, not all members of this class of estimators are calibrated in the sense of fulfilling Property 5 of Section 1. The estimators do bear a kinship to the RKM estimator, often being asymptotically equivalent.

One set of methods for calibrating relies on the difference between a population total and a weighted sample sum whose terms are some function $g(.)$, possibly vector valued, of x : $\Delta_w(x_U) = \sum_{k\in U} g(x_k) - \sum_{i\in s} w_i g(x_i)$; g can depend on the sample y, and x other than the explicit x_k, as well. There are at least three methodological schemes for doing this: (a) explicit regression models (GREG) (see Chapter 25), (b) the use of calibration weights (also Chapter 25), and (c) the use of pseudo empirical likelihood (see Chapter 30).

For estimating the distribution function, all take the dependent variable as $z_i = I(y_i \le t)$.

(a) In the case of GREG, $w_i = d_i$, and $\Delta_d(x_U)$ is an explicit component of the estimator. The estimator can be written in the form $\hat{F}_{\text{GREG}}(t) = N^{-1}\{\sum_{i\in s} d_i I(y_i \le t) + \{\sum_{k\in U} g(x_k) - \sum_{i'\in s} d_i g(x_i)\}\hat{B}_\pi\}$, where $\hat{B}_\pi = \sum_{i\in s} d_i(g(x_i) - \bar{g}_\pi)I(y_i \le t)/\sum_{i\in s} d_i(g(x_i) - \bar{g}_\pi)^2$ and $\bar{g}_\pi = \sum_{i\in s} d_i g(x_i)/\sum_{i\in s} d_i$.

(b) In the case of calibration weights, the $w_i = p_i$ are such as to satisfy (i) $\Delta_p(x_U) = 0$, (ii) $\sum_{i\in s} p_i = N$ and at the same time minimize a chosen function measuring the distance between the sample w_i and d_i. The estimator is of the form $\hat{F}_{\text{cal}}(t) = N^{-1}\sum_{i\in s} p_i I(y_i \le t)$. In the particular case of the chi-square distance function $D(w, d) = \sum_{i\in s}(w_i - d_i)^2/(q_i d_i)$, (where the q_i are some user controlled weights, usually just a constant), the calibration estimator reduces to the GREG, mentioned earlier. The p_i need not be positive in this case.

(c) The pseudo empirical likelihood estimator can be regarded as exactly like the calibration situation, except now the p_i are such as maximize the empirical likelihood $l(p) = \sum_{i\in s} d_i \log p_i$ and there is an additional condition (iii) $0 < p_i$. (This follows by taking the standard formulation of pseudo empirical likelihood as in Chapter 30 and multiplying the p_i in that formulation by N.) Then pseudolikelihood maximum likelihood estimators are, again, of the form $\hat{F}_{pl}(t) = N^{-1}\sum_{i\in s} p_i I(y_i \le t)$, but now the p_i are conveniently positive, and satisfy Property 1a of Section 1.3. For some choices of $g(x)$ the p_i will depend on t, so Property 1b is not necessarily satisfied in general.

For a given choice of g, the calibration and pseudolikelihood estimators are asymptotically equivalent, and they tend to behave similarly to each other in simulations.

Wu and Sitter (2001a) (henceforth WS) suggest taking $g(x_i) = \hat{G}_\pi\big((t - x_i^T\hat{\beta}_\pi)/v_i\big)$, as in the RKM estimator. In that case we get: $\hat{F}_{\text{GREG}}(t) = N^{-1}\{\sum_{i\in s} d_i I(y_i \le t) + \{\sum_{k\in U}\hat{G}_\pi\big((t - x_k^T\hat{\beta})/(v_k^{1/2})\big) - \sum_{i'\in s} d_i\hat{G}_\pi\big((t - x_{i'}^T\hat{\beta})/(v_{i'}^{1/2})\big)\}\hat{B}_\pi\}$ where $\hat{B}_\pi = \sum_{i\in s} d_i\big(\hat{G}_i - \bar{G}_\pi\big)I(y_i \le t)/\sum_{i\in s} d_i\big(\hat{G}_i - \bar{G}_\pi\big)^2$, with $\hat{G}_i$ shorthand for $g(x_i) = \hat{G}_\pi\big((t - x_i^T\hat{\beta}_\pi)/v_i\big)$ and $\bar{G}_\pi = \sum_{i\in s} d_i\hat{G}_i/\sum_{i\in s} d_i$. [compare WS, Eq. (19)]. The major difference between this and RKM is the $\hat{B}_\pi$ [compare Eqs. (10–12)].

Kovacevic (1997) suggests two calibration estimators. First, $\hat{F}_{\text{KOV1}}$ using $g(x) = x$ and with distance function $D(w, d)$ other than the chi-square distance, designed to

guarantee that the p_i are nonnegative. This estimator does not require that all x in the population be known, but only the population totals. A second estimator $\hat{F}_{\mathrm{KOV2}}$ modifies $\hat{F}_{\mathrm{KOV1}}$ by including a bias corrected term, which does require full population information: $\hat{F}_{\mathrm{Kov2}}(t) = N^{-1}\sum_{i\in s} p_i\{z_i + N^{-1}\sum_{k\in U}\hat{G}_k - \hat{G}_i\}$, with the $\hat{G}_i$ defined as earlier. This estimator behaves very similarly to RKM.

Chen and Wu (2002) (henceforth CW) suggest three pseudolikelihood estimators. The first relies on a model

$$y_i = \mu(x_i, \beta) + v_i^{1/2}\varepsilon_i, \tag{19}$$

which, of course, includes the linear Model (4), except that they posit that the cdf of the errors is a normal distribution (cf. Kuk (1993), discussed in Section 3.7.2). They take $g(x_i) = \Phi\left(\frac{t_0-\mu(x_i,\hat{\beta}_\pi)}{v_i^{1/2}\hat{\sigma}_\pi}\right)$, where Φ is the standard normal distribution, for some fixed suitable preselected value t_0. The resulting estimator $\hat{F}_{\mathrm{pl,CW1}}(t) = N^{-1}\sum_{i\in s} p_i(t_0) I(y_i \le t)$ is asymptotically equivalent to a calibration estimator, satisfies Property 1b, and so is a distribution function. The estimator is design consistent for all t. It is model unbiased for $F(t_0)$, but not necessarily for $F(t)$, $t \neq t_0$, so there might be some loss of efficiency if t is far from t_0.

CW's second estimator $\hat{F}_{\mathrm{pl,CW2}}(t)$ relies on a model for $z_i = I(y_i \le t_0)$. CW adopt in particular a logistic model and take $g(x_i) = \exp\left(x_i^T\hat{\beta}_\pi\right)/\left(1+\exp\left(x_i^T\hat{\beta}_\pi\right)\right)$. $\hat{F}_{\mathrm{pl,CW2}}(t)$ has similar characteristics to $\hat{F}_{\mathrm{pl,CW1}}(t)$. Their third estimator $\hat{F}_{\mathrm{pl,CW3}}(t)$ returns to the model (19) and takes $g(x_i) = I\left(\hat{y}_{\pi i} \le t_0\right)$, where $\hat{y}_{\pi i} = \mu\left(x_i, \hat{\beta}_\pi\right)$ are the fitted values from the regression. In a simulation study on artificial populations CW find $\hat{F}_{\mathrm{pl,CW1}}(t)$ and $\hat{F}_{\mathrm{pl,CW2}}(t)$ outperforming $\hat{F}_{\mathrm{pl,CW3}}(t)$.

Rueda et al. (2007a) (hereafter RMMA) develop a calibration estimator, also based on $\hat{y}_{\pi i}$ and a fixed vector $t^* = (t_1, t_2, \ldots t_P)^T$, with $t_1 < t_2 < \cdots < t_P$, leading to a vector valued $g(x_i) = I\left(\hat{y}_{\pi i} \le t^*\right) = \left[I\left(\hat{y}_{\pi i} \le t_1\right), \ldots, I\left(\hat{y}_{\pi i} \le t_P\right)\right]^T$. RMMA use the chi-square distance and show that, if $q_i = c$, a constant, then the weights p_k take a surprisingly simple form and are positive (see RMMA, Section 4). If t_P is taken large enough, then $F_{\mathrm{cal,RMMA}}(t) \to 1$, as $t \to \infty$. This estimator is a cdf. In simulation studies on some actual populations, the estimator performs on a level with the CD estimator, and better than RKM. It has not been established how robust it is to model failure.

Rueda et al. (2007b) replaces $\hat{y}_{\pi i} = \mu\left(x_i, \hat{\beta}_\pi\right)$ in the above by a nonparametric fit $\hat{y}_{\pi i} = m(x_i)$. See Section 3.7 described earlier for a discussion of nonparametric regression.

Harms and Duchesne (2006) suggest a calibration estimator based on a population x-quantile, as an intermediate step to estimating quantiles. Their estimator differs from the above in replacing

$$z_i = I(y_i \le t) \quad \text{by} \quad \varphi(y_k; t) = \begin{cases} \frac{t-y_{k-1}}{y_k-y_{k-1}}, & y_{k-1} \le t < y_k \\ I(y_k \le t) & \text{else} \end{cases}.$$

The resulting estimator

$$\hat{F}_{\mathrm{HD}}(t) = N^{-1}\sum_s w_i\varphi(y_k; t) \tag{20}$$

is an interpolated continuous linear function in t. As in the above estimators the weights minimize a distance function $D(w, d)$. The constraints are $\sum_{i\in s} w_i = N$ and $\hat{Q}_{\mathrm{HD},x}(p_0) = Q_x(p_0)$, where $Q_x(p_0)$ is a chosen (or available) population quantile on the x-variables. They allow x to be vectorial, and so in general $Q_x(p_0)$ is a vector of length $J \geq 1$. The sample estimates are defined in terms of a sample cdf of the form (20): $\hat{Q}_{\mathrm{HD},x}(p_0) = \hat{F}^{-1}_{\mathrm{HD},x}(p_0)$. Their primary focus is on estimating quantiles; see Section 5.2 below for further discussion.

Kuk and Mak (1994) (henceforth KM) replace the difference form of the abovementioned calibration estimators by a relation between distribution functions. They offer two estimators, one which does not rely on an explicit model, the other, like $\hat{F}_{\mathrm{RKM}}$, which does. The first, the simple transformation model, is

$$\hat{F}_{\mathrm{KM1}}(t) = G\Big(\hat{G}_\pi^{-1}\big(\hat{F}_\pi(t)\big)\Big)$$

where G is the population distribution function for x, and $\hat{G}_\pi$ is its Hajek estimator. They suggest making both $\hat{F}_\pi$ and $\hat{G}_\pi$ continuous by linear interpolation before applying this transformation. KM point out that there is a tacit weak model here, namely y being some monotone increasing function of x. Where this condition is met, but an explicit working model is false (e.g., the working model assumes linearity, but there is strong curvature) this estimator can do considerably better than estimators like $\hat{F}_{\mathrm{RKM}}$ which incorporate the model (cf. KM, Figure 1, and Table 1, Factory Population).

KMs second estimator is

$$\hat{F}_{\mathrm{KM2}}(t) = H\Big(\hat{H}_\pi^{-1}\big(\hat{F}_\pi(t)\big)\Big),$$

where $H(t) = \sum_{k\in U} \hat{G}_\pi\Big(\frac{t - x_k^T\hat{\beta}_\pi}{v_k^{1/2}}\Big)$ and $\hat{H}_\pi(t) = \sum_{i\in s} d_i \hat{G}_\pi\Big(\frac{t - x_i^T\hat{\beta}_\pi}{v_i^{1/2}}\Big)\Big/\sum_{i\in s} d_i$. Especially when the model is correct or nearly correct, this estimator behaves much like $\hat{F}_{\mathrm{RKM}}$, with the added advantage of being a strict distribution function.

4. Estimating the distribution function using partial auxiliary information

Several estimators of Section 3 can be applied with a limited amount of population information on the auxiliary. The poststratified estimator $\hat{F}_{\mathrm{ps}}$ of Nascimento Silva and Skinner (1995) requires only interval information. The calibration estimator $\hat{F}_{\mathrm{Kov1}}$ of Kovacevic (1997) requires only means or totals. The calibration estimator $\hat{F}_{\mathrm{HD}}$ of Harms and Duchesne (2006) requires only knowledge of population quantiles.

Dunstan and Chambers (1989) adapt the CD estimator to the situation in which group information is available on x. Assumed known is the number N_h of x in each of H intervals $[x_{hL}, x_{hU}]$, $h = 1, \ldots, H$, the boundaries themselves, as well as the within interval means $\bar{x}_h$. Positing within interval cdf $C_h(x)$ for the x, and conditioning on the known residuals from the sample fit, they derive an expression for the expectation of the double sum in (8) as $\sum_h (N_h - n_h)\left\{1 - n^{-1}\sum_{i\in s}\Gamma_{ht}\left(\hat{\varepsilon}_i\right)\right\}$, where Γ_{ht} is the cdf of the transformed variable $\big(t - \hat{\beta}x_{hi}\big)\big/v_{hi}^{1/2}$, it being assumed that $v_{hi}^{1/2}$ is a well-defined function of x_{hi}. They work out details for the case where $v_{hi} = x_{hi}$ and C_h

is a split uniform distribution on $[x_{hL}, x_{hU}]$ with mean $\bar{x}_h$. They also derive expressions for variance estimation. In simulations on a population where the linear model holds pretty well, behavior of the resulting estimator $\hat{F}_{\mathrm{CD,lim}}$ closely approximates that of $\hat{F}_{\mathrm{CD}}$. It would be interesting to see how their estimator compares to $\hat{F}_{\mathrm{ps}}$, in the case where just interval boundary information is available.

Kuk and Mak (1989) suggest an estimator of the population cdf which relies only on knowing the population median $Q_x(1/2)$ of an auxiliary x. Break the sample into two components: s_1, those with $x_i \leq Q_x(1/2)$, and s_2, those with $x_i > Q_x(1/2)$. (Under unequal probability sampling, these groups could be quite different in size.) Let n_{x1} be the number of units in the first group and n_{x2}, in the second, and define $\hat{F}_{\mathrm{KM},Q(1/2)}(t) = \frac{1}{2}\big(\hat{F}_1(t) + \hat{F}_2(t)\big)$, where $\hat{F}_l(t) = \frac{1}{n_l}\sum_{i\in s_l} I(y_i \leq t)$. This is a well defined cdf. Their primary concern is estimating the population median for y, which can be readily gotten by inverting $\hat{F}_{\mathrm{KM},Q(1/2)}(t)$. In simulations it does not do quite as well as their direct "position" estimate of $Q_y(1/2)$ (see Section 5.2).

5. Quantile estimation

The primary way to get estimates of quantiles is by inverting estimates of distribution functions, as we briefly describe in Section 5.1. In addition to estimates by inversion, there are a number of direct estimates of quantiles suggested in the literature, described in Section 5.2. Also, there is a class of what we shall here call "hinge estimates," which avoid extensive repeated calculations on a complicated cdf estimator by making use of $\hat{F}_\pi$ or other simply calculated estimator (Section 5.3).

5.1. *Inversion of estimates of the distribution function*

Estimates of quantiles are most readily achieved through inverting an estimate of the distribution function. Thus all of the estimates $\hat{F}_*(t)$ of distribution function described earlier yield a corresponding quantile estimate

$$\hat{Q}_*(p) = \min\{t : \hat{F}_*(t) \geq p\}. \tag{21}$$

Basically, one gets a grid of values $\hat{F}_*(t_\nu)$, $t_1 < t_2 < \cdots < t_\nu < \cdots < t_{\nu *}$ so that the values $\hat{F}_*(t_\nu)$ surround p and takes the smallest of the t_ν satisfying (21). It is desirable that the estimator have the monotonicity property: $p < p' \Rightarrow \hat{Q}_*(p) \leq \hat{Q}_*(p')$. This requires that $\hat{F}_*(t)$ is a proper distribution function; in some cases this may require isotonization.

The estimators $\hat{F}_{\mathrm{KM},Q(1/2)}$ of Kuk and Mak (1989) and $\hat{F}_{\mathrm{HD}}$ of Harms and Duchesne (2006) are specifically intended for quantile estimation (see comment in Section 5.2).

5.2. *Direct estimates*

By a direct estimate of quantile, we mean one that does not require an explicit expression for the cdf, $\hat{Q}_w(p)$ in the next paragraph being the prime example.

We have already noted that estimators of the form (1) with fixed positive w_i are particularly easy to invert: we order the sample y by size and take $\hat{Q}_w(p)$ to be the smallest y_i such that $\sum_1^i w_{i'} \geq p$. $\hat{Q}_n(p)$, the sample quantile when the weights are all

equal, is especially simple. In the case of small samples, it may be worthwhile to interpolate between the sample y, effectively linearizing the estimate of cdf. Additionally, one might want to modify slightly the weights themselves, to achieve analogues to the quantile estimates available in standard software packages (compare Hyndman and Fan (1996) and Section 2.2 earlier).

Rao et al. (1990) suggest readily calculated ratio and difference estimators:

$$\hat{Q}_{\text{rat}}(p) = \left\{\hat{Q}_{\pi t}(p) \big/ \hat{Q}_{x\pi}(p)\right\} Q_x(p)$$
$$\hat{Q}_{\text{dif}}(p) = \hat{Q}_{\pi t}(p) + \hat{\beta}_\pi\left(Q_x(p) - \hat{Q}_{x\pi}(p)\right),$$

with $\hat{\beta}_\pi = \sum_s d_i y_i \big/ \sum_s d_i x_i$. The ratio estimator requires that x be a scalar, but the difference estimator readily generalizes to vector x (Kovacevic, 1997). In any case the only information required beyond the sample is the appropriate population quantile for x. It may noted that in neither is monotonicity guaranteed. In rare cases, one might have to make adjustments. Empirical work suggests both do better than $\hat{Q}_\pi(p)$. Harms and Duchesne (2006) carried out an empirical study on several populations and under two sampling plans, comparing $\hat{Q}_\pi$, $\hat{Q}_{\text{rat}}$, $\hat{Q}_{\text{dif}}(p)$, $\hat{Q}_{\text{HD}} = \hat{F}_{\text{HD}}^{-1}$ and $\hat{Q}_{\text{CD}} = \hat{F}_{\text{CD}}^{-1}$. Typically, apart from the CD estimator, which could strongly lead or lag, the winners with respect to mean square error were $\hat{Q}_{\text{rat}}$ or $\hat{Q}_{\text{HD}}$. There were instances where $\hat{Q}_{\text{rat}}$ was a good deal worse than $\hat{Q}_{\text{HD}}$, but not the reverse, so overall $\hat{Q}_{\text{HD}}$ performed well.

Kuk and Mak (1989) suggest a position estimator $\hat{Q}_{\text{pos}}(1/2)$ of the population median of y that relies on knowing only the population median $Q_x(1/2)$ of an auxiliary x. The basic idea is this: there exists some p between 0 and 1 such that the sample distribution function evaluated at the unknown population median is $p : \hat{F}_n\left(Q(1/2)\right) = p$. Thus, a good estimate of p will yield a good estimate of $Q(1/2)$. This is achieved as follows: Break the sample into two components: s_1', those with $x_i \leq \hat{Q}_x(1/2)$, the x sample median, and s_2', those with $x_i > \hat{Q}_x(1/2)$. Let n_{x1} be the number of units in s_1' and n_{x2}, in s_2'. Let $\hat{p}_{1|1}$ be the proportion of points in the first group s_1' for which y is less than the sample median, and $\hat{p}_{1|2}$ be the like proportion of the points in s_2'. Then $\hat{p} = n^{-1}\left(n_{x1}\hat{p}_{1|1} + n_{x2}\hat{p}_{1|2}\right)$ is an estimate of the fraction p of points i in the sample having y_i less than or equal to $Q_y(1/2)$, and we take the position estimator $\hat{Q}_{\text{pos}}(1/2) = Q_n\left(\hat{p}\right)$, the $\hat{p}$th sample quantile. In simulations the position estimator does better than the straight sample median, than a ratio-type estimator, and than an estimate based on $\hat{F}_{\text{KM},Q(1/2)}$. Note that it makes no use of selection probabilities. It would be of interest to compare it to the Hájek estimator $\hat{Q}_\pi(1/2)$.

Meeden (1995) uses an urn model simulating a posterior distribution of nonsample y_i that rests on the idea of exchangeability of the ratios $r_i = y_i/x_i$, as would hold under the model

$$y_i = \beta x_i + x_i \varepsilon_i. \tag{22}$$

The method assumes availability of nonsample x_j, which we can assume set out in order $x_{n+1}, \ldots, x_N$. A sample unit i is selected at random, $r_{n+1}^* = r_i$ recorded, the unit is cloned, and both it and its clone returned to the urn. The process is repeated $N-n$ times to get $r_{n+j}^*, j = 1, \ldots, N-n$, (the eventual size of the urn is N) and nonsample values of y are taken as $y_{n+j}^* = r_{n+j}^* x_{n+j}$. These, combined with the known sample values gives one realization of a population of y compatible with the given data and assumed

structure. The whole process is repeated R times and the average of the medians of the R generated populations taken as the estimated Polya posterior median $\hat{Q}_{\text{pp}}(1/2)$. (Clearly, the same process supplies estimates of other quantiles as well, and these would fulfill the monotonicity property.) In simulation studies on three real and several artificial populations, $\hat{Q}_{\text{pp}}(1/2)$ appears to be robust to failure of the model (22) and to perform on a par with the inverse of the CD estimator.

5.3. *Hinge estimates of quantiles*

The following idea arises out of Mak and Kuk (1993). Suppose we have an estimator $\hat{F}_*$, typically based on the population values of an auxiliary x, and so presumed accurate, but complicated enough that we might prefer not to calculate it repeatedly on an extensive grid $\{t_v\}$ to get $\hat{F}_*^{-1}(p)$. Suppose available also a simple estimator $\hat{F}_w$, say $\hat{F}_\pi$. Then, as in the rationale for the position estimator $\hat{Q}_{\text{pos}}$, described earlier, it would be enough to determine q so that $\hat{F}_w(Q(p)) = q$, for then we can take $\hat{Q}_{\text{adj}}(p) = \hat{F}_w^{-1}(q)$.

The difference $\Delta \equiv q - p$ can be written $\Delta = \hat{F}_w(Q(p)) - F(Q(p))$, the difference between the value of the simple estimator and the actual population distribution evaluated at the population quantile. It suffices to estimate Δ. This is done by replacing $Q(p)$ by an initial estimate such as $\hat{Q}_w(p)$ and F by $\hat{F}_*$, to yield a one-step adjusted estimator

$$\begin{aligned}\hat{Q}_{\text{adj}}^{(1)}(p) &= \hat{F}_w^{-1}\Big(p + \hat{F}_w\Big(\hat{Q}_w(p)\Big) - \hat{F}_*\Big(\hat{Q}_w(p)\Big)\Big) \\ &= \hat{F}_w^{-1}\Big(2p - \hat{F}_*\Big(\hat{Q}_w(p)\Big)\Big) \qquad (23)\end{aligned}$$

This requires calculation of $\hat{F}_*$ at only one value of t. The process can be iterated, replacing $\hat{Q}_w(p)$ by $\hat{Q}_{\text{adj}}^{(1)}(p)$ on the right in expression (23); this will require a calculation of $\hat{F}_*$ at a second value; and further iteration is possible. Mak and Kuk focus on $\hat{F}_* = \hat{F}_{\text{RKM}}$, but there is no reason to limit the idea to this case and indeed in a slightly variant version of this estimator, they take $\hat{F}_* = \hat{F}_{\text{CD},\Phi}$ (see Section 3.6.1). It is neither clear what the impact is of the choice of $\hat{F}_*$ or of the degree of iteration, nor how close the result would be to $\hat{F}_*^{-1}(\alpha)$.

6. Variance estimation and confidence intervals for distribution functions

We shall not in this chapter pursue the possibilities for variance estimation in great detail. Each of the estimators described earlier will require some particular corresponding variance estimator, and it seems wisest to refer the reader to the originating papers. Instead, we shall make some general remarks, focus on the CD and RKM estimators, and point to one or two papers whose primary concern is variance estimation. In general, there are three basic approaches: (1) plug-in model based, estimating asymptotic variances, (2) design-based, and (3) replication methods, which are widely regarded as falling into both the model-based and design-based camps.

All three are considered in Wu and Sitter (2001b) (henceforth WS). The model assumption is the simple linear model $y_i = \alpha + \beta x_i + \varepsilon_i$, and they get an estimate $v_m = \hat{\text{var}}\big(\hat{F}_{\text{CD}}(t) - F(t)\big)$ by plugging in sample estimates of the I terms in (14), replacing the separate expressions I_2 and I_3 by an expression for $I_{23} \equiv I_2 - I_3^2$ and estimating that directly, for the sake of stability and positivity.

WS suggest several jackknife estimates for the CD estimator, but best appears to be a jackknife hybrid estimator which estimates the (less important) $\mathrm{var}(F(t))$ component by a simple plug-in estimate of I_4, and $\mathrm{var}\big(\hat{F}_{\mathrm{CD}}(t)\big)$ by the jackknife $v_{J1}(t) = \frac{n-1}{n}\sum_{i\in s}\big(\hat{F}_{\mathrm{CD},-i}(t) - \bar{F}(t)\big)^2$, where $\hat{F}_{\mathrm{CD},-i}(t)$ is $\hat{F}_{\mathrm{CD}}(t)$ computed without making any use of the ith sample point, and $\bar{F}(t) = n^{-1}\sum_i \hat{F}_{\mathrm{CD},-i}(t)$ (note that $\hat{F}_{\mathrm{CD}}(t)$ itself would probably do as well.) We note that leave one out estimates of the necessary parameters and residuals are readily available without needing complete recalculation (Miller, 1974).

For the design-based RKM estimator WS take RKM design-based variance estimator $v_d = \hat{\mathrm{var}}\big(\hat{F}_{\mathrm{RKM}}\big) = N^{-2}\sum_{i\in s}\sum_{j>i}\frac{\pi_i\pi_j-\pi_{ij}}{\pi_{ij}}\big\{\frac{u_i(j)}{\pi_i} - \frac{u_j(i)}{\pi_j}\big\}$, where the residuals $u_i(j) = I\big(y_i \le t\big) - \hat{G}_{ic}(j)$, with $\hat{G}_{ic}(j) = \big[\sum_{k\in s}\frac{\pi_{ij}}{\pi_{ijk}}I\{y_k \le t - \hat{\beta}_\pi\,(x_i - x_k)\}\big]\big/\sum_{k\in s}\frac{\pi_{ij}}{\pi_{ijk}}$. Typically, simplifying assumptions will be necessary to calculate the higher order inclusion probabilities.

WS give two variants of a jackknife estimator of $\mathrm{var}\big(\hat{F}_{\mathrm{RKM}}(t)\big)$: v_{Jd2}, based on $\hat{F}_{\mathrm{RKM},-i}(t)$, as $v_{J1}(t)$ was based on $\hat{F}_{\mathrm{CD},-i}(t)$ and v_{Jd1}, slightly easier to calculate, based on the full sample $\hat{G}_\pi\Big(\frac{t-x_k^T\hat{\beta}_\pi}{v_k^{1/2}}\Big)$ and $\hat{G}_{\pi C}\Big(\frac{t-x_{i'}^T\hat{\beta}_\pi}{v_{i'}^{1/2}}\Big)$ of (10), and using deleted versions only of the weights.

In testing on artificial data, the plug-in estimator v_m for CD was consistently less variable than the hybrid jackknife (INST, WS, Table 1), and tended to be biased lower (RB, WS, Table 1). This could lead, we suspect, to somewhat closer to nominal coverage for the jackknife. For RKM, v_{Jd1} and the design-based estimator v_d were indistinguishable; v_{Jd2} was consistently less variable than they were, and tended to be biased higher (WS, Table 2). This should lead to better coverage for v_{Jd2}. It would be of interest to assess coverage properties. WS avoided considering a model-based plug-in estimate for $\mathrm{var}\big(\hat{F}_{\mathrm{RKM}}(t) - F(t)\big)$, but it would be interesting to see how the one given in Wang and Dorfman (1996) compares empirically to v_d and v_{Jd2}.

Lombardía et al. (2003) suggest a bootstrap methodology for generating bias and variance estimates and confidence intervals for the CD estimator. The method involves Monte Carlo generation of B bootstrap populations based on the fitted model and residual distribution from the sample, and R samples from each population, and assumes homoscedasticity and that the working model is correct. It is not clear how robust the methodology is to deviations from those conditions.

Lombardía et al. (2004) use a similar bootstrap methodology for generating estimates of bias, variance, and mean square error, for the nonparametric CD estimator $\hat{F}_{\mathrm{np,CD}}$, as ingredients in a comprehensive bandwidth selection procedure (Section 3.7.3). Again, homoscedasticity is assumed. It would be of interest to see how their methodology can be adapted to other estimators, like the Kuo. It is very computer intensive.

Empirical likelihood confidence intervals for the distribution function are discussed in Chapter 30.

7. Confidence intervals and variance estimates for quantiles

Woodruff (1952) describes a straightforward method for transforming estimates of precision for a cdf to estimates of precision for the corresponding quantiles. Assume a

cdf estimator $\hat{F}(t)$ is itself a cdf and that $(1-\alpha)100\%$ two-sided confidence intervals $\big(\hat{F}_L(t), \hat{F}_U(t)\big)$, based say on the normal assumption and an estimate of $\text{var}\big(\hat{F}(t) - F(t)\big)$. Suppose we want a $(1-\alpha)100\%$ two-sided confidence interval for $t_p = Q(p)$, which we have estimated by $\hat{t}_p = \hat{F}^{-1}(p)$. Let $t_L = \hat{F}^{-1}\big(\hat{F}_L(\hat{t}_p)\big)$ and $t_U = \hat{F}^{-1}\big(\hat{F}_U(\hat{t}_p)\big)$. Then Woodruff argues that $[t_L, t_U]$ is an approximate $(1-\alpha)100\%$ confidence interval for $Q(p)$. It will not in general be symmetric about $\hat{Q}(p)$.

Francisco and Fuller (1991) (henceforth FF) offered an alternative construction, taking $t_L = \hat{F}_U^{-1}\big(\hat{F}(\hat{t}_p)\big)$ and $t_U = \hat{F}_L^{-1}\big(\hat{F}(\hat{t}_p)\big)$. The resulting interval is basically a standard calibration confidence interval in the sense of Carroll and Ruppert (1988), Section 2.9.3. It is more complicated to calculate than the Woodruff, because it requires values of the upper and lower bounds over a range of values of t near t_p, as well as various transformations and smoothing operations.

The Woodruff has held up well in empirical studies, and there is little evidence of an advantage to the FF. Sitter and Wu (2001) note that, especially at low or high p, the actual coverage of the Woodruff interval for $Q(p)$ can be better than that for the corresponding interval for $F[Q(p)]$. (Consider small p; by the nature of a binomial distribution $\hat{p} = \hat{F}(Q(p))$ is more likely than not to lie below p and because the variance estimator of a binomial is monotonically increasing for small p, the variance estimate associated with $\hat{p}$ will be lower than it would have been had $\hat{p}$ equaled p. This can lead to low coverage of the corresponding confidence interval for the cdf. On the other hand, on the same premise, working backwards, one can anticipate that $\hat{Q}(p)$ is likely to lie above $Q(p)$, with $\hat{F}\big(\hat{Q}(p)\big) \geq \hat{F}(Q(p))$ and the corresponding variance estimate of the cdf, from which the Woodruff is generated, is larger than what it would have been if $\hat{Q}(p) = Q(p)$.)

For low or high p, Shah and Vaish (2006) make adjustments to $\hat{F}(t)$ and adopt modifications of the variance estimate, due to Korn and Graubard (1998c), prior to applying the Woodruff.

Dorfman and Valliant (1993) (DV), in the context of quantile estimation for domains, some of them quite sparse, compare Woodruff and FF confidence intervals, as well as confidence intervals derived from balanced repeated replication (BRR) variance estimation applied directly to the quantile estimates. In terms both of bias for the actual mean square error and coverage, BRR was better than Woodruff, which was better than FF.

Making use of the well known Bahadur (1966) representation $\hat{Q}_n(p) - Q(p) \approx F(Q(p)) - \hat{F}(Q(p))/f(Q(p))$, DV also base a variance estimator on the variance estimate for the cdf: $\hat{\text{var}}\{\hat{Q}_n(p) - Q(p)\} \approx \hat{\text{var}}\big\{\hat{F}\big(\hat{Q}(p)\big) - F\big(\hat{Q}(p)\big)\big\}/\hat{f}\big(\hat{Q}(p)\big)^2$, where $\hat{f}\big(\hat{Q}(p)\big)$ is estimated using nonparametric density estimation. The overall behavior of the resulting estimator is somewhere between BRR and Woodruff. Chen and Wu (2002) have further discussion of the Bahadur representation. Wheeless and Shah (1988) give an alternate approach to estimating the density f.

Using Woodruff $1-\alpha$ confidence intervals, Rao et al. (1990) suggest variance estimates $\big(L_\alpha/2z_{\alpha/2}\big)^2$ where L_α is the length of the interval and $z_{\alpha/2}$ is the $\alpha/2$ percentile of the standard normal distribution. Variance estimates for ratios, etc. can be built up from these. DV offer a few modifications of this.

8. Further results and questions

The CD and other estimators we have discussed assume independence of errors. This is not always a safe assumption, particularly in cluster sampling. Mayor (2002) has made a suggestion on estimation of the cdf for cluster sampling, using a Horvitz Thompson type estimator, which, however, violates Property 1a. It is not clear how in general to do cdf estimation in the presence of correlated errors.

Singh et al. (2001) discuss generic methods for estimating quantiles based on double sampling. It might be expected that most of the methods of this chapter can be extended to the double sampling situation, with the larger sample replacing the population in the various formulas.

Ranked set sampling is an approach to double sampling designed to give an even spread of the y among the sampled x. Lam et al. (2002) give an estimator of the cdf appropriate to ranked set sampling patterned after the kernel method of Kuk (1993).

Stefanski and Bay (1996) consider the interesting situation of estimating the cdf when there is measurement error in the variable of interest.

It would be interesting to combine the nonparametric regression estimation of variance structure (Section 3.7.1: Lombardía et al. (2005)) with nonparametric estimation of the mean for a completely parameter free estimator of the cdf. It would be of interest to compare the resulting estimator to the CD estimator with mean and variance structure chosen by regression diagnostics.

There has been no comprehensive comparison of the many available alternatives for cdf and quantile estimation. Such a study, breaking up estimators into comparison groups based on auxiliary information used, carried out on the many real and artificial populations found in the cdf literature, and tracking relative efficiencies and computing times, would be very useful.

Sample Surveys: Inference and Analysis, Vol. 29B
ISSN: 0169-7161
© 2009 Published by Elsevier B.V.
DOI: 10.1016/S0169-7161(09)00237-5

Scatterplots with Survey Data

Barry I. Graubard and Edward L. Korn

1. Introduction

The scatterplot is one of the most useful graphical displays of bivariate data. It allows one to see general trends and atypical points simultaneously, as well as other aspects of the data. Data collected in a survey, however, have some additional features that can make a simple scatterplot less useful. One such feature is that individuals in the sample represent different numbers of individuals in the population. The sample weights of the sampled individuals effectively estimate these numbers. A second feature of survey data is that some of it may be imputed to account for item nonresponse. A third feature is that the sample sizes can be large. A fourth feature is that the observations may have intraclass correlation due to cluster sampling. As will be shown below, standard scatterplots that are used for simple random samples that ignore these features can be misleading or hard to interpret. We know of no "super plot" that will be as successful in the survey setting as the simple scatterplot is in the nonsurvey setting. Instead, we present in this chapter different modifications of the scatterplot, demonstrated by examples, that can improve the presentation of survey data. By and large, these modified plots are not new, but their application to survey data may not be well known. There has been little new literature on scatterplots with complex survey data since Korn and Graubard (1998) and Korn and Graubard (1999). Most of this chapter is taken from those two sources with some minor updates.

2. Modifications of scatterplots for survey data

In this section, we present some techniques that can be used to modify a scatterplot to incorporate various aspects of survey data. First, we describe the use of bubble plots in which the sizes of the plotted circles are proportional to the sample weights of the points. Examples are given showing that such bubble plots can perform better than a simple scatterplot in (a) describing the population distribution and (b) identifying influential points in a weighted analysis (which is typically used when analyzing survey data). However, for moderate-to-large sample sizes, a bubble plot can be hard to interpret

because of the overlapping bubbles. For this situation, we consider in Section 2.2 using a "sampled scatterplot," in which the sampled data is resampled proportionally to the sample weights, yielding a data set that can be plotted without circles but still represents the population distribution.

Plots of large data sets can be problematic because of overlapping plotted points. This can especially be a problem when the raw data has been implicitly or explicitly rounded. An example is given in Section 2.3, along with the possible solution of "jittering" the data, that is, adding a small amount of random noise to the data before plotting. In Section 2.4, we discuss scatterplots in which some of the plotted points represent imputed data values to account for item nonresponse. In Section 2.5, we consider using conditional mean and percentile curves constructed using kernel smoothing for nonparametrically displaying the relationship between Y and X when the sample sizes are large. Finally, in Section 2.6, a modeling alternative approach is considered that uses regression splines to investigate relationships between Y and X.

2.1. Accounting for the sample weights: bubble plots

Survey designs typically specify that individuals are to be sampled with unequal probabilities of selection. The sample weight associated with an individual is the inverse of that individual's probability of being included in the sample, adjusted, if necessary, for nonresponse. There is often an additional poststratification to ensure that the sum of the sample weights equals known population values for various subgroups (e.g., age/race/sex subgroups). The sample weights effectively represent the number of individuals in the population that the sampled individual represents.

Figure 1 is a scatterplot of daughter's birthweight versus mother's birthweight for mothers aged 30–39 years at the time of birth; the data are from the 1988 National Maternal and Infant Health Survey which sampled vital records corresponding to live births, late fetal deaths, and infant deaths in the United States (Sanderson et al., 1991). For the live birth component of the survey, mothers corresponding to sampled birth certificates were mailed a questionnaire. The birthweight of the child was taken from the birth certificate (reported in grams) and the birthweight of the mother was taken from the mother's questionnaire (reported in ounces, converted to grams for the plot). Relationships between the birthweights of mothers and their children have been studied previously using data from this survey (Wang et al., 1995). We restrict attention to first births that were daughters, and mother-daughter pairs with nonmissing birthweights ($n = 225$). Figure 1 is a misleading representation of the population because it ignores the sample weights; this survey oversampled low birthweight babies and black babies (Table 1) (Nonresponse and poststratification adjustments to the sample weights were relatively small.). One possibility to more accurately reflect the population is displayed by the bubble plot in Fig. 2; the areas of the circles are proportional to the sample weights.

Another use for using the size of bubbles to designate sample weights is to help identify influential points in an analysis. We now give an example using an analysis of the association of developing cancer with baseline transferrin saturation values based on women participating in the epidemiologic follow-up of the first National Health and Nutrition Examination Survey (National Center for Health Statistics et al., 1987). This

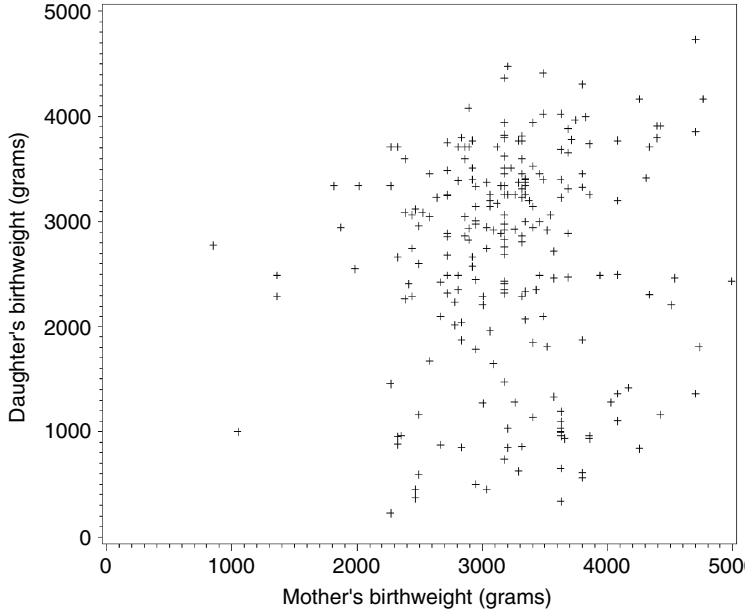


Fig. 1. Simple scatterplot based on data from mothers aged 30–39 surveyed in the 1988 National Maternal and Infant Health Survey.

Table 1
Sampling strata and sampling rates of 1988 National Maternal and Infant Health Survey

Strata		Sampling Rate
Race	Birth Weight (grams)	
Black	<1500	1/14
	1500–2499	1/55
	≥2500	1/113
Nonblack	<1500	1/29
	1500–2499	1/160
	≥2500	1/720

association has been previously studied by us Korn and Graubard (1995) and others (Stevens et al., 1988). We follow the previous analyses and remove women from the analysis who had cancer at the baseline or who developed it within four years of the baseline survey; this leaves 197 women who developed cancer and 5073 who did not. The sample weights ranged from 611 to 186,062 (coefficient of variation = 97%), with the distribution being similar for the women who developed cancer and for those who did not. We consider a logistic regression of the probability of developing cancer on transferrin saturation and other covariates described in footnote 1 of Table 2. A classical survey analysis uses weighted estimators; the weighted logistic regression coefficient for transferrin saturation is given in the first line of Table 2.

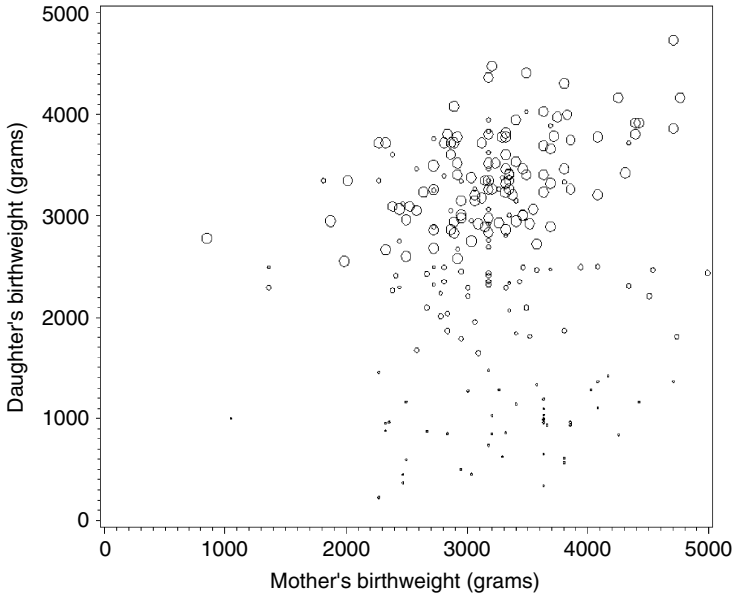


Fig. 2. Bubble plot of data plotted in Fig. 1; areas of circles are proportional to the sample weights.

Table 2
Weighted logistic regression coefficient (±standard error) for transferrin saturation from a multiple logistic regression of the probability of developing cancer on transferrin saturation and other covariates[a], dropping certain data points

Point[b] Dropped from the Analysis	Sample Size	$\beta \pm$ SE[c]
None	5270	$.025 \pm .014$
Point A	5269	$.009 \pm .009$
Point B	5269	$.024 \pm .014$
Point C	5269	$.028 \pm .014$

[a]Covariates included in the model are age at the baseline examination; smoking (never smoked, former smoker, current smoker, and unknown); race (white and nonwhite); senior status (age ≥65 and age <65 years); living in poverty census Enumeration District (yes, no); and family income (<$3000, $3000–6999, $7000–9999, $10,000–14,999, and ≥$15,000).
[b]Points are designated in Fig. 3.
[c]To account for the complex sampling design, the computer program SUDAAN (Shah et al., 1995) was used to calculate the standard errors.

An added variable plot, also known as a partial regression leverage plot, is useful for identifying influential points in a multiple linear regression of Y on X and Z (Cook and Weisberg, 1994, Chapter 12.1; Atkinson, 1985, Chapter 5.2-3). It is a plot of the residuals from the regression of the dependent variable Y on the covariate vector Z (which includes the intercept) versus the residuals from the regression of the independent variable currently under study (X) on Z. The slope of the least-squares line based on this plot is the same as the regression coefficient for X in the multiple linear regression. For a multiple *logistic* regression of a binary Y on X and Z, O'Hara Hines and Carter (1993) suggest calculating the residuals from the linear regression of $\sqrt{p(1-p)}\left[\log\frac{p}{1-p} + \frac{Y-p}{p(1-p)}\right]$ on

$\sqrt{p(1-p)}X$ and $\sqrt{p(1-p)}Z$ and plotting these residuals against the residuals from the linear regression of $\sqrt{p(1-p)}X$ on $\sqrt{p(1-p)}Z$, where p is the predicted probability that $Y = 1$ based on the multiple logistic regression. The slope of the least-squares line through this plot will equal the logistic regression coefficient of X from the multiple regression.

In our application, a *weighted* multiple logistic regression is used since the observations have sample weights. To account for this in the added variable plot, the linear regressions used to obtain the residuals above need to be weighted linear regressions, and the predicted values p need to be obtained from the weighted logistic regression. With these modifications, the slope from a weighted least-squares regression through the added variable plot will equal the regression coefficient of X from the weighted logistic regression of Y on X and Z.

Figure 3 is the added variable plot for transferrin saturation; the areas of the circles are proportional to the sample weights. The dashed line in Fig. 3 is the weighted least-squares line; its slope is .025, the same at the logistic regression coefficient for transferrin saturation (Table 2). The mass of plotted points on the bottom left of the plot is not aesthetically pleasing, but for the purpose of identifying influential points is not troublesome. The point labeled A would appear to be highly influential. This is confirmed by noting that when this point is dropped from the analysis, the logistic regression coefficient for transferrin saturation changes from .025 to .009 (Table 2). This point is also highly influential for estimating the standard error of the coefficient; it changes from .014 to .009 with removal of the point.

A simple scatterplot without the circles would not be as successful as Fig. 3 in identifying influential points. For example, without the circles, the point labeled B

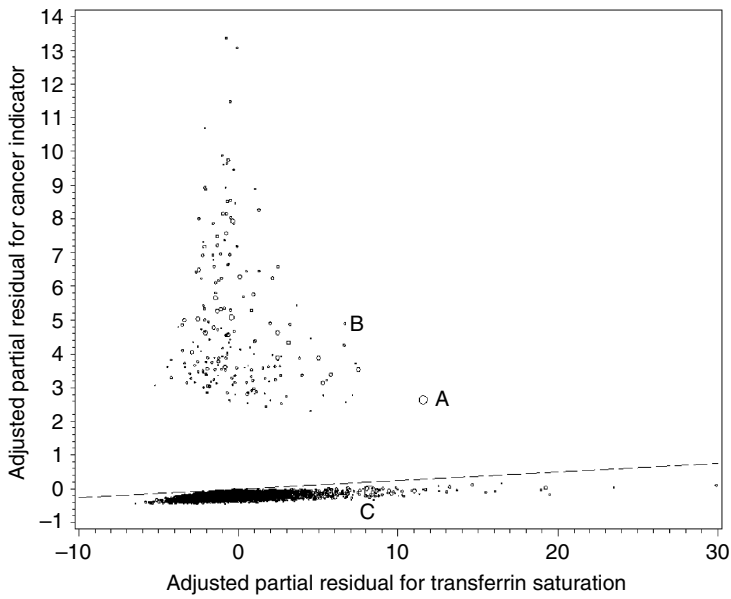


Fig. 3. Added variable plot for transferrin saturation based on weighted multiple logistic regression described in Table 2. Dashed line is weighted least-squares line; labeled points are described in the text.

might appear about as influential as point A. However, because of its small sample weight, it has very little influence on the coefficient (Table 2). On the other hand, it is not sufficient to ignore the plot and assume that observations with large sample weights will be influential. For example, the observation above the label *C* in Fig. 3 has a larger sample weight than point *A*. From its plotted position, however, we would not expect it to be influential and it is not (Table 2).

2.2. *Accounting for the sample weights: sampled scatterplots*

An alternative strategy to using a bubble plot is to use a "sampled scatterplot." The idea is to sample the data with probabilities proportional to the sample weights; the resulting sampled data is then approximately representative of the population and can be plotted ignoring the sample weights. Figure 4 ($n = 100$) is a sampled scatterplot of the data displayed in Fig. 2. The ith observation from the original data set was included in Fig. 2 if a uniform (pseudo-)random number was less than $w_i/w_{\max}$, where w_i is the sample weight of the ith observation and $w_{\max}(= 1008.515)$ is the largest sample weight of the 225 observations in Fig. 2. In general, one samples the ith data point to be plotted an expected number of times equal to $w_i/(cw_{max})$, where c is chosen to control the expected sample size of the plot. In some cases, the same observations may be sampled multiple times resulting in overlapping points on a plot. In these cases, one might consider jittering the data in the plot to separate the overlapping points as described in Section 2.3. The idea of resampling survey data to eliminate the effects of the sample weights in further analysis has been used by Murthy and Sethi (1965) and Hinkins et al. (1994) to use conventional nonsurvey methods of analysis for survey data.

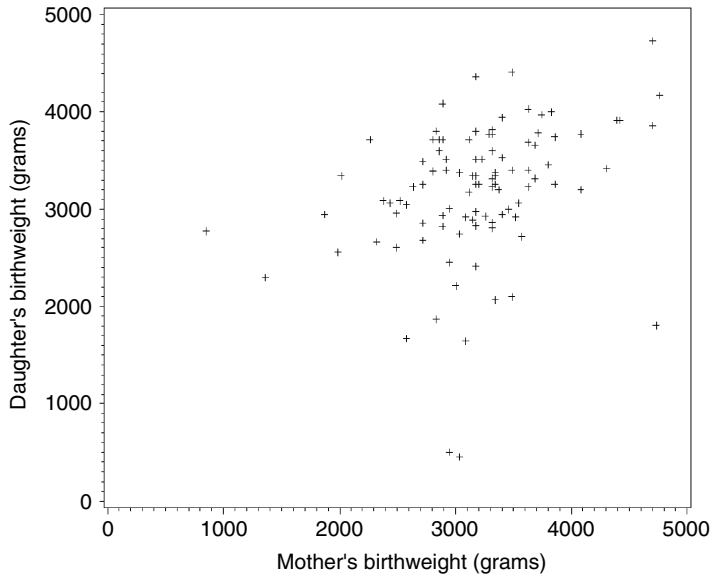


Fig. 4. Sampled scatterplot of data plotted in Fig. 2. Points were chosen for plotting with probability proportional to their sample weights.

There is no question that there is a loss of information in going from Fig. 2 to Fig. 4. Therefore, Fig. 2 would be the preferred plot for data cleaning. Additionally, weighted estimation using the full data set should be used for estimating population parameters. However, as a visual display of the population, we prefer Fig. 4 to Fig. 2, and this preference would become stronger if the sample size were larger, see the height/age example given below.

For some applications, it may be useful to sample points for a sampled scatterplot not just proportionally to the sample weights. For example, suppose we are interested in the relationship of mother's and daughter's birthweights for black and nonblack daughters. Only four of the data points in Fig. 4 correspond to black daughters, and this is reflective of the population. Since black babies were oversampled in the survey, there is much more information available. Figure 5 is a sampled scatterplot in which data points corresponding to black daughters were sampled with probability $w_i/166.642$ (166.642 is the largest sample weight corresponding to a black baby in the original data), whereas data points corresponding to nonblack daughters were sample with probability $w_i/1008.515$. Therefore, although Fig. 5 is not representative of the population, it is representative of the black and nonblack populations separately. It appears from Fig. 5 that there is a stronger positive correlation among the nonblack mother-daughter pairs than among the black mother-daughter pairs. This can also be demonstrated numerically by comparing the weighted correlations using all the sampled data for the nonblack and black pairs: $0.32(n = 170)$ versus $0.07(n = 55)$, respectively.

Figure 5 also displays an additional characteristic of the data that may not have been apparent before—there are many observations with mother's birthweight equal to 3175.133 grams, converted from 7 pounds, 0 ounces. A better representation of the

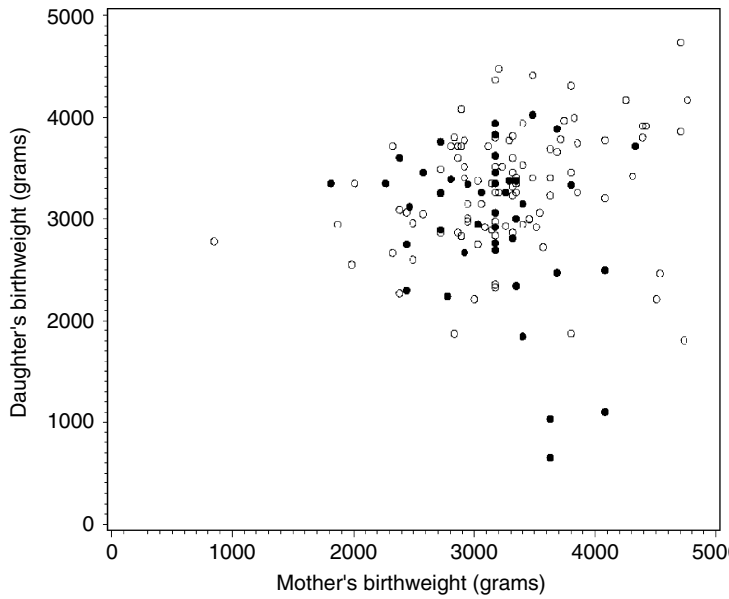


Fig. 5. Sampled scatterplot of data plotted in Fig. 2. Black daughters (filled-in circles) were sampled for plotting at approximately six times the rate as nonblack daughters (open circles).

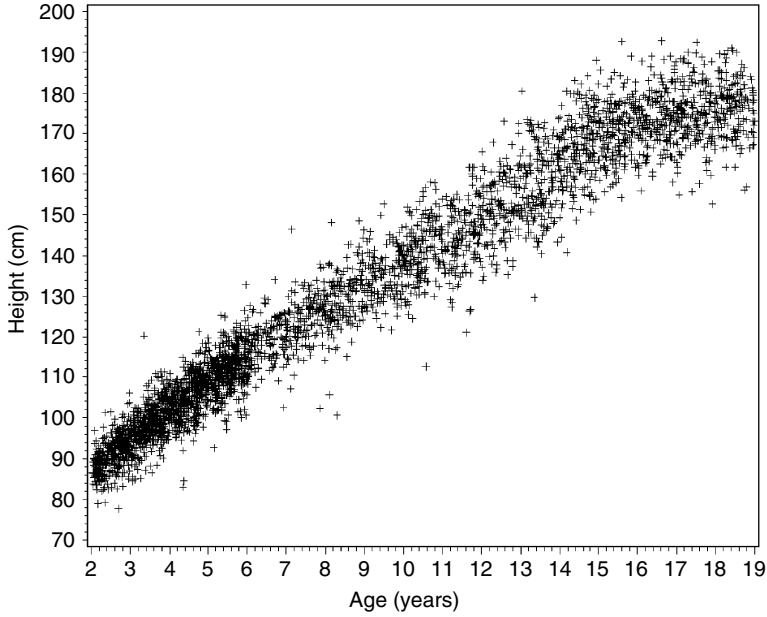


Fig. 6. Simple scatterplot of height versus age for boy aged less than 19 years sampled in the second National Health and Nutrition Examination Survey.

population might be obtained by randomly jittering the data to account for the rounding in the reporting (see Section 2.3 below).

Another application of the sampled scatterplot is when the sample size is large. Figure 6 is a simple scatterplot of height versus age for the 3667 boys aged 2–19 years sampled in the second National Health and Nutrition Examination Survey. The sample weights for these boys ranged from 1359 to 47,385, with a coefficient of variation of 71%, see McDowell et al. (1981) for full details of this survey. Besides being an unappealing plot because of the mass of points being plotted, the plot is also not representative of the population because of the differing sample weights. In particular, boys aged five years or younger were sampled in this survey at three times the rate of boys six years or older. This is reflected in Fig. 6 in the increased density of plotted points for age less than six. Because of the large number of plotted points, a bubble plot version of Fig. 6 would not be useful. We can solve the two problems of excessive density and representativeness at once by using a sampled scatterplot, see Fig. 7 in which $n = 699$ points are plotted.

2.3. Accounting for overlap and rounding: jittering

In plotting a small number of observations, occasionally multiple observations will have values so close (or identical) as to make their plotted points indistinguishable. The easy solution to this problem is to displace by a small amount such points. With larger data sets, the problem can become more acute. For example, Fig. 8 is a bubble plot of systolic blood pressure versus the logarithm of blood lead values for 595 white

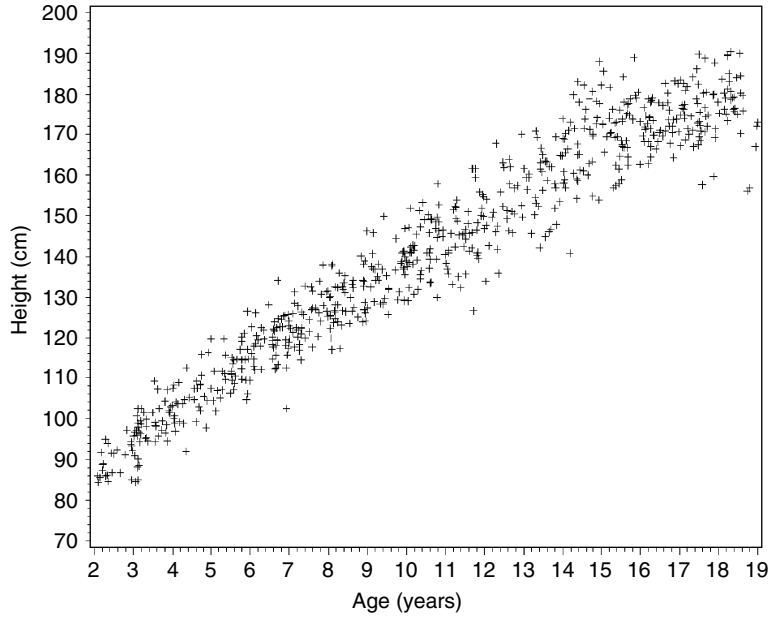


Fig. 7. Sampled scatterplot of data plotted in Fig. 6. Points were chosen for plotting with probability proportional to their sample weights.

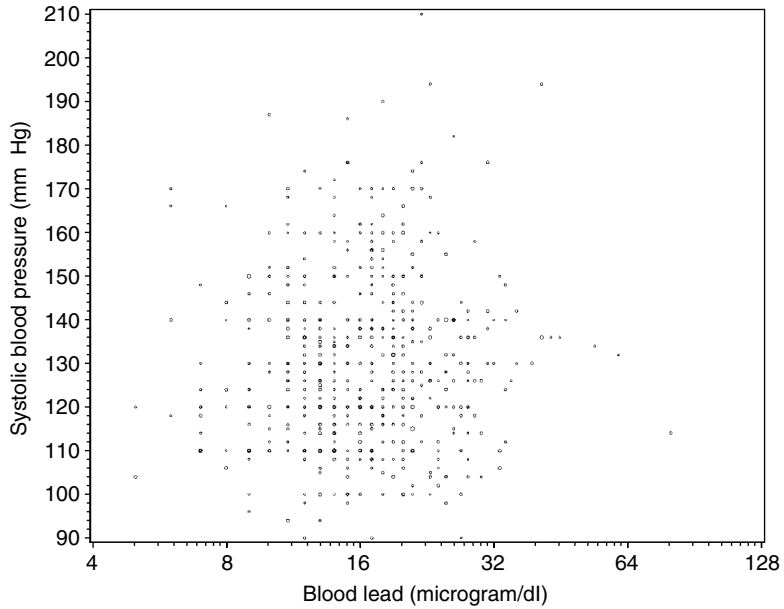


Fig. 8. Bubble plot based on data from white males aged 40–59 years sampled in the second National Health and Nutrition Examination Survey; areas of circles are proportional to the sample weights. There are many overlapping circles in this plot.

males aged 40–59 years. The data are from the second National Health and Nutrition Examination Survey, with the areas of the bubble being proportional to the sample weights (range = 11601–79176, coefficient of variation = 41%). The relationship of blood pressure and lead levels has been previously studied using these data by Pirkle et al. (1985). The lattice pattern of Fig. 8 is because blood pressure was recorded to the nearest mm Hg and blood lead values were recorded to the nearest microgram/deciliter. The overlap of the circles gives a misleading impression of the distribution of values. With this type of "rounding" of the data, a natural solution to the problem of overlapping points is to jitter the data (Chambers et al., 1983, pp. 106–107). In this case, random uniform $(-1/2, +1/2)$ variates are added to the blood pressure and lead values before plotting because it is reasonable to treat the observed values as if they had been rounded to the nearest integer from true values. The jittered plot displayed in Fig. 9 not only avoids the overlap of plotted points but also gives a better representation of the prerounded blood lead levels.

An alternative solution to the overlap problem is to sum the sample weights for points that are plotted at the same location. Figure 10 is the bubble plot using these summed sample weights. This approach has been suggested in the nonsurvey setting, in which "sunflowers" (with the number of lines in the sunflowers equal to the number of data points at the location) are used instead of bubbles (Cleveland and McGill, 1984). Additionally, continuous data can be artificially rounded to apply this approach (Cleveland and McGill, 1984). In the survey setting, this approach is less attractive than jittering because one cannot distinguish in the plot single individuals with large sample weights versus many individuals with small sample weights plotted at the same location.

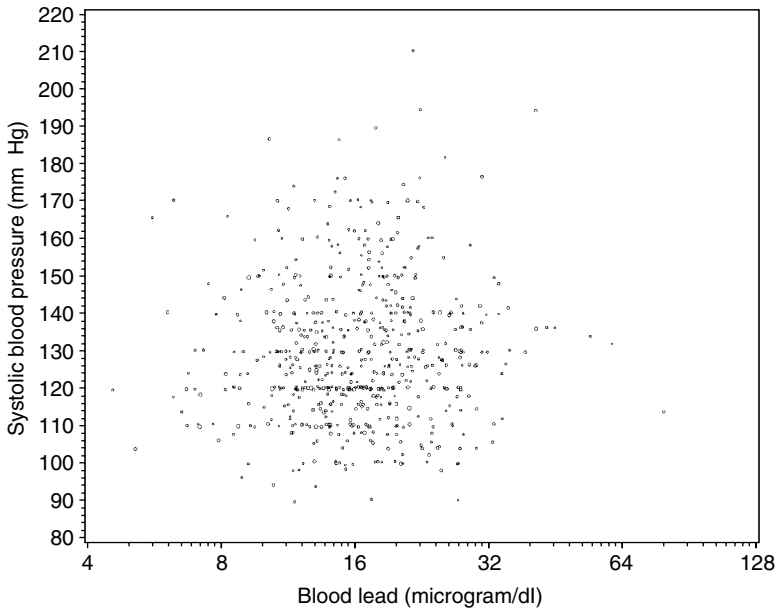


Fig. 9. Jittered bubble plot of data plotted in Fig. 8.

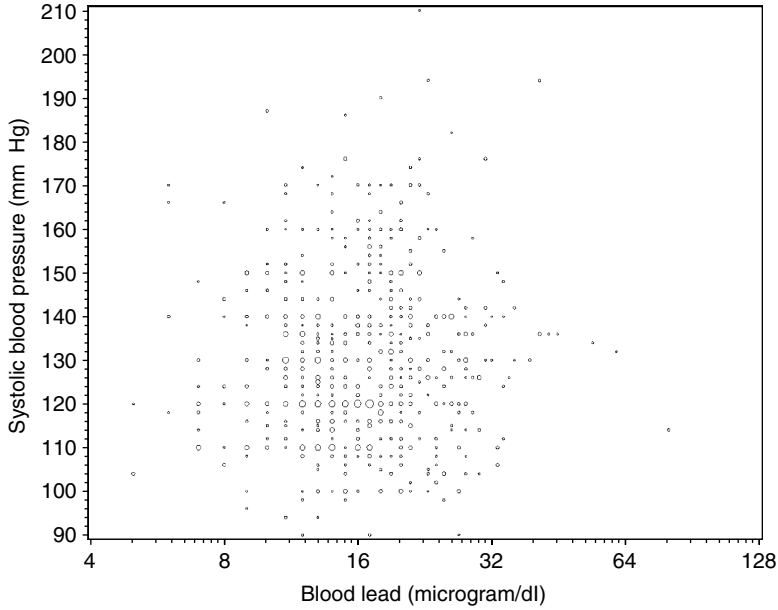


Fig. 10. Summed bubble plot of data plotted in Fig. 8. Areas of circles are proportional to the sum of the sample weights of the individuals with the same data values to be plotted.

2.4. Accounting for missing data: imputation

Although missing data can be a problem in any data analysis, survey data are especially susceptible because of the possibility of nonresponse. Data can be missing completely from a sampled individual (unit nonresponse) or partially missing because some questions remain unanswered (item nonresponse). A nonresponse adjustment to the sample weights is frequently used for unit nonresponse; the sample weights are adjusted upwards for respondents with values of other variables similar to those of nonrespondents. The sample weights can be accounted for in a scatterplot as described in Sections 2.1 and 2.2. Item nonresponse is sometimes handled by imputing values for the missing values. There are many ways to do this (Little and Rubin, 2002, Chapters 4 and 5), one of which is described below.

As a preliminary, it can be useful to plot the data without any imputations. Returning to the mother-daughter birthweight data (Fig. 2), the full sample size is 286 of which 225 observations have both mother's and daughter's birthweight nonmissing. Sixty observations are solely missing mother's birthweight and one observation is solely missing daughter's birthweight. Figure 11 displays the sampled scatterplot of Fig. 4, but now also contains (modified) box plots for the estimated distributions of daughter's birthweight for observations not missing, and missing, mother's birthweight (For plotting, the single observation missing daughter's birthweight is ignored.). For these box plots, the edges of the boxes represent the 25th and 75th percentiles, the line in the box represents the median, and the lines extending from the box represent the 10th and 90th percentiles. These percentiles are estimated from using weighted percentiles of the complete samples and not just the (re)sampled observations displayed on the left-hand side of Fig. 11. The

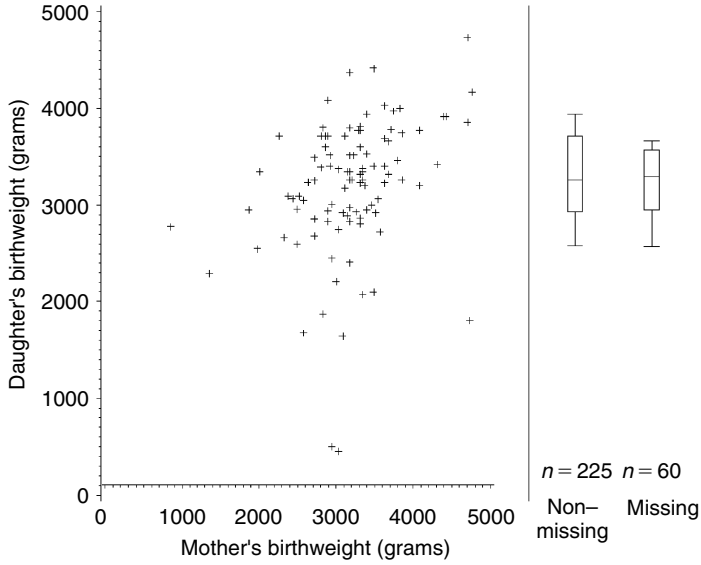


Fig. 11. Sampled scatterplot of nonmissing data with weighted box plots of nonmissing and missing data. Data are from mothers aged 30–39 surveyed in the 1988 National Maternal and Infant Health Survey.

box plots suggest that missingness of mother's birthweight may be less prevalent for high birthweight daughters, but the two-sided p-value for comparing the means is 0.18. An alternative to using the box plots in Fig. 11 would be to display weighted histograms of the distributions. As mentioned above, there are many ways for imputing values for missing data. For graphical displays, it is important that the variability of the imputed values should be consistent with the population variability. We will demonstrate the point with the mother-daughter birthweight data (no imputed values were supplied by the National Center for Health Statistics for mother's birthweight). We use the regression model

$$\begin{aligned}\text{mother's birthweight} = \alpha &+ \beta_{\text{M-HT}} X_{\text{M-HT}} + \beta_{\text{M-RACE}} X_{\text{M-RACE}} \\ &+ \beta_{\text{D-BW}} X_{\text{D-BW}} + \text{ error}, \end{aligned} \tag{1}$$

where $X_{\text{M-HT}}$ and $X_{\text{M-RACE}}$ denote mother's height and race (1 = nonblack, 2 = black) and $X_{\text{D-BW}}$ denotes the daughter's birthweight. The regression coefficients in model (1) are estimated using (sample-)weighted least-squares for those observations with nonmissing mother's birthweight (the one observation of missing daughter's birthweight was assigned the mean daughter's birthweight). The fitted regression was

$$\begin{aligned}\text{predicted mother's birthweight} = &- 202 + 37.3 X_{\text{M-HT}} + 123 X_{\text{M-RACE}} \\ &+ 0.270 X_{\text{D-BW}}. \end{aligned} \tag{2}$$

To impute a mother's missing birthweight, we substitute the mother's height and race and her daughter's birthweight into (2) to obtain the predicted mother's birthweight, and then add on an error term obtained as follows. The error terms for the imputed values were obtained by sampling the residuals from the fitted model (2) using probability

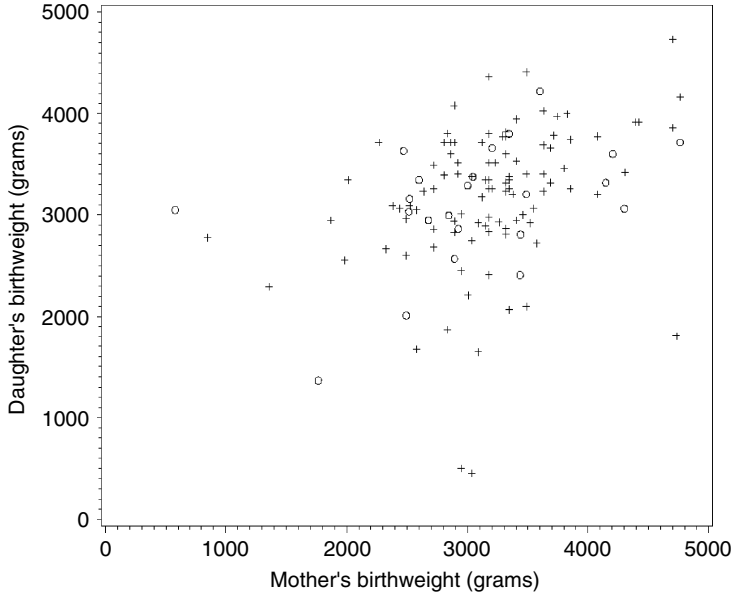


Fig. 12. Sampled scatterplot based on data from mothers aged 30–39 surveyed in the 1988 National Maternal and Infant Health Survey (circles = imputed values, + = nonimputed values).

proportional-to-size sampling, where the inclusion probabilities were proportional to the sampling weights. Figure 12 is a sampled scatterplot of the mother-daughter pairs in which the pairs with imputed mothers' birthweights are designated by o and the nonimputed values by +. If one used for the imputed values the predicted mothers' birthweights from (2) without adding the error term, the sampled scatterplot would be Fig. 13. The spread of the imputed values is misleadingly small in Fig. 13, demonstrating the importance of including an error term in the imputed values.

In both Figs. 12 and 13, the imputed values were highlighted by using a dramatically different symbol in the plots. For many applications, we may want the distinction between imputed and nonimputed values to be visible but not to overpower the display. This can be accomplished by using different symbols that are somewhat similar. For example, one could use x instead of o to denote the imputed values in Fig. 12.

2.5. Conditional mean and percentile curves: kernel smoothing

Although one might typically use a polynomial regression to display the X-Y relationship on a scatterplot of a small-to-moderate number of observations, the large number of observations sometimes available with survey data allows for the consideration of less model-dependent approaches. As a simple example, Fig. 14 is a strip box plot (Chambers et al., 1983, pp. 87–91) of height as a function of age for boys aged 2–19 years sampled in the second National Health and Nutrition Examination Survey, see Fig. 7 for a sampled scatterplot of this data. Each box plot displays the sample-weighted 10th, 25th, 50th, 75th, and 90th percentiles of height of those individuals at a particular year of age at the time of examination. The number of observations included for each year of age

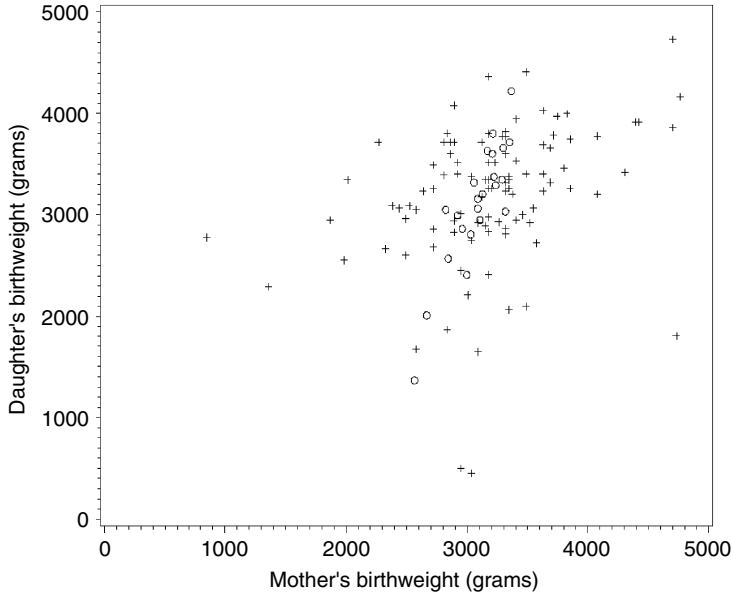


Fig. 13. Sampled scatterplot based on data from mothers aged 30–39 surveyed in the 1988 National Maternal and Infant Health Survey (circles = imputed values without error included, + = nonimputed values).

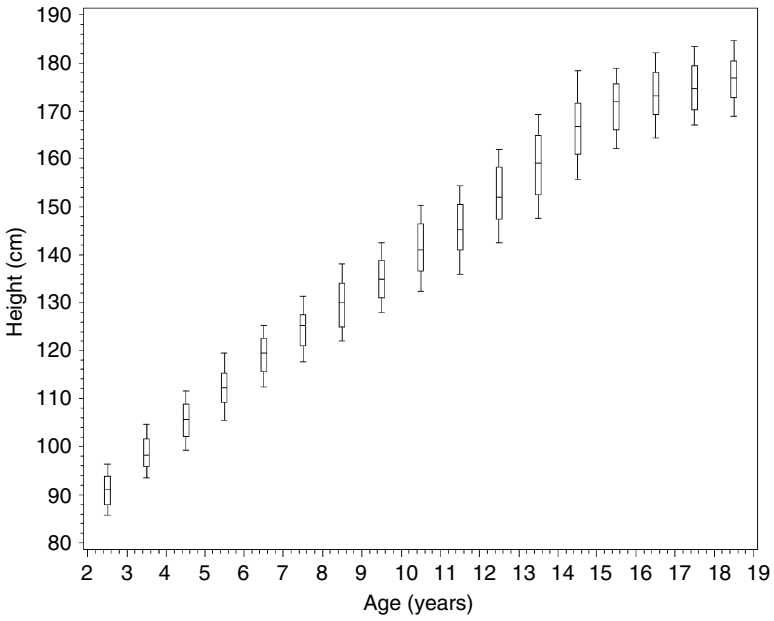


Fig. 14. Strip box plot of height versus age for data plotted in Fig. 6. Box plots show weighted 10th, 25th, 50th, 75th, and 90th percentiles for each year of age.

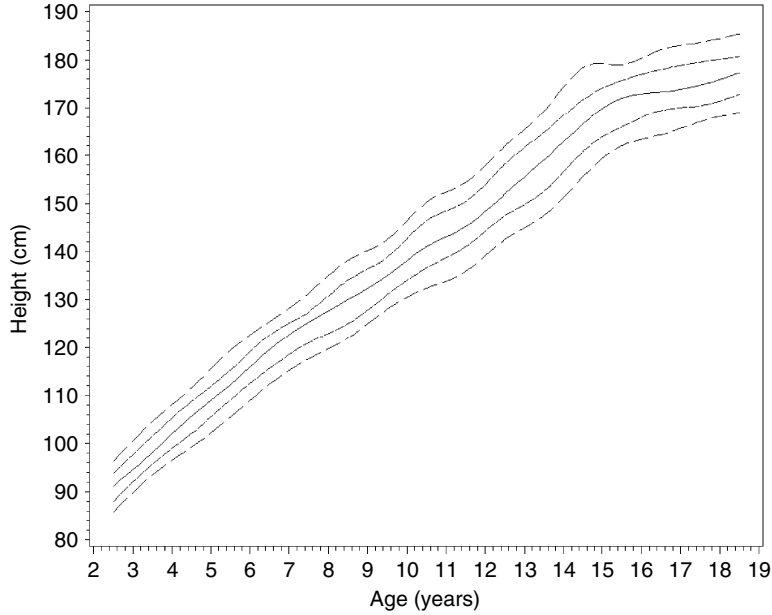


Fig. 15. Cubic spline interpolation of weighted percentiles are shown in Fig. 14. Solid line is the median, dashed lines are the quartiles, and outer dashed lines are the 10th and 90th percentiles.

range from 144 to 429. Figure 14 is not a particularly pleasing display of the percentiles as a function of age. One can remove the boxes and generate smooth curves through the percentiles for the different ages for a better plot. For example, Fig. 15 displays a piecewise cubic spline (SAS, 1990) that fits third-degree polynomials through the percentiles between adjacent years of age; this was the type of approach used in an early presentation of growth curves by the National Center for Health Statistics (1976). Guo et al. (1990) discuss alternative methods for smoothing percentiles for this type of grouped data.

More direct approaches to estimating smooth conditional percentile or mean curves are possible using the original ungrouped data. There are many different ways to do this (Härdle, 1990); we briefly describe a kernel method. Let $\{(x_i, y_i, w_i) \mid i = 1, \ldots, n\}$ be the sampled (X, Y) data with their corresponding sample weights. The idea behind a kernel estimator of the conditional mean of Y given $X = x$ is to evaluate the weighted mean of the y_i whose corresponding x_i are near x. The weights used for this weighted mean incorporate the sample weights and can also weight points with x_i close to x more than points x_i further from x by the choice of a "kernel function." We describe in the next section how to incorporate the sample weights into a particular kernel smoother. The end result is that one can express an estimator of the conditional mean as

$$\text{mean}(y|x) = \sum_{i=1}^{n} w_i^{\text{LS}} y_i, \tag{3}$$

where the weights w_i^{LS} incorporate the sample weights as well as the choice of the kernel function, local regression smoothing, and bandwidth. Figure 16 is a replot of the

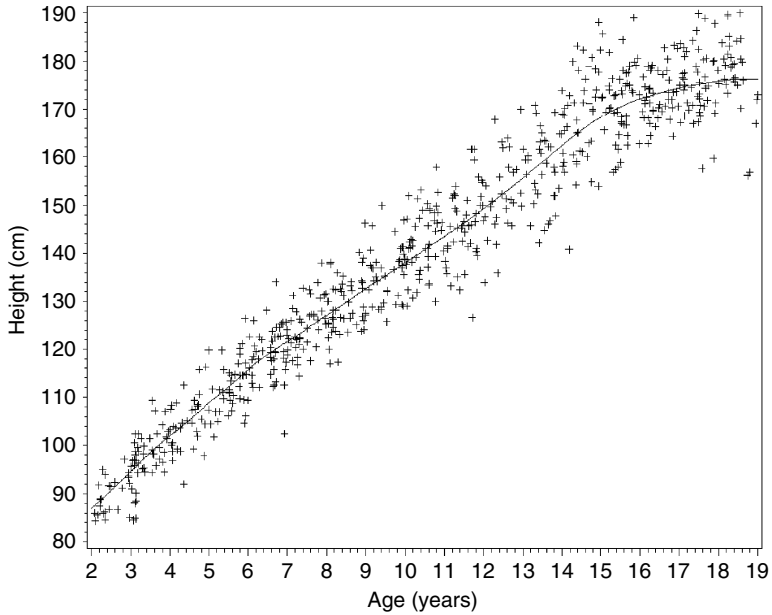


Fig. 16. Replot of Fig. 7 with the local-linear kernel estimator of the conditional mean using a triangular kernel with a bandwidth determined by a one-sided sample size of 350.

sampled scatterplot of Fig. 7 with the local-linear kernel estimator of the conditional mean using a triangular kernel with a bandwidth determined by a one-sided sample size of 350 (see the next section) (The conditional mean estimator uses the full sample of size 3667 and not just the points plotted in Fig. 7.).

2.5.1. Details of kernel smoothing

Let the kernel function $K(u)$ be a non-negative symmetric function that integrates to one, for example, the triangular kernel $K(u) = 1 - |u|$ for $|u| \leq 1$ and 0 otherwise. In the nonsurvey setting, one possible kernel estimator of the conditional mean is given by

$$\text{mean}^K(y|x) = \sum_{i=1}^{n} w_i^K y_i, \tag{4}$$

where $w_i^K = K\left(\frac{x - x_i}{h_x}\right) \Big/ \sum_{j=1}^{n} K\left(\frac{x - x_j}{h_x}\right)$ and h_x is the bandwidth that essentially determines how far the x_i can be from x and still be included in the estimator $\text{mean}^K(y|x)$. A potential problem with the curve $\text{mean}^K(y|x)$ is at the boundaries of the X data. One way to avoid this problem is to use a locally weighted regression (Cleveland, 1979), with a local-linear smoother being a special case: instead of using the weighted mean (4), one fits a weighted linear regression to the data around x using the w_i^K weights. Then, one defines $\text{mean}^L(y \mid x)$ to be the predicted value of Y at $X = x$ from this regression. The estimator $\text{mean}^L(y \mid x)$ can still be defined as a weighted

mean, namely,

$$\text{mean}^L(y|x) = \sum_{i=1}^{n} w_i^L y_i \tag{5}$$

with weights equal to

$$w_i^L = w_i^K \left[1 + \frac{\left(x_i - \overline{x}^K\right)\left(x - \overline{x}^K\right)}{\sum_{j=1}^{n} w_j^K \left(x_j - \overline{x}^K\right)^2} \right],$$

where $\overline{x}^K = \sum_{j=1}^{n} w_j^K x_j$. The additional possibility of downweighting points with large residuals ("lowess," Cleveland, 1979) is not pursued here.

In the survey setting, to account for the sample weights (w_i), one let

$$w_i^{\text{KS}} = w_i K\left(\frac{x - x_i}{h_x}\right) \Big/ \sum_{j=1}^{n} w_j K\left(\frac{x - x_j}{h_x}\right)$$

and

$$w_i^{\text{LS}} = w_i^{\text{KS}} \left[1 + \frac{\left(x_i - \overline{x}^{\text{KS}}\right)\left(x - \overline{x}^{\text{KS}}\right)}{\sum_{j=1}^{n} w_j^{\text{KS}} \left(x_j - \overline{x}^{\text{KS}}\right)^2} \right],$$

where $\overline{x}^{\text{KS}} = \sum_{j=1}^{n} w_j^{\text{KS}} x_j$. The local-linear smoother is then defined by (3). The use of the sample weights implies that (3) is estimating what (5) would be estimating if all the population values were available and used for the estimation.

Bellhouse and Stafford (2001, 2003) have proposed using local-polynomial regression to estimate a smooth conditional mean curve and have given asymptotic bias and variance properties of their estimators. This approach is a generalization to the local-linear smoother described above, where a weighted polynomial regression is used instead of the weighted simple linear regression.

The choice of the bandwidth is critical in determining how smooth the resulting conditional mean curve will be. There are various ways to choose the bandwidth (Härdle, 1990, Chapter 5; Ruppert et al., 1995). We describe two simple approaches here: one approach is to fix h_x to be a constant that is meaningful to the scale of the data at hand and a second approach is to choose h_x so that a certain minimum sample size is contained in $x \pm h_x$, for example, 100 observations. A modification of this second approach, which we prefer, is to choose h_x so that a certain minimum sample size is contained in either $[x, x - h_x]$ or $[x, x + h_x]$, for example, 50 observations. Without this modification, h_x will tend to increase as x approaches a boundary of the data.

A benefit of the development of the conditional mean estimator (3) as a weighted mean of the $y_i s$ is that the approach extends naturally to other functionals of the conditional distribution of Y given X, for example, percentiles. This was suggested by Stone (1977) and studied extensively by Owen (1987). The idea is to estimate the cumulative distribution function (CDF) for Y using the y_i, whose x_i are near x. In the present context, to estimate the conditional percentiles, one can use for each x the (weighted) percentile estimated from the weighted empirical CDF of the y_i using the w_i^{LS} weights. Unfortunately, this approach has a serious drawback for quantiles other than the median: even if the relationship of the quantiles and x was linear (but not horizontal), the larger

the bandwidth, the more the estimated quantiles will be biased away from the median. This is because the changing values of the conditional percentiles as a function of x causes the spread of y values to be larger when a larger bandwidth is considered.

To avoid this bias in the estimated conditional percentiles other than the median, we modify the approach analogously to that used for estimating "upper and lower smoothings" based on conditional means (Cleveland and McGill, 1984). We first estimate the conditional median using the weighted CDF as described above, denote it by $q_{50}(y|x)$ and let $z_i = y_i - q_{50}(y|x_i)$. To estimate a conditional percentile greater than the median, say the 90th percentile, use the weighted CDF approach to estimate the conditional 80th percentile of the zs given x using only the data points for which $z_i > 0$. If we denote this conditional 80th percentile by $q_{80}(z|x)$, then the desired conditional 90th percentile is estimated by $q_{50}(y|x) + q_{80}(z|x)$. In general, one estimates the conditional ηth percentile for $\eta > 50$ by $q_{50}(y|x) + q_{\gamma}(z|x)$, where $\gamma = 2\eta - 100$. This modification works for conditional percentiles less than the median in the obvious fashion. Figure 17 displays selected conditional percentiles for the height/age data using a local-linear kernel estimator using a triangular kernel with a bandwidth determined by a one-sided sample size of 350.

With large data sets, the discreteness of the scale of the measurement of Y can sometimes become noticeable in the conditional percentile curves. For example, consider the blood lead data described in Section 2.3. A plot of the smoothed conditional percentiles of blood lead versus age will take on only integer values since blood lead is recorded

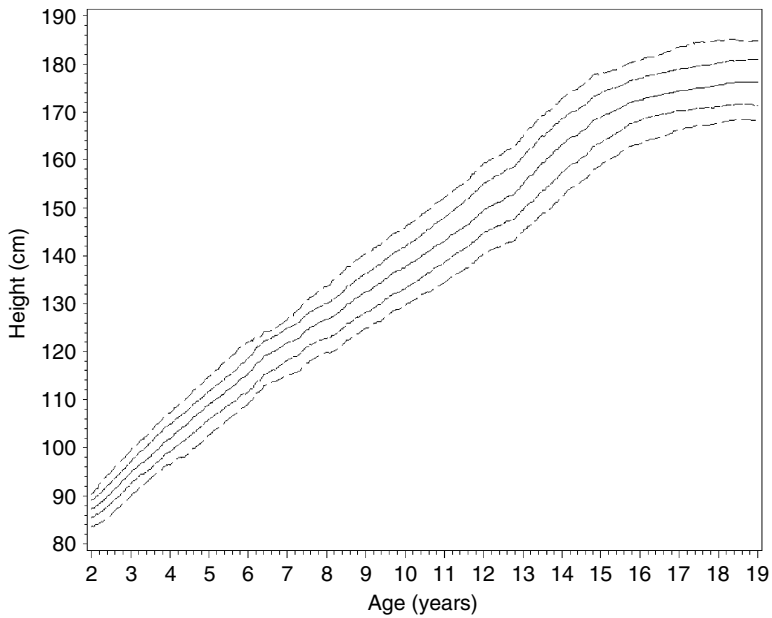


Fig. 17. Weighted conditional percentiles of height as a function of age of data plotted in Fig. 6. Solid line is the median, dashed lines are the quartiles, and outer dashed lines are the 10th and 90th percentiles. Conditional percentiles are estimated using a local-linear kernel estimator using a triangular kernel with a bandwidth determined by a one-sided sample size of 350 (see text).

to the nearest integer (plot not shown). If this is a problem, the weighted empirical CDF calculated at each x can itself be smoothed before estimating the percentiles, see Woodruff (1952) and Korn et al. (1997) for some simple methods of doing this.

The last issue we address is the calculation of standard errors for kernel estimators. A simple approach is to use one of the replication methods of variance estimation (Korn and Graubard, 1999, pp. 29–35). For example, with a balanced half-sample replication, the kernel estimator would be calculated using the data from each half-sample of primary sampling units. The bandwidth used for the replicated kernel estimators should be approximately the same as the bandwidth used for the kernel estimator for the original data (at each x). In particular, if a variable-length bandwidth involving a minimum sample size was used for the original data, you should *not* use a variable-length bandwidth involving the same minimum sample size for the replicates for a jackknife or a balanced half-sample replication. Instead, for example, you should, for a balanced half-sample replication, use a variable-length bandwidth involving half the minimum sample size used for the estimator on the original data or fix the bandwidth for the replicates at the value used for the original data. The rationale for this is that balanced half-sample replication or the jackknife is derived assuming that fewer observations are used in the replicate estimators. For example, balanced half-sample replication yields a reasonable variance estimator because the variability of the half-sample estimators is about twice that of the full-sample estimator. It should be noted that jackknife variance estimators should not be used for conditional percentile curves because they are not differentiable functions of the data. However, jackknife estimators can be used for conditional mean curves.

We caution the reader when using standard errors for kernel estimators. Although they can be interpreted as representing the variability one would see in the estimators if they were calculated from repeated independent surveys of the population, they cannot automatically be used to derive confidence intervals. This is because the smoothed estimators are biased (This bias is hard to quantify because it depends on the amount of smoothing and the curvature of the true curves.). This problem can become especially noticeable when a variable-width bandwidth is used and the data are scarce in a region of the horizontal axis. The bandwidth will be large in this region to capture a sufficient sample size. Therefore, the replicated standard errors of the smoothed curve will be no larger than at other regions of the curve where the data density is higher, presenting a potentially misleading picture. With cautious interpretation, however, we still believe that the presentation of the standard errors of kernel estimator is worthwhile. For example, if they are large, then the kernel estimators are not useful no matter what the size of the bias. If the sample size is so large that the bandwidth is quite small, then the bias of the smoothed estimators will be small.

As an alternative to presenting standard errors, we present a different method for examining whether a smoothed conditional mean or percentile curves is reflecting a property of the underlying distributions rather than just noise. As an example, Fig. 18 is a partial residual plot for the logarithm of blood lead from a (sample-) weighted multiple linear regression of systolic blood pressure on log lead, age, and body mass index using the data described previously. Partial residual plots for a particular independent variable x_1, also known as component-plus-residual plots, are plots of $r_i + \hat{\beta}_1 x_{i1}$ versus x_1, where $\hat{\beta}_1$ and the residuals r_i are estimated from the multiple linear regression model (Atkinson, 1985, Chapter 5.4; Cook and Weisberg, 1994, Chapter 9). These plots are useful for examining possible needed transformations of the independent variable. The

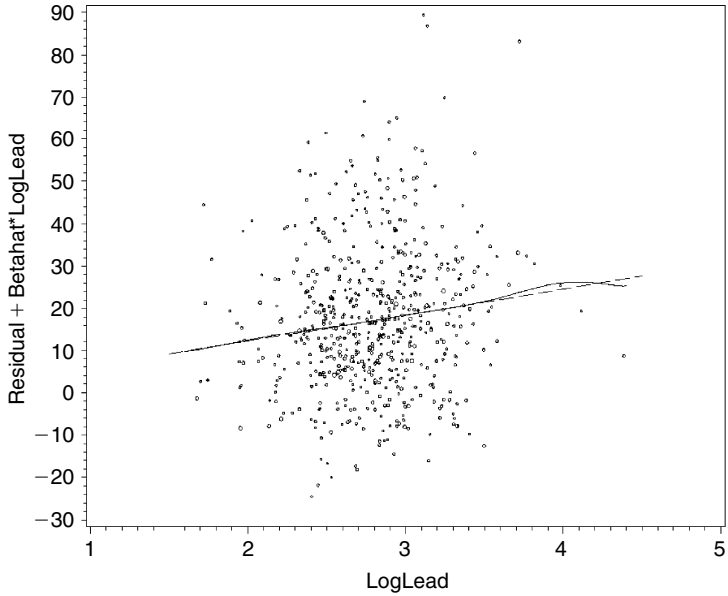


Fig. 18. Partial residual plot of the logarithm of blood lead (log lead) from a weighted regression of systolic blood pressure on log lead, age, and body mass index using data from 595 white males aged 40–59 sampled in the second National Health and Nutrition Examination Survey. Areas of circles are proportional to the sample weights. Dashed line is the weighted least-squares line. Solid line is the local-linear kernel estimator of the conditional mean using a triangular kernel with a fixed bandwidth of ± 1.5 units of log lead.

dashed line in Fig. 18 is the weighted least-squares line; its slope is identical to the estimated regression coefficient of log lead in the weighted multiple linear regression. Analogous partial residual plots can also be constructed for other types of regression analyses of complex survey data, including logistic regression and proportional hazard regression (Korn and Graubard, 1999, pp. 111–113, 124–126).

The smooth curve in Fig. 18 is a local-linear kernel estimator of the conditional mean using a triangular kernel with the fixed bandwidth of ± 1.5 units of log lead. The curve shows no great nonlinearity although there is the suggestion of a rise and then fall of the curve for log lead values greater than 3.5. As an ad hoc check of the reality of this nonlinearity, we simulated five data sets in which the linear regression model holds exactly—the values of the independent variables and the sample weights were taken as in the observed data set, and Y values were simulated with normal distributions around the predicted values (with standard deviation equal to the residual standard deviation from the observed data set computed without regard to possible sample weighting and correlation from cluster sampling). There should be no structure in the residuals from the weighted linear regressions using these simulated data sets. The top five curves in Fig. 19 are the estimated conditional mean plots from the partial residuals from these five simulated data sets; the bottom curve is a replot of the conditional mean curve from Fig. 18. The structure seen in these curves is at least as great as that seen in the curve calculated from the actual data, suggesting that the structure seen in the curve based on the actual data can be safely ignored.

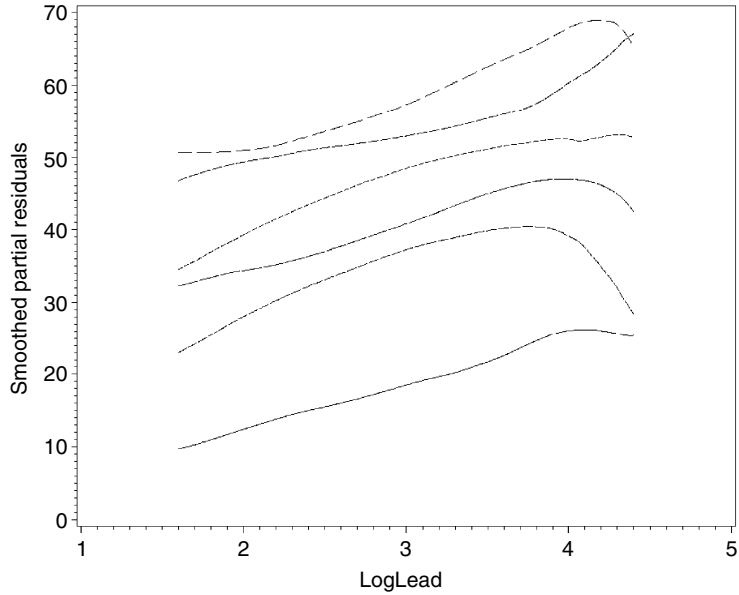


Fig. 19. Replot of the kernel estimator of the conditional mean from Fig. 18 (solid curve) with kernel estimators of the conditional mean based on five simulated data sets for which the conditional mean should be linear (dashed and dotted curves, translated in the vertical direction to avoid overlapping curves).

Although this section has focused on using kernel smoothing to obtain conditional means or quantiles in scatterplots with survey data, we note that there has been work in kernel smooth methods for density estimation with survey data (Bellhouse and Stafford, 1999, 2003; Buskirk and Lohr, 2005). The density estimation work may help to frame the theoretical basis for the kernel smoothing methods that we have presented. Finally, in addition to Härdle (1990) other references on topic of kernel smoothing for simple random samples are Wand and Jones (1995), Eubank (1999), and Simonoff (1996).

2.6. Regression splines

Regression splines is an alternative approach to kernel smoothing that use modeling when investigating the functional relationship between an outcome Y and a variable X. In the context of linear regression, the question is whether the simple inclusion of X as an independent variable is sufficient to model the relationship. A common modeling approach to this problem is to include powers of X as additional independent variables to allow for a polynomial relationship. For example, inclusion of X and X^2 in the model allows for Y to be a quadratic function of X and inclusion of X, X^2, and X^3 allows for Y to be a cubic function of X, etc. Frequently, however, this approach does not work because to adequately fit the data over the whole range of X may require a high-degree polynomial. This would lead to a nonparsimonious model with many independent variables involving X.

An alternative approach is to use a regression spline, which involves a fixed set of "knots" $t_1 < t_2 < \cdots < t_K$ in the range of X. The spline function of X is a piecewise

polynomial that is smoothly joined at the knots. Spline functions can be fit to the data by adding a small number of independent variables to the (linear) regression model. A very convenient type of spline (and the one we will discuss) is called "restricted cubic regression splines" (Stone and Koo, 1985), which are defined as follows: in between each pair of adjacent knots, the function is a cubic polynomial, with possibly different cubic functions between the different knot pairs. To the left of t_1 and to the right of t_K, the spline is straight lines (These linear constraints rather than allowing cubic functions in the tails are those that distinguish *restricted* cubic regression splines from ordinary cubic regression splines.). The cubic and linear functions are constrained so that the functional values and their first and second derivatives coincide at each knot; this ensures that the spline is a continuous smooth function.

Typically, a small number of knots (e.g., 3–5) are sufficient to model most data. Durrelman and Simon (1989) use knots at the following percentiles of the X data: {5, 50, 95}, {5, 25, 75, 95}, and {5, 25, 50, 75, 95} for 3, 4, and 5 knots, respectively. In survey data with sample weights, the knots can be placed at the weighted percentiles.

It is simple to express a restricted cubic regression spline in terms of functions of X that are included as independent variables in a regression. Details of the construction of these independent variables are given in Durrelman and Simon (1989) or Korn and Graubard (1999, Appendix C). With 3 knots, besides X, one need only to include one additional independent variable; with 4 knots, two additional independent variables; etc (For details about other approaches to estimating splines in simple random samples see Eubank, 1999.)

In most applications, there will be other independent variables in the model in addition to X and the spline variables that are functions of X. The interpretation of the spline function in these situations is the usual conditional one for a regression. A nice feature of using regression splines is that one can easily test whether a linear relationship is adequate by testing whether the estimated regression coefficients of the spline variables are significantly different from zero. For the restricted cubic regression spline with 3 knots, this involves testing if one regression coefficient equals zero; with 4 knots, a simultaneous test of whether two regression coefficients equal zero is used; etc. For survey data, these tests can be performed by estimating the coefficients using a sample-weighted regression and using a Wald statistic that incorporates the survey design. Standard linear regression software for survey data can be used for the analysis (although the analyst may be required to generate the spline variables).

When a linear relationship is inadequate to model the data, plotting the regression spline can suggest alternative nonlinear relationships. With no other independent variables in the model, a plot of the predicted values versus X can be overlaid on a scatterplot of the data. When there are other independent variables in the model $(Z_1, \ldots, Z_p)$, the following procedure can be applied. For example, for multiple linear regression modeling, suppose there are K knots with the corresponding $K-2$ spline variables in X being $S_{K,1}, S_{K,2}, \ldots, S_{K,K-2}$. Plot

$$\hat{\alpha} + \hat{\beta}_X X + \hat{\gamma}_{K,1} S_{K,1} + \cdots + \hat{\gamma}_{K,K-2} S_{K,K-2} + \hat{\beta}_{Z_1} c_1 + \cdots + \hat{\beta}_{Z_p} c_p$$

versus X, where $\{\hat{\alpha}, \hat{\beta}_X, \hat{\gamma}_{K,1}, \ldots, \hat{\gamma}_{K,K-2}, \hat{\beta}_{Z_1}, \ldots, \hat{\beta}_{Z_p}\}$ are the estimated regression coefficients from the model including the spline variables, and $c_1, \ldots, c_p$ are a set of constants representing possible values of $Z_1, \ldots, Z_p$. This plot is interpreted as the

predicted value for an individual with covariate values $Z_1 = c_1, \ldots, Z_p = c_p$ and $X = x$, as a function of x. A confidence band for this plot can be calculated using the estimated covariance matrix of the estimated regression coefficients (The width of the confidence band will depend on the particular values $c_1, \ldots, c_p$ chosen.). This plot can be overlaid with a plot of $\tilde{\alpha} + \tilde{\beta}_X X + \tilde{\beta}_{Z_1} c_1 + \cdots + \tilde{\beta}_{Z_p} c_p$ versus X (which is a straight line), where $\{\tilde{\alpha}, \tilde{\beta}_X, \tilde{\beta}_{Z_1}, \cdots, \tilde{\beta}_{Z_p}\}$ are the estimated coefficients from the linear regression model without the spline variables. The resulting plot allows for a comparison of the linear and spline-modeled associations of Y and X.

3. Discussion

In the nonsurvey setting, the simple scatterplot is an excellent overall graphical display of bivariate data. In the survey setting, different purposes may be best suited by different plots. For example, is the plot to describe the sample for data cleaning purposes or is to describe the population for population inference? With large sample sizes, is the plot to describe general trends or is to identify possible outliers and influential points? We have given examples in this chapter of some modifications of the simple scatterplot that we have found useful for displaying survey data. Other modifications are possible, and may be advisable, depending on the survey and the purpose of the display.

Introduction to Part 6

Danny Pfeffermann

1. Motivation

A common perception of survey sampling is that it is solely engaged with inference on finite population quantities such as means and proportions, with the inference based on the randomization distribution over repeated sampling from the fixed, finite population from which the original sample has been drawn. This perception is obviously false, as easily concluded from just cursory reading of the many chapters of this volume. Although inference about finite population quantities based on the randomization distribution is still a major component of the work of government bureaus producing official statistics, survey data are increasingly used for pure modeling purposes, and statistical models are often used for estimating finite population quantities, such as in small area estimation, a growing undertaking by statistical bureaus throughout the world (Chapter 32). The use of statistical models is inevitable when dealing with nonsampling errors such as coverage errors, nonresponse, and measurement errors, the focus of Chapters 8, 10, 12, and 25.

The use of the randomization distribution for inference is generally conceived as being robust, by not relying on an underlying statistical model assumed to generate the finite population values, which, as stated above, are considered as fixed numbers. Thus, an estimator $T(s)$ is unbiased under the randomization distribution for the population total T in a finite population U, if $\sum_s \Pr(s)T(s) = T$, where $\Pr(s)$ denotes the probability of drawing the sample s and the summation is over all the samples possibly drawn under the sampling design used to draw the sample. The randomization variance of the unbiased estimator $T(s)$ is defined accordingly as, $\mathrm{Var}_R[T(s)] = \sum_s \Pr(s)[T(s) - T]^2$. Brewer and Särndal (1983) noted that the use of the randomization distribution is "robust by definition." "Since no model is assumed, there is no need to discuss what happens under model breakdown." However, for inference beyond point estimation, such as the construction of a confidence interval, the use of this approach requires having sufficiently large samples (and hence sufficiently large populations), such that the randomization distribution of the estimator can be approximated by the normal distribution, based on an appropriate version of the central limit theorem. This is so because the exact randomization distribution of a sample statistic is generally unknown as it depends on the unknown population values. Smith (1994) states that "without the normal approximation there could be no randomization-based inference." Chapter 40 in Part 6 examines the conditions under which randomization-based estimators can be regarded as approximately normally distributed.

The estimation of finite population quantities can alternatively be based on statistical models assumed to generate the finite population values, which are now viewed as random variables. See, for example, Chapters 23, 24, 32, 36, and 39. Under this approach, the sampled units are considered fixed, and the estimation of the finite population total, for example, turns into a prediction problem of predicting the unobserved values for the nonsampled units under the assumed model. Denoting the population values by $\{Y_i, i \in U\}$ and the model-based predictors of the Y-values for the nonsampled units by $\{\hat{Y}_j, j \notin s\}$, the "model-based" predictor of the finite population total is $\hat{T}(s) = \left(\sum_{k\in s} Y_k + \sum_{j\notin s} \hat{Y}_j\right)$. It is unbiased if $E_{\mathrm{M}}\left[\left(\sum_{j\notin s} \hat{Y}_j - \sum_{j\notin s} Y_j\right) | S\right] = 0$, where E_{M} defines the expectation under the model. The model-based prediction variance (assuming model unbiasedness) is now $E_{\mathrm{M}}\left[\left(\sum_{j\notin s} \hat{Y}_j - \sum_{j\notin s} Y_j\right)^2 | S\right]$.

A third approach, often used for comparing different randomization-based strategies (sampling design and estimator) when direct comparisons are not feasible, is to combine the randomization and model distributions. For example, for comparing different randomization variances, one may compare their model expectations, known as the anticipated variance. Denoting a strategy for estimating the population total T by $[\Pr(s), T(s)]$, where $\Pr(s)$ defines the sampling design and $T(s)$ the (unbiased) estimator, the anticipated variance is $E_{\mathrm{M}}\mathrm{Var}_{\mathrm{R}}[T(s)] = E_{\mathrm{M}}\left\{\sum_s \Pr(s)[T(s) - T]^2\right\}$. See Chapter 24 and Chapters 39 and 41 in Part 6 for further discussion and uses of the combined distribution.

The use of statistical models is in line with conventional statistical inference, and the question arising is why survey sampling inference should be based on the randomization distribution and thus be different from "mainstream" statistics. There are many articles in the statistical literature discussing the pros and cons of the use of either one of the two approaches for estimating finite population quantities, see, for example, Smith (1994, 1997) and Chapters 23, 24, 32, 36, and 39. A related interesting question, however, with theoretical and practical implications, is whether randomization-based strategies can be justified by classical statistical theory by being, for example, minimax or Bayes rules. Recall in this respect the result of Godambe (1955) that states that no minimum variance linear unbiased estimator with coefficients that depend on the sample exist with respect to the randomization distribution for an arbitrary population. This result, however, does not rule out the consideration of other optimality criteria, the topic of Chapter 41 in Part 6. As mentioned earlier, this chapter considers also the combined model-randomization distribution for inference.

There is an overall agreement even among proponents of randomization-based inference that the use of models is inevitable for dealing with nonsampling errors such as coverage errors, nonresponse, and measurement errors, or in situations where the samples are too small, producing estimators with unacceptable large randomization variances and not allowing the use of asymptotic inference via the central limit theorem. Correspondingly, although model-based inference does not require in principle random sampling, there seems to be an agreement among modelers that for survey sampling inference, the sample should be selected at random, preferably within strata, either because there is usually not sufficient prior information for selecting a more efficient purposive sample, or because large scale surveys are practically always multipurpose, and a purposive sample that could be suitable for one target variable might be very inefficient for another variable. A more theoretical argument in favor of randomization

is that randomization can be seen as a mixed strategy in game theoretic terminology, and minimax results follow naturally, see Chapter 41. It is also often advocated that randomization protects against possible selection bias effects when fitting models to data. However, as discussed later, selection bias could be present even with random samples if the sampling design is informative, and failure to account for it can bias the inference very substantially, the focus of Chapters 38 and 39.

We considered so far the estimation (or prediction) of finite population quantities, commonly referred to as descriptive inference, but in recent decades, survey data are increasingly used for analytic inference about statistical models assumed to generate the corresponding population values, without necessarily employing the models for descriptive purposes. Familiar examples include the estimation of income elasticities from household surveys, the analysis of labor market dynamics from labor force surveys, and the study of the relationships between risk factors and disease incidence from health surveys. How should statistical models be fitted to survey data selected by probability sampling designs? Can the sample selection be ignored and the model be fitted as if the sample was actually a census? Is there a role for randomization-based inference when fitting models? Chapters 38 and 39 examine this issue in detail, with Chapter 39 reviewing several plausible approaches for fitting models to complex survey data and Chapter 38 discussing modeling procedures for case–control studies; one of the most common designs in health research.

To understand why modeling of complex survey data can be problematic, let us suppose that the target model assumed to generate the population values (the *census model*) is specified accurately. If the sampling design is simple random sampling and all the sampled units respond, the same model can be assumed to hold also for the sample data after selection, and it may be fitted using Bayesian, likelihood, or least squares methods, as found appropriate. The goodness of fit of the census model can be tested in this case using conventional techniques. But large scale surveys generally involve multistage cluster sampling and unequal selection probabilities, at least at some stages of the sampling process, with possibly not missing at random (NMAR) nonresponse. Consequently, the model holding for the sample data (the *sample model*) can be very different from the target population model, even if the ultimate sampling units are sampled with equal probabilities. Failure to account for the difference between the sample model and the census model can result in biased and inconsistent parameter estimators, poor coverage of confidence intervals, wrong predictions, and ultimately erroneous conclusions.

We illustrate the difference between the census model and the sample model, and the possible implications of ignoring the sample selection by considering the following simple example taken from Chapter 39. Let the census model be $Y_i|i \in U \sim \text{Mult}(\{p_k\}, K)$, such that $\Pr_U(Y_i = k) = p_k, k = 1, \ldots, K$; $\sum_{k=1}^{K} p_k = 1$, and suppose that unit $i \in U$ is sampled with probability $\Pr(i \in s|Y_i = k) = \pi_k$. Assume for convenience full response. Selection based on the values of the outcome variable underlies the use of case–control studies considered in Chapter 38. By Bayes rule, the sample model is $Y_i|i \in s \sim \text{Mult}\left(\{p_k^*\}, K\right)$, where $p_k^* = P_S(Y_i = k) = \Pr(Y_i = k|i \in S) = \pi_k p_k \big/ \sum_{j=1}^{K} \pi_j p_j$. We conclude that the population and sample models are different, unless the π_k are the same for all k. A sampling design under which the sample selection probabilities depend on the values of the outcome variable is informative because as we have just seen, under such designs the population and the sample distributions

are different. Ignoring the sample selection in our example and estimating the population probabilities $\{p_k\}$ by the "ordinary" estimates $\{\hat{p}_k = (n_k/n)\}$, where n_k is the number of sampled units with $Y_i = k$ and $n = \sum_{k=1}^{K} n_k$, yields in this case unbiased estimators of p_k^*, but biased estimators for the population probabilities p_k. Note, however, that for known selection probabilities π_k, one can construct the estimator $\tilde{p}_k = (\hat{p}_k/\pi_k)\big/\sum_{j=1}^{K}(\hat{p}_j/\pi_j)$, which is consistent for p_k under mild conditions. See Chapters 38 and 39 for further discussion.

2. Overview of chapters in Part 6

Chapter 38 considers case–control studies, which as mentioned earlier are used extensively for health research. The problem discussed is the fitting of a logistic regression model (the census model), with the response variable Y taking the value $Y = 1$ for a "case," (for example, an individual who contracted a certain disease), and the value $Y = 0$ for a "control" (who has not contracted the disease). It is assumed that the Y-values are known for every unit in the population. The sampling designs considered in this chapter consist of sampling independently from the subpopulations of cases and controls, and then collecting the information on the unknown explanatory variables $\mathbf{x}$ in the model for the sampled units. The samples of cases and controls are drawn by probability sampling using standard sampling techniques that may involve complex stratified multistage designs.

The major problem of case–control studies is that the sample selection probabilities depend directly on the model response variable (the most important design variable and often the only design variable), and the selection probabilities differ enormously between the two subpopulations, often by several orders of magnitude. Thus, case–control studies are an extreme example of informative sampling and the sample selection cannot be ignored at the inference process. In addition, when cluster sampling is involved, intracluster correlation needs to be taken into account when computing standard errors that are often used for the construction of confidence intervals or for testing hypotheses.

The first estimation procedure considered by the authors (and in many other chapters of the volume, see, in particular, Chapters 24, 26, and 39) is the use of what is known in the survey sampling literature as pseudolikelihood. The idea here is that instead of solving the census model score equations that would be obtained in the case of a census, one solves a randomization unbiased estimator of these equations. For the logistic model considered in this chapter, the census model score equations are, $\sum_{i=1}^{N} \mathbf{x}_i[Y_i - p_1(\mathbf{x}_i;\beta)] = 0$, where $p_1(\mathbf{x}_i;\beta) = \exp(\beta_0 + \mathbf{x}_i'\beta_1)\big/\big[1 + \exp(\beta_0 + \mathbf{x}_i'\beta_1)\big]$. The solution of these equations yields the maximum likelihood estimator (mle) of $\beta = (\beta_0, \beta_1')'$. The pseudo maximum likelihood estimator (pmle) of β is obtained by solving instead, $\sum_{k\in s} w_k\mathbf{x}_k[Y_k - p_1(\mathbf{x}_k;\beta)] = 0$, where $w_k = 1/\Pr(k \in s)$ is the base sampling weight, possibly adjusted for nonresponse or poststratification. Thus, the randomization distribution is seen to have a possible role even for analytic inference about a population model!

The major problem with the use of this approach is that the pmle may have a very large variance due to the extreme variability of the sampling weights. The authors consider therefore other, more efficient estimation procedures that include semiparametric maximum likelihood estimates and various ways of rescaling of the base weights. This is

shown to improve the efficiency very significantly. In the rest of the chapter the authors consider special sampling designs in common use for case–control studies and study appropriate estimation procedures that take into account stratification and clustering. An important feature of this chapter is that the authors show in some detail how available computer software can be used for implementing the various estimation procedures, taking into account the particular sampling designs used to collect the data. See also Chapter 13.

Chapter 39 reviews several plausible approaches for fitting models to complex survey data, with emphasis on informative sampling designs. As illustrated earlier, under informative sampling the population (census) model and the sample model (the model holding for the sampled units) are different, requiring special methods for proper inference about the target census model. The chapter begins by examining the conditions that warrant the ignorability of the sampling (response) process for likelihood, Bayesian, or sampling distribution inference. As implied by these conditions, a possible way to account for the sampling and response effects is to include among the model covariates all the variables and interactions that determine the sample and response probabilities. When this information is not available to the analyst fitting the model, as is often the case, it may be possible to use instead the sampling weights. However, as discussed in the chapter, for general (large scale) surveys, the sampling weights may not be an adequate summary of the missing design information (or the variables determining the response).

The main (commonly used) approach to deal with informative sampling is to weight the sample observations by the sampling weights, often referred to as probability weighting. For example, the probability weighted estimator of the simple regression slope coefficient is, $\hat{\beta}_w = \sum_{i\in s} w_i\big(y_i - \hat{\bar{Y}}_{\mathrm{HT}}\big)\big(x_i - \hat{\bar{X}}_{\mathrm{HT}}\big)\big/\sum_{i\in s} w_i\big(x_i - \hat{\bar{X}}_{\mathrm{HT}}\big)^2$, where $\big(\hat{\bar{Y}}_{\mathrm{HT}}, \hat{\bar{X}}_{\mathrm{HT}}\big)$ are the familiar Horvitz–Thompson estimators of the population means of Y and X. Chapter 39 reviews several variants of this approach for estimating population model parameters with many examples, including the estimation of the fixed parameters of two-level models under informative sampling of first and second-level units.

Another approach discussed at length in Chapter 39 is the use of the sample distribution for inference. The sample distribution (model) is the distribution of the outcomes given the selected sample of units and it is modeled as a function of the population distribution (model) and the sample selection probabilities. See the example in the first section of the introduction. Basing the inference on this distribution overcomes many of the problems underlying the other approaches reviewed in this chapter, but it requires modeling the expectations of the sampling and response probabilities as functions of the observed data. Several models and estimation procedures of these expectations and subsequent (or simultaneous) estimation of the population model parameters are reviewed and illustrated in the chapter.

Two other important topics discussed in Chapter 39 are as follows: (i) the use of the sample distribution for prediction, with application to small area estimation under informative sampling of areas and within the selected areas, and (ii) testing the informativeness of the sample selection and response. For complex sampling designs in common use, it is generally difficult and often impractical to check directly the conditions under which the sampling process can be ignored, evoking the need for test procedures that

can guide the analyst in determining whether the sampling process is ignorable for the type of inference intended.

Chapter 40 examines the asymptotic properties of sample survey estimators under the randomization distribution framework. The randomization distribution depends on the particular sampling design chosen by the sampler, and so does generally the asymptotic behavior of the estimator under consideration. The authors define the concept of a U-statistic (of which the simple sample mean and variance as special cases), and state that for simple random sampling without replacement (SRSWOR) and under very general regularity conditions, the U-statistic has asymptotically a normal distribution around the corresponding population parameter as the population and sample sizes grow to infinity. Extensions of this result to vector U-statistics and to the case of stratified SRSWOR are considered. Another interesting case is simple random sampling with replacement (SRSWR), but with the estimators based only on the distinct units.

As pointed out earlier, the asymptotic normality of sample survey estimators under the randomization distribution is crucial for probabilistic inference such as the construction of confidence intervals. Alternatively, the asymptotic distribution of U-statistics or functions of them, such as the ratio or regression estimators can possibly be approximated by the jackknife or bootstrap distributions, and the authors examine conditions that justify the use of these approximations. A somewhat different problem but often of practical importance is the estimation of a population size. Here the familiar (Petersen) method of "capture, mark, release, and recapture" is extended to more than one round and the authors consider alternative estimators with identical asymptotic behavior. In particular, the asymptotic normality of the estimators is established.

All the results mentioned so far basically assume simple random sampling designs or simple extensions of them. Often, however, the sample is drawn with unequal probabilities giving rise to the use of Horvitz–Thompson (HT) type estimators or functions of these estimators. There are many procedures in common use for sampling with unequal probabilities (see Chapter 2), and the authors study in detail many of these procedures, stating the conditions under which the corresponding standardized HT estimator with an appropriate standard deviation (SD) estimator converges to the standard normal distribution. The SD estimators considered do not generally require knowledge of the joint selection probabilities. In the rest of the chapter the authors consider extensions to stratified, multistage sampling designs.

Chapter 41 studies decision-theoretic aspects of strategies under the randomization distribution, and under the combined model-randomization distribution. As pointed out at the beginning of the chapter, "decision theory provides tools and insights for understanding, comparing, and selecting sampling and estimation procedures." The emphasis in the chapter is on optimality criteria, which when considering strategies, require optimizing over both the estimator and the sampling design. For example, Let $R(P_S, t_S; y) = E_R[L(t_S, T)]$ denote the risk associated with a loss function L when estimating the population total $T = \sum_{i=1}^{N} y_i$ by the estimator t_S, using the sampling design P_S, (the expectation of the loss under the randomization distribution). A strategy (P_{0S}, t_{0S}) is minimax for a given class of strategies if $\sup_y R(P_{0S}, t_{0S}; y) \leq \sup_y R(P_S, t_S; y)$ for every other strategy (P_S, t_S) in the class. The supremum is over all possible vectors $y = (y_1, \ldots, y_N)$ of the finite target population, considered as a fixed unknown parameter in some parameter space. A minimax strategy guarantees the smallest risk for the worst possible y.

The chapter proves and illustrates many interesting results on minimax strategies. For example, it is shown that the strategy (SRSWOR, $\bar{y}_S$), where $\bar{y}_S$ is the simple sample mean is minimax with respect to a convex loss function L for estimating the population mean $\bar{y}$ in the class of unbiased strategies $\left(P, \hat{\bar{y}}\right)$ having a fixed sample size. Some of the results are extended to the case of stratified sampling and probability proportional to size (PPS) sampling. The relationship between minimax estimators and Bayes estimators (the posterior mean under a quadratic loss) is shown to yield further interesting results, which are illustrated by considering the estimation of a proportion.

As stated earlier, no uniformly minimum variance unbiased estimator with weights that depend on the sample exists under the randomization distribution. Consequently, weaker optimality criteria have to be considered instead. We mentioned already the minimax that minimizes the maximum risk over the parameter space. The Bayes risk averages the risk over the parameter space with respect to a prior distribution. Another weak criterion discussed in Chapter 41 is admissibility. An estimator t_{0S} is admissible in a given class of estimators under a given sampling design P_{S0} if there is no other estimator t_S in the class satisfying $R(P_{S0}, t_S; y) \leq R(P_{S0}, t_{0S}; y)$ for all y, with strict inequality for at least one y. Several results relating to admissibility are given in the chapter. For example, it is shown that for any design, the familiar HT estimator is admissible in the class of unbiased estimators.

Chapter 41 considers also the use of superpopulation models for survey sampling inference. In superpopulation models the vector $y = (y_1, \ldots, y_N)$ of population values is considered as a random realization from some distribution g, and as mentioned earlier, model-based inference uses this distribution for descriptive inference on finite population quantities. The emphasis in the chapter, however, is on the use of the combined model-randomization distribution, reviewing and proving several results related to strategies that use the simple sample mean for estimating the population mean and to linear estimators defined as linear combinations of the sample observations with weights that may depend on the sample.

Sample Surveys: Inference and Analysis, Vol. 29B
ISSN: 0169-7161

© 2009 Elsevier B.V. All rights reserved
DOI: 10.1016/S0169-7161(09)00238-7

Population-Based Case–Control Studies

Alastair Scott and Chris Wild

1. Introduction to case–control sampling

This chapter discusses the design and analysis of case–control studies conducted with a view to fitting binary regression models. Let Y denote a binary response variable which can take value $Y = 1$ (corresponding to a "case," for example someone who contracts a disease of interest) or $Y = 0$ (corresponding to a "control," someone who does not), and let $\boldsymbol{x}$ be a vector of explanatory variables or covariates. Our purpose is to fit binary regression models of the general parametric form $\text{pr}(Y = 1 \mid \boldsymbol{x}) = p_1(\boldsymbol{x}; \boldsymbol{\beta})$ modeling the probabilistic behavior of Y as a function of the observed values of the explanatory variables recorded in $\boldsymbol{x}$. We focus particularly on the logistic regression model because this is the usual model of choice in applications.

Simple population-based case–control sampling is depicted in Fig. 1. In such studies, separate samples are drawn independently from the case- and control-subpopulations of a real, finite target population represented as the central element of Fig. 1. Covariate information, $\boldsymbol{x}$, is then ascertained for sampled individuals. Descriptive inferences about the finite population itself are virtually never of interest, however. Instead, interest centers on the process that turns some individuals into cases (e.g., they contract the disease) and others into controls. This is represented on the left of Fig. 1. We behave as if our data has been produced by the two-phase sampling process depicted by Fig. 1 as a whole.

Case–control sampling is a cost-reduction device. If we could afford to collect data on the whole finite population we would do so since that is the data we would really

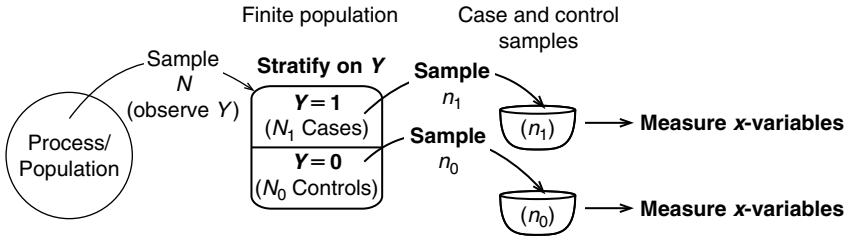


Fig. 1. Simple case–control sampling.

like to have to fit our binary regression model. In other words, we are interested in characteristics of models of the superpopulation. This has implications, in particular, for the calculation of standard errors. Case–control data is special because the response (outcome) variable is the most important design variable (so the design is not ignorable, see Chapter 39), and because the selection probabilities typically differ enormously, often by several orders of magnitude.

We note that there are other types of case–control studies that we will not deal with here. In particular, we will not discuss matched case–control studies, in which each case is individually matched with one or more controls (see Section 5 of Breslow, 1996). Korn and Graubard (1999, p. 307) call this "set matching." The subject of this chapter is unmatched studies, in which the case and control samples are drawn independently, and more particularly with population-based studies in which the controls (and occasionally the cases as well) are selected using standard survey sampling techniques. Our treatment will, however, accommodate loose "frequency matching" in which the control sample is allocated across strata defined by basic demographic variables in such a way that the distribution of these variables in the control sample is similar to their expected distribution in the case sample. For discussion about where this might be desirable, see Breslow (2005, pp. 298–302).

In principle, the most straightforward way of obtaining data from which to build regression models for $\text{pr}(Y \mid \boldsymbol{x})$ is to employ so-called prospective sampling designs. A cohort of individuals is selected and their covariate information is ascertained. They are then tracked through time and whether they become cases ($Y = 1$) or do not ($Y = 0$) is recorded. With prospective sampling designs, observation proceeds from covariates (explanatory variables) to response, corresponding to the logic underlying the modeling. A case–control study is counterintuitive in that data collection takes place in the reverse direction from response to covariate. Despite this, the case–control study is one of the most common designs in health research. In fact, Breslow and Day (1980) described such studies as "perhaps the dominant form of analytical research in epidemiology" and since that time the rate of appearance of papers reporting on case–control studies appears to have gone up by a factor of about 20. These designs are also used in other fields under other names. In econometrics, for example, the descriptor "choice-based" is used rather than "case–control" because the designs were developed in a research field that grew out of the work of Nobel Prize-winner Daniel McFadden in the late 1960s and 1970s on discrete economic choices (e.g., choice of a mode of transport); see Manski (2001).

There are several reasons for the popularity of case–control studies. The first two reasons concern efficiency; efficiency of time and statistical efficiency. The former comes from being able to use historical information immediately rather than having to follow individuals through time and then wait to observe an outcome as in a prospective study. We will not discuss this, or the attendant risks, but refer the reader to the first chapter of Breslow and Day (1980). The statistical efficiency advantages can be huge. For example, suppose that we have a condition that affects only 1 individual in 20 on average and we wish to investigate the effect of an exposure that affects 50% of people. In this situation, a case–control study with equal numbers of cases and controls has the same power for detecting a small increase in risk as a prospective study with approximately five times as many subjects. If the condition affects only one individual in 100 then the prospective

study needs 25 times as many subjects, while if it affects one individual in 1000 over the time period of interest then the prospective study needs 250 times as many subjects! Thus, case–control based sampling designs are very efficient for investigating the effects of explanatory variables on a comparatively rare response.

A third factor that has influenced the uptake of case–control studies is the simplicity of analysis. When fitting a logistic regression model including an intercept term to case–control data obtained from simple random samples, it is well known, following the landmark papers of Anderson (1972) for discrete covariates, and Prentice and Pyke (1979) for general $\boldsymbol{x}$, that valid inferences about all coefficients except the intercept can be obtained by fitting a logistic regression model using standard software as if it had been obtained prospectively. The intercept is completely confounded with the relative sampling rates of cases and controls but can be recovered using additional information such as the finite population totals of cases and controls (see Scott and Wild, 1986, for example).

Excellent well-referenced introductions to the strengths and potential pitfalls of case–control sampling are given by Breslow (1996, 2005). One of the most important and difficult challenges confronting anyone designing such a study is to ensure that controls really are drawn from the same population, using the same protocols, as the cases. In the words of Miettinen (1985), cases and controls "should be representative of the same base experience." Failure to ensure this adequately in some early examples led to case–control sampling being regarded with some suspicion by many researchers. Because the essence of survey sampling lies in methods for drawing representative samples from a target population, it became natural at some stage to think about using survey methods for obtaining controls. Increasingly over the last 25 years or so, studies are being conducted in which the controls (and occasionally the cases as well) are drawn using complex stratified multistage designs. A good history of this development can be found in Chapter 9 of Korn and Graubard (1999) and more recent work is reviewed in Scott (2006). These studies retain all the efficiency advantages of simple case–control studies but the analysis is no longer quite so simple. It is this aspect that we want to discuss here. We start with two examples to illustrate the sort of problem that we want to handle.

> **Example 1.** In 1977–78, the National Cancer Institute and the US Environmental Protection Agency conducted a population-based case–control study to examine the effects of ultraviolet radiation on nonmelanoma skin cancer over a 1-year period. This is typical of many large scale studies conducted by the National Cancer Institute whose personnel have been responsible for much of the development of population-based case–control studies (see Hartge et al., 1984a,b who also give a description of a number of other similar studies). The study was conducted at eight geographic locations with varying solar ultraviolet intensities. Samples of patients with nonmelanoma skin cancer aged 20 to 74 and samples of general population controls from each region were interviewed by telephone to obtain information on risk factors. At each location, a simple random sample of 450 patients and an additional sample of 50 patients in the 20–49 age group were selected for contact. For the controls, 500 households were sampled at each location using random-digit dialing. An attempt was made to interview all adults aged 65–74 as well as a randomly selected individual of each sex

aged from 20 to 64. In addition, a second telephone sample of between 500 and 2100 households was taken at each location and information gathered on all adults aged 65–74. After allowing for nonresponse, this resulted in samples of approximately 3000 cases and 8000 controls, with the sampling rate for cases being roughly 300 times the rate for controls, depending on age.

Example 2. The Auckland Meningitis Study was commissioned by the New Zealand Ministry of Health and Health Research Council to study risk factors for meningitis in young children, which was reaching epidemic proportions in Auckland at that time (see Baker et al., 2000). The target population was all children under the age of nine in the Auckland region in 1997.

All cases of meningitis in the target age group over the 3-year duration of the study were included in the study, resulting in about 200 cases. A similar number of controls was drawn from the remaining children in the study population using a complex multistage design. At the first stage of sampling, 300 census mesh blocks (each containing roughly 70 households) were drawn with probabilities proportional to the number of houses in the block. At the second stage, a systematic sample of 20 households was selected from each chosen mesh block and children from these households were selected for the study with varying probabilities that depended on age and ethnicity and were chosen to match the expected frequencies among the cases.

Cluster sample sizes varied from one to eight and a total of approximately 300 controls was achieved. This corresponds to a sampling fraction of about 1 in 400 on average, so that cases are sampled at a rate that is 400 times that for controls.

These two studies are fairly typical of the sort of study that we want to discuss. They also illustrate the two main sampling methods used, namely random digit dialing and area sampling. A lively discussion of the relative merits of these two strategies is given in Brogan et al. (2001) and DiGaetano and Waksberg (2002). Note that the response rates for telephone surveys have dropped substantially since 2002 so the outcome of that discussion might well be different if repeated now. (See Chapter 7 for more on random digit dialing.)

2. Basic results

2.1. Setup

We have a binary response variable, Y, with $Y = 1$ denoting a case and $Y = 0$ denoting a control, and a vector of potential explanatory variables, $\boldsymbol{x}$. We assume, as in Fig. 1, that the value of Y is known for all N units in some target population but that at least some components of $\boldsymbol{x}$ are unknown. We split the population into cases and controls, draw a sample from each subpopulation using a sample design based on the variables that we know for all units, and measure the values of the missing covariates for the sampled units. We want to use the sample data to fit a binary regression model for the marginal probability of being a case as a function of the covariates. The model used is almost always logistic with

$$\text{logit}\,\{\text{pr}(Y = 1|\boldsymbol{x})\} = \log\left\{\frac{P(Y = 1|\boldsymbol{x})}{P(Y = 0|\boldsymbol{x})}\right\} = \beta_0 + \boldsymbol{x}^T\boldsymbol{\beta}_1 \tag{1}$$

say, and we shall assume model (1) throughout the chapter. Extensions to more general regression models are straightforward in principle (see Scott and Wild, 2001b) but the resulting expressions can be considerably clumsier than those for the logistic model.

We are interested here in methods that allow for complex sampling, including stratified multistage sampling, of cases and controls in Fig. 1. It is important, when complex sampling is undertaken, that it should be taken into account in the analysis. Failure to do so can lead to all the usual problems that arise from ignoring survey structure. Varying selection probabilities can distort the mean structure if not taken into account and estimates produced by standard programs may be inconsistent. The use of different strata for cases and controls is an example of this type. In addition, intracluster correlation can reduce the effective sample size so that methods that ignore this correlation can lead to standard errors that are too small, confidence intervals that are too short, P-values that are too low, and so on. One simple strategy that has been adopted by some researchers to minimize the effect of intracluster correlation is to keep the numbers of subjects in each cluster small (see Graubard et al., 1989, for example). This reduces the design effect and hence the impact of clustering, but it can be a very expensive remedy.

For situations where it would make scientific sense to fit model (1) to data from the whole finite population, the methods currently available for the estimation of data obtained from complex sampling are variants of the standard weighted estimating equation approach discussed by Binder (1983) and now embodied in most modern packages for analyzing survey data. We will now briefly review this approach as it applies to the current context. (See Chapters 24 and 26 for a more detailed discussion of the general approach.)

2.2. *Basics of design-weighted estimation*

If we wanted to to fit model (1) to the whole finite population, then we could estimate $\boldsymbol{\beta}$ by solving the whole-population (or census) estimating equations

$$\boldsymbol{S}(\boldsymbol{\beta}) = \sum_{1}^{N} \boldsymbol{x}_i \{Y_i - p_1(\boldsymbol{x}_i; \boldsymbol{\beta})\} = 0, \tag{2}$$

where $p_1(\boldsymbol{x}; \boldsymbol{\beta}) = e^{\beta_0 + \boldsymbol{x}^T \boldsymbol{\beta}_1} \big/ \left(1 + e^{\beta_0 + \boldsymbol{x}^T \boldsymbol{\beta}_1}\right)$ and $p_0(\boldsymbol{x}; \boldsymbol{\beta}) = 1 - p_1(\boldsymbol{x}; \boldsymbol{\beta})$. These equations are the score equations from the log-likelihood if the N population Y-values are assumed to be independent but the resulting estimators are consistent under much more realistic population structures as long as model (1) holds marginally (see Rao et al., 1998, for further discussion). For any fixed value of $\boldsymbol{\beta}$, $\boldsymbol{S}(\boldsymbol{\beta})$ is simply a vector of population totals. This means that we can estimate $\boldsymbol{S}(\boldsymbol{\beta})$ from the sample by

$$\widehat{\boldsymbol{S}}(\boldsymbol{\beta}) = \sum_{\text{sample}} w_i \boldsymbol{x}_i \{y_i - p_1(\boldsymbol{x}_i; \boldsymbol{\beta})\}, \tag{3}$$

where the design weight w_i is the inverse of the selection probability, perhaps adjusted for nonresponse and poststratification. Setting $\widehat{\boldsymbol{S}}(\boldsymbol{\beta})$ equal to 0 gives us our estimator, $\widehat{\boldsymbol{\beta}}$. We could use linearization or the jackknife directly on $\widehat{\boldsymbol{\beta}}$ to get standard errors. Alternatively, we can expand $\widehat{\boldsymbol{S}}(\widehat{\boldsymbol{\beta}})$ about the true value, $\boldsymbol{\beta}$, and obtain as our estimated

covariance matrix the "sandwich" estimator

$$\widehat{\mathrm{Cov}}\left\{\widehat{\boldsymbol{\beta}}\right\} \approx \mathbf{J}(\widehat{\boldsymbol{\beta}})^{-1}\widehat{\mathrm{Cov}}\left\{\widehat{\boldsymbol{S}}(\widehat{\boldsymbol{\beta}})\right\}\mathbf{J}(\widehat{\boldsymbol{\beta}})^{-1}, \tag{4}$$

where $\mathbf{J}(\boldsymbol{\beta}) = -\frac{\partial \widehat{\boldsymbol{S}}}{\partial \boldsymbol{\beta}^T} = \sum_{\text{sample}} w_i\, p_1(\boldsymbol{x}_i;\boldsymbol{\beta})\, p_0(\boldsymbol{x}_i;\boldsymbol{\beta})\boldsymbol{x}_i\boldsymbol{x}_i^T$. Here we are ignoring the additional (lower order) error incurred by replacing the true value, $\boldsymbol{\beta}$, by the estimate, $\widehat{\boldsymbol{\beta}}$. Because $\widehat{\boldsymbol{S}}(\boldsymbol{\beta})$ is a vector of totals, $\widehat{\mathrm{Cov}}\{\widehat{\boldsymbol{S}}(\boldsymbol{\beta})\}$ should be available as a matter of course for any standard design.

Most major statistical packages can now handle logistic regression with complex sampling in this way. In this sense, then, producing weighted estimates of logistic regression parameters and making associated inferences from case–control data, even with very complex sampling, is now quite routine using widely available software packages – modulo, of course, the myriad difficulties and subtleties of accurately characterizing and describing any complex design to software, and adequately analyzing any large and complex set of data. The aforementioned is the approach recommended by Korn and Graubard (1999, Chapter 9) and it provides a very powerful and flexible set of tools as an even cursory reading of that chapter will reveal. We note in passing that variance estimation in standard software does not take account of the variation in the weights w_i that is due to the fact that the selection probabilities used are estimates obtained using N_1 and N_0, the total number of cases and controls in the population, which are themselves random rather than known fixed constants. In most cases the differences, which are of order $(1/N)$, are negligible. Section 2.6 contains further discussion and gives correction terms.

Unfortunately, design weighting used in this way can be very inefficient (see Scott and Wild, 2001a, for example). A great deal of efficiency can be recaptured, however, by a simple rebalancing of the relative weights given to cases and controls. We take the weights that would normally be used in (3) and rescale them so that the weights given to cases add to the number of cases sampled and likewise for the weights of controls. We justify these statements in the following subsection.

The inefficiency of standard design weighting for case–control data should not be unexpected. It is well known that weighting in general tends to be inefficient when the weights are highly variable. In case–control studies, the variation in weights is about as extreme as it can get and no experienced survey sampler would be surprised to find that weighting is not very efficient under these circumstances. When estimating means, a common measure of efficiency relative to the unweighted mean is $e = 1/(1+c^2)$, where c is the coefficient of variation of the weights (see Korn and Graubard, 1999, p. 173, for example). This works out to about $e = 0.7$ in Example 1 and $e = 0.3$ in Example 2. Even lower values than this are common. The effect on more complex statistics such as logistic regression coefficients tends to be smaller than for means but can still be substantial.

2.3. *Improving efficiency by reweighting*

Where the case and control samples are simple random samples, maximum likelihood estimates are available and easy to find. Scott and Wild (2002) simulated case–control sampling from a model with $\mathrm{logit}\{\mathrm{pr}(Y = 1 \mid x)\} = \beta_0 + \beta_1 x$, where x-values were generated from a standard Normal distribution. They showed that, when the case to

control ratio in the population was 1:400 and equal-sized random samples of cases and controls are taken, the efficiency of design-weighted (DW) estimation could be as low as 12% relative to that of maximum likelihood. Even with a 1:20 population case–control ratio, relative efficiencies of around 50% are not uncommon. These and other simulations showed that the relative efficiency of design weighting reduces as the case-to-control ratio gets smaller and also as $|\beta_1|$ increases.

Suppose that we have a simple random sample of size n_1 from the case stratum and an independent simple random sample of size n_0 from the control stratum. Here all units in Stratum ℓ have weight $w_i \propto \frac{W_\ell}{n_\ell}$, where W_ℓ denotes the proportion of the population in the stratum, for $\ell = 0, 1$. Thus, recalling that $Y_i = 1$ for cases and $Y_i = 0$ for controls, the estimating equation (3) can be written in the form

$$W_1 \frac{\sum_{\text{cases}} \boldsymbol{x}_i p_0(\boldsymbol{x}_i; \boldsymbol{\beta})}{n_1} - W_0 \frac{\sum_{\text{controls}} \boldsymbol{x}_i p_1(\boldsymbol{x}_i; \boldsymbol{\beta})}{n_0} = \mathbf{0}, \tag{5}$$

The likelihood for case–control data is based upon probabilities of the form $\text{pr}(\boldsymbol{x} \mid Y)$, which depend on the marginal distribution of $\boldsymbol{x}$ as well as the logistic regression parameters $\boldsymbol{\beta}$. The semiparametric maximum likelihood estimates of $\boldsymbol{\beta}$ are obtained by maximizing this likelihood over both $\boldsymbol{\beta}$ and the marginal distribution of $\boldsymbol{x}$ treated nonparametrically. Prentice and Pyke (1979) showed that, for all coefficients except the constant term, the resulting estimates satisfy the ordinary prospective likelihood equations

$$\sum_{\text{sample}} \boldsymbol{x}_i \{y_i - p_1(\boldsymbol{x}_i; \boldsymbol{\beta})\} = \mathbf{0}$$

which can be rewritten

$$\omega_1 \frac{\sum_{\text{cases}} \boldsymbol{x}_i p_0(\boldsymbol{x}_i; \boldsymbol{\beta})}{n_1} - \omega_0 \frac{\sum_{\text{controls}} \boldsymbol{x}_i p_1(\boldsymbol{x}_i; \boldsymbol{\beta})}{n_0} = \mathbf{0}, \tag{6}$$

where $\omega_\ell = \dfrac{n_\ell}{n}$, with $n = n_1 + n_0$, for $\ell = 0, 1$. Breslow et al. (2000) showed that this leads to semiparametric efficient estimators (i.e., having the smallest possible variance in the class of all consistent, asymptotically linear estimators).

Both (5) and (6) are special cases of the general set of estimating equations

$$\lambda_1 \frac{\sum_{\text{cases}} \boldsymbol{x}_i p_0(\boldsymbol{x}_i; \boldsymbol{\beta})}{n_1} - \lambda_0 \frac{\sum_{\text{controls}} \boldsymbol{x}_i p_1(\boldsymbol{x}_i; \boldsymbol{\beta})}{n_0} = \mathbf{0}. \tag{7}$$

As $n_0, n_1 \to \infty$ the solution of (7) converges almost surely to the solution of

$$\lambda_1 E_1 \left\{ \boldsymbol{X} p_0(\boldsymbol{X}; \boldsymbol{\beta}^*) \right\} - \lambda_0 E_0 \left\{ \boldsymbol{X} p_1(\boldsymbol{X}; \boldsymbol{\beta}^*) \right\} = \mathbf{0}, \tag{8}$$

where $E_\ell\{\cdot\}$ denotes the conditional expectation given that $Y = \ell$ for $\ell = 0, 1$. If model (1) is true, then equation (8) has solution $\boldsymbol{\beta}_1^* = \boldsymbol{\beta}_1$ and $\beta_0^* = \beta_0 + b_\lambda$ with $b_\lambda = \log\{\lambda_1 Q_0/(\lambda_0 Q_1)\}$, where $Q_\ell = pr(Y = \ell)$, for any positive λ_ℓ for $\ell = 0, 1$. This can be seen directly by expanding (8). For simplicity, suppose that $\boldsymbol{X}$ is continuous

with density function $f(\boldsymbol{x})$. Then the conditional density of $\boldsymbol{X}$ given that $Y = \ell$ is $f(\boldsymbol{x} \mid Y = \ell) = p_\ell(\boldsymbol{x}; \boldsymbol{\beta}) f(\boldsymbol{x})/Q_\ell$, with $Q_\ell = p_\ell(Y = \ell)$ and (8) is equivalent to

$$\int \frac{\boldsymbol{x} e^{\boldsymbol{x}^T\boldsymbol{\beta}+b_\lambda} f(\boldsymbol{x})}{(1+e^{\boldsymbol{x}^T\boldsymbol{\beta}^*})(1+\boldsymbol{x}^T\boldsymbol{\beta})} d\boldsymbol{x} = \int \frac{\boldsymbol{x} e^{\boldsymbol{x}^T\boldsymbol{\beta}^*} f(\boldsymbol{x})}{(1+e^{\boldsymbol{x}^T\boldsymbol{\beta}^*})(1+\boldsymbol{x}^T\boldsymbol{\beta})} d\boldsymbol{x}$$

and the result follows immediately. Thus the solution to equation (7) produces consistent estimators of all the regression coefficients, apart from the constant term, for any $\lambda_\ell > 0$, $\ell = 0, 1$.

It is easy to correct the inferences about the constant term by estimating Q_ℓ with W_ℓ, the proportion of cases in the population, provided that it is known. However, this is not always necessary. In medical applications, interest is often centred on the relative risks associated with whether or not an individual was exposed to a putative hazard or with an increase in the value of a continuous x-variable. The elements of $\boldsymbol{\beta}_1$ are the log odds-ratio parameters that tell us about the relative effects of changes in an x-variable. For keeping all x-variables but the jth fixed (and dropping the others from the notation for simplicity of exposition)

$$\begin{aligned}\beta_{1j} &= \log \frac{\text{pr}(Y=1 \mid x_j+1)}{\text{pr}(Y=0 \mid x_j+1)} \Big/ \frac{\text{pr}(Y=1 \mid x_j)}{\text{pr}(Y=0 \mid x_j)} \\ &\approx \log \frac{\text{pr}(Y=1 \mid x_j+1)}{\text{pr}(Y=1 \mid x_j)} \text{ if cases are rare.}\end{aligned}$$

The rightmost expression is the relative risk associated with a 1-unit increase in x_j. If we want to estimate absolute levels of risk we also need β_0. If, however, we are content to work with relative risks only (there is debate about whether this is really ever entirely adequate) and have a mechanism for taking separate samples of cases and controls in a way that they are both "representative of the same base experience," we can use (7) and (8) without correction and therefore any need to know anything about population sizes. This is part of the reason that we have de-emphasized the superpopulation aspects of Fig. 1.

We now adapt (7) and (8) to more complex sampling schemes. Because the left-hand side of equation (8) just involves two subpopulation means, we can still estimate these means for any standard survey design. This suggests an estimator, $\widehat{\boldsymbol{\beta}}_\lambda$ say, for general sampling schemes satisfying

$$\widehat{\boldsymbol{S}}_\lambda(\boldsymbol{\beta}) = \lambda_1 \widehat{\boldsymbol{\mu}}_1(\boldsymbol{\beta}) - \lambda_0 \widehat{\boldsymbol{\mu}}_0(\boldsymbol{\beta}) = \boldsymbol{0}, \tag{9}$$

where $\widehat{\boldsymbol{\mu}}_\ell(\boldsymbol{\beta})$ is the sample estimate of the subpopulation mean $E_\ell\{\boldsymbol{X}(1 - p_\ell(\boldsymbol{X}; \boldsymbol{\beta}))\}$, $\ell = 0, 1$. The covariance matrix of $\widehat{\boldsymbol{\beta}}_\lambda$ can then be obtained by standard linearization arguments. This leads to an estimated ("sandwich") covariance matrix

$$\widehat{\text{Cov}}\{\widehat{\boldsymbol{\beta}}_\lambda\} \approx \mathbf{J}_\lambda(\widehat{\boldsymbol{\beta}}_\lambda)^{-1} \widehat{\text{Cov}}\{\widehat{\boldsymbol{S}}_\lambda(\widehat{\boldsymbol{\beta}}_\lambda)\} \mathbf{J}_\lambda(\widehat{\boldsymbol{\beta}}_\lambda)^{-1}, \tag{10}$$

with $\mathbf{J}_\lambda(\boldsymbol{\beta}) = \left(-\frac{\partial \widehat{\boldsymbol{S}}_\lambda(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T}\right)$. Note that $\widehat{\text{Cov}}\{\widehat{\boldsymbol{S}}_\lambda(\boldsymbol{\beta})\} = \lambda_1^2 \widehat{\text{Cov}}\{\widehat{\boldsymbol{\mu}}_1(\boldsymbol{\beta})\} + \lambda_0^2 \widehat{\text{Cov}}\{\widehat{\boldsymbol{\mu}}_0(\boldsymbol{\beta})\}$ because the samples are taken independently from the case and control subpopulations. Here, $\widehat{\text{Cov}}\{\widehat{\boldsymbol{\mu}}_\ell(\boldsymbol{\beta})\}$ denotes the usual survey estimate that should be available routinely for any standard survey design since $\widehat{\boldsymbol{\mu}}_\ell(\boldsymbol{\beta})$ is just an estimated subpopulation mean.

Estimation can also be carried out straightforwardly using any package that can handle logistic regression for complex survey designs by specifying the appropriate vector of weights. Suppose that

$$\widehat{\boldsymbol{\mu}}_\ell(\boldsymbol{\beta}) = \frac{\sum_{i \in s_\ell} w_i \boldsymbol{x}_i \left(y_i - p_1(\boldsymbol{x}_i; \boldsymbol{\beta})\right)}{\sum_{i \in s_\ell} w_i}, \tag{11}$$

where s_1 denotes the case sample and s_0 denotes the control sample. Then the estimating equation (9) can be written in the form

$$\widehat{\boldsymbol{S}}_\lambda(\boldsymbol{\beta}) = \sum_{\text{sample}} w_i^* \boldsymbol{x}_i \left(y_i - p_1(\boldsymbol{x}_i; \boldsymbol{\beta})\right) = \boldsymbol{0}, \tag{12}$$

with $w_i^* \propto \frac{\lambda_\ell w_i}{\sum_{i \in s_\ell} w_i}$ for units in sample s_ℓ ($\ell = 0, 1$). In other words, we scale the case weights and control weights separately so that the sum of the case weights is proportional to λ_1 and the sum of the control weights is proportional to λ_0 and put them, along with the usual specification of the design structure (strata, primary sampling units), into our program of choice. (This is a little oversimplified but adequate for most practical applications – see Section 2.5). If $\sum_{i \in s_\ell} w_i = N_\ell$, then $w_i^* \propto \frac{\lambda_\ell w_i}{N_\ell}$. We assume this to be true in the rest of this chapter.

We still have to decide on good values for λ_1 and λ_0. We can get consistent, and sometimes very large, gains using $\lambda_\ell = \frac{n_\ell}{n}$ (i.e., $w_i^* \propto \frac{n_\ell w_i}{N_\ell}$), which are the maximum likelihood weights for simple random sampling – we shall call these pseudo maximum likelihood (PML) weights – compared to using design weights, $\lambda_\ell = W_\ell$ (i.e., $w_i^* \propto w_i$). Scott and Wild (2002) report efficiency gains of over 600% with a 1:300 case–control ratio and a single regression variable, x. The gains became larger as strength of the relationship between Y and x increased, and as the effect of clustering increased. Moreover, the coverage of confidence intervals was closer to the nominal value for PML weighting than for design weighting in the simulations. This is consistent with having a larger effective sample size.

These simulation results need to be treated with some caution, however. Korn and Graubard (1999, p. 306) comment that, in their experience, the PML weighting strategy rarely produces quite such big gains in efficiency in practice and the empirical results for the meningitis study in the next section give some support to this comment. It seems that the gains may depend on the particular problem under examination. More empirical work is needed here and, until this has been done, it seems prudent to fit the model using both PML weights and design weights routinely. If the coefficient estimates are similar, then we can make a judgement based on the estimated standard errors. However, substantial differences in the coefficient estimates may indicate that the model has been misspecified. If we are unable to fix up the deficiencies in the model, then we need to think very carefully about just what it is that we are trying to estimate. We look at this again in Section 2.5.

Using PML weights is the most efficient possible strategy when we have simple random samples of cases and controls but this is not necessarily true for more complex schemes. We might, for example, expect weights based on some form of equivalent sample sizes to perform better. We have done some limited simulation and this does

indeed produce some gain in efficiency. However, the gains are relatively small, at least when the control sample design effect is less than 2, because Cov $\{\widehat{\boldsymbol{\beta}}_\lambda\}$ is very flat as a function of λ near its minimum. Considerations of robustness that we discuss in the next subsection are possibly more important in the choice of λ.

Although the weighted methods discussed in this subsection can cope with any form of stratified sampling, we can gain still more efficiency in situations where the same design variables are used in constructing sampling strata, or for poststratification, in both the case and control samples. We return to this in Section 3.

2.4. *Example: Auckland Meningitis Study*

We illustrate by fitting a logistic model for the probability of contracting meningitis during the 3-year period of the study to the data collected from the Auckland Meninigitis Study (Example 2), using the DW and PML methods. More details of the study are given in Baker et al. (2000). Estimated coefficients and their standard errors are given for the modifiable risk factors in the table later. (Note that our model differs slightly from that in the paper because some variables are no longer available.)

The estimates from the two schemes lead to the same general conclusions, with the PML standard errors about 12% smaller than their DW equivalents on average. The results from both methods suggest that household overcrowding (as measured by the number of adults per room) and frequent attendance at large social gatherings are both highly significant factors.

If we calculate the efficiency ($e = 1/(1+c^2)$) for the two weighting schemes, we get $e = 0.31$ for design weights and $e = 0.63$ for PML weights. This would correspond to a 30% reduction in the standard errors with PML weighting compared to design weighting. The reduction actually achieved is considerably less than this. Note that the efficiency of the PML scheme is still relatively low here, with a fair amount of variability in the control weights, which suggests that it might be possible to do better still. We look at this again in Section 3.

If we had ignored the clustering, the calculated standard errors would be reduced by about 5% so the effect of intracluster correlation is fairly small here. In part, this is because the average cluster size was kept small deliberately in this study to minimize the design effect. This made the sampling more expensive, of course, and it may have been more efficient to sample more children in fewer clusters.

Table 1
Comparison of PML and DW estimates

Risk factor[a]	PML(s.e.)	DW(s.e.)
No. of adults/room	2.30 (0.56)	2.29 (0.57)
Attends substantial social gatherings	0.37 (0.16)	0.73 (0.25)
No. of smokers in usual HH	0.20 (0.13)	0.26 (0.19)
Shares food, drink or pacifier	0.39 (0.26)	0.24 (0.25)
Respitory infection in HH member	0.41 (0.25)	0.59 (0.28)
Bed sharing	0.51 (0.26)	0.45 (0.28)

[a] Base model includes age, ethnic group, year and month of interview, tertiary education of a parent, and possession of a community services card.

2.5. Reweighting and robustness

The analyses of the previous sections assume that the model we are using is true. Given the trial-and-error processes used in arriving at a model, this will almost never be the case. DW analyses have a certain type of robustness against model misspecification. The question arises as to what price in terms of robustness do we pay for the gains in efficiency made by using PML rather than design weights. How compelling the type of robustness that design weighting confers in practical applications is also a question that merits investigation.

By its construction, the DW estimator always estimates the linear-logistic approximation that we would get if we had data from the whole population. By contrast, what the more efficient PML-weighted estimator is estimating depends on the particular sample sizes used. We suspect that very few people would regard it as completely satisfactory to have the target of their inference depend on the arbitrary choice of sample size.

Our general estimator $\widehat{\boldsymbol{\beta}}_\lambda$ satisfying (9) converges to the solution of equation (8), $\boldsymbol{B}_\lambda$ say, with $\lambda = \lambda_0/(\lambda_0 + \lambda_1)$, which depends on the true model and distribution of the covariates, as well as on λ. Scott and Wild (2002) examined what happens to $\boldsymbol{B}_\lambda$ under mild deviations from the assumed model. (Interest is centred on small-to-mild deviations since large ones should be picked up by routine model-checking procedures and the model then improved.) For simplicity, suppose that we fit a linear model with a single explanatory variable for the log odds ratio but that the true model is quadratic, say

$$\text{logit}\{P(Y = 1|x)\} = \beta_0 + \beta_1 x + \delta x^2 \tag{13}$$

with δ small.

The slope of the function, $\beta_1 + 2\delta x$, changes as we move along the curve and $B_{\lambda 1}$ is equal to the actual slope at some point along the curve for **any** $0 < \lambda < 1$. Denote this value by $x = x_\lambda$. Let x_0 be the expected value of x in the control population and let x_1 be the expected value of x in the case population. We shall assume that $\beta_1 > 0$ so that $x_0 < x_1$. It turns out that x_λ always lies between x_0 and x_1 and that x_λ increases as λ increases from 0 to 1 (see Scott and Wild, 2002, for details of the proof). Recall that design weighting corresponds to $\lambda = W_0$ and PML weighting to $\lambda = \omega_0 = n_0/n$. Typically, W_0 is much larger than ω_0 so that design weighting gives an estimate of the slope at larger values of x, where the probability of a case is higher, while the slope estimated from PML weighting is closer to the slope at the average value of x in the population. Figure 1, adapted from Scott and Wild (2002), illustrates the position in two scenarios, one with positive curvature and one with negative, based roughly on the meningitis example of the previous section. The value of δ is chosen so that it would be detected with a standard likelihood ratio test about 50% of the time if we took simple random samples of $n_0 = n_1 = 200$ from the population.

In both scenarios, the value of β_0 is set so that the proportion of cases in the population is 1 in 400, so that $W_0 = 0.9975$. The overall density of x is shown at the top of the graph and the conditional densities for cases and controls are shown at the bottom. Values of x_λ and $B_{\lambda 1}$ are shown for $\lambda = W_0$ (labeled "population") and $\lambda = 0.5$ (labeled "equal"). The latter value corresponds to PML weighting if we draw equal numbers of cases and controls. Clearly, design weighting is estimating the appropriate slope for values of x further out in the upper tail of the distribution (i.e., for individuals at higher risk) than equal weighting in both scenarios.

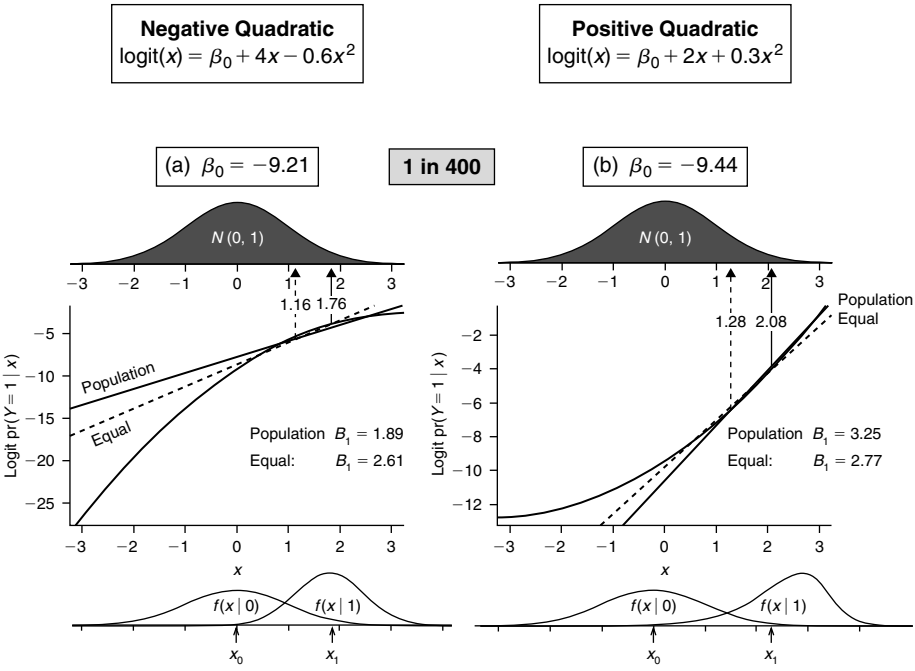


Fig. 2. Comparison of population and equal weights.

If we take simple random samples of $n_0 = n_1 = 200$ from the population in Fig. 2a, it turns out that the relative efficiency of design weighting is only about 16%, and the small sample bias is 0.24. In this case, even if we take the population value as our target, design weighting leads to a larger mean squared error than PML weighting.

More results are given in Scott and Wild (2002) where we also look at the effect of omitted covariates. This turns out to have a similar, but somewhat smaller, effect to omitting a quadratic term.

Which is the right value of λ to use? That clearly depends on what we want to use the resulting model for. If our primary interest is in using the model for estimating odds ratios at values of x where the probability of a case is higher, and the sample is large enough so that variance and small sample bias are less important, we might use design weights. For smaller sample sizes, or if we are interested in values of x closer to the population mean, PML weights would be better. A value intermediate between design weighting and PML weighting might sometimes be a sensible compromise. For example trimming the weights to 10:1 (i.e., setting $\lambda \approx 0.91$) in the example, instead of 1:1 (PML weighting) or 400:1 (design weighting), leads to an efficiency of 70% and a small sample bias of 0.04. The corresponding values for design weighting were 16% and 0.24. The value of $x_{0.91}$ lies almost exactly half way between $x_{0.5}$ and $x_{0.9975}$.

2.6. Variance estimation for super-population parameters

Graubard and Korn (2002) point out that we have to take some care in deriving the properties of DW estimators of $\boldsymbol{\beta}$ since the weights employed, $w_i = N_i/n_i$, involve

N_0 and N_1 which are random variables in the super-population framework adopted here rather than fixed constants as in the standard finite population set-up. In sampling terminology, we can think of our situation as being equivalent to two-phase sampling as depicted in Fig. 1. In the first phase, the finite population is generated as a random sample of size N from an (infinite) super-population and N_0 and N_1 are recorded. In the second phase we draw a simple random samples of size n_i from the N_i units in the $Y = i$ stratum ($i = 0, 1$), with the values of n_0 and n_1 depending only on N_0 and N_1, and we observe $\boldsymbol{x}$. In Scott and Wild (2007) we use the results for two-phase sampling developed by Rao (1973) to show that we can estimate $\mathrm{Cov}(\widehat{\boldsymbol{\beta}}_W)$ in a way that validly accounts for this additional source of variation using

$$\widehat{\mathbf{J}}^{-1}\left(\sum_i W_i^2 \widehat{\mathrm{Cov}}\,\{\widehat{\boldsymbol{\mu}}_i\}\right)\widehat{\mathbf{J}}^{-1} + \frac{1}{N}\widehat{\mathbf{J}}^{-1}\left(\sum_i W_i\{\widehat{\boldsymbol{\mu}}_i - \widehat{\boldsymbol{\mu}}\}\{\widehat{\boldsymbol{\mu}}_i - \widehat{\boldsymbol{\mu}}\}^T\right)\widehat{\mathbf{J}}^{-1},$$

where $\widehat{\mathbf{J}} = \widehat{\mathbf{J}}(\widehat{\boldsymbol{\beta}})$, $\widehat{\boldsymbol{\mu}}_i = \widehat{\boldsymbol{\mu}}_i(\widehat{\boldsymbol{\beta}})$, and $\widehat{\boldsymbol{\mu}} = \sum W_i\widehat{\boldsymbol{\mu}}_i / \sum W_i$. The first term, which is of order $\frac{1}{n}$, is the variance estimate we would use if we assumed that the N_i were fixed and is what we get out of a standard survey regression program. The second term measures the effect of not knowing the N_i in advance. This second term, which is of order $\frac{1}{N}$, will be negligible in most applications.

For other choices of λ_i in (7), such as PML weights $\lambda_i = n_i/n$, the estimating equations do not depend on the N_i. The estimates of all coefficients in $\boldsymbol{\beta}_1$ and their covariances are thus unaffected and the variance estimate for $\widehat{\beta}_0$ can be corrected by adding $(1/N_1 + 1/N_2)$. This extra variance component results from the estimated correction term, $\widehat{b}_\lambda = \log\{\lambda_1 W_0/(\lambda_0 W_1)\} = \text{const.} + \log(N_0/N_1)$, which has to be subtracted to correct the estimate of β_0 obtained from the program output.

2.7. Related designs: Case-augmented studies

Case-augmented sampling designs, represented in Fig. 3, are closely related to the simple case–control design and can be treated as such when cases are rare. Here, a sample of cases (represented in Fig. 3 as the lower sample) is supplemented by an independent sample from the parent population or process. In "Design 1" (called "case-supplemented" by Cosslett, 1981) only information on $\boldsymbol{x}$ is collected from the whole-population sample and Y is unobserved. In "Design 2" (called "case-enriched" by Cosslett, 1981) both Y and $\boldsymbol{x}$ are observed for the whole-population sample.

Obviously there is very little difference between the designs if cases are rare in the population, but one of the advantages of either of the case-augmented designs is that they allow us to estimate relative risks without invoking the "rare disease" assumption.

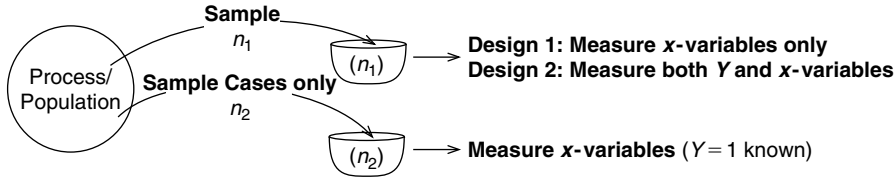


Fig. 3. Case-augmented sampling.

For, with a simple application of Bayes Theorem, we can write the risk at covariate value $\boldsymbol{x}$ relative to that at some baseline, $\boldsymbol{x}_0$, as

$$\frac{\text{pr}(Y=1 \mid \boldsymbol{x})}{\text{pr}(Y=1 \mid \boldsymbol{x}_0)} = \frac{g(\boldsymbol{x} \mid Y=1)}{g(\boldsymbol{x}_0 \mid Y=1)} \Big/ \frac{g(\boldsymbol{x})}{g(\boldsymbol{x}_0)},$$

where $g(\boldsymbol{x})$ denotes the marginal density of $\boldsymbol{x}$, etc. Clearly, the first likelihood-ratio on the right-hand side can be estimated directly from the case sample and the second directly from the whole-population sample.

Examples of case-augmented sampling are widespread in medical studies. In fact, many population-based case–control studies are really case-augmented designs, with the cases in the reference sample either being transferred to the case sample (when case status is recorded) or treated as controls otherwise. Rothman and Greenland (1998) advocate the use of (what are essentially) case-augmented designs, partly because of the ability to estimate relative risks directly, but more fundamentally because it helps ensure that controls are drawn from the same population as the cases. Design 2 is formally equivalent to a case-cohort study, where a simple random sample of a larger cohort is selected for more intensive investigation at the beginning of the study and the cases are added as they occur. The main emphasis in case-cohort studies is almost always on survival analysis but, in his seminal paper on the subject, Prentice (1986) also considered fitting a logistic regression to a binary response, and showed that the maximum likelihood estimates of the regression coefficients (apart from the constant term) are obtained by fitting an ordinary prospective model to the pooled sample, just as in ordinary case–control studies.

Case-augmented sampling is also used in other fields. For example, Millar (1992) describes a standard design in fisheries for investigating the size-selectivity of fishing gear. In such studies, two nets are dragged behind a trawler. One net has a fine mesh that lets no fish escape while the other has a coarser, test mesh designed to let smaller, immature fish escape. The object of the experiment is to estimate the relationship between fish size (as measured by length) and the probability of capture in the coarse test-net. This is an example of the first design (case-supplemented), in which the fish caught in the coarse test-mesh are assumed to be a random sample of capture-cases ($Y = 1$) with their size information ($\boldsymbol{x}$) being observed, and the fine mesh net being regarded as a random sample from the whole fish population giving us information on the marginal distribution of fish sizes but no information on Y. An example of the second (case-enriched) design in ecology, is discussed by Manly et al. (2002). They consider a study to determine the factors that influence the selection of nest sites in fernbirds. Covariates were measured on 24 nest sites and also on a random selection of 25 possible sites (clumps of vegetation) in the same area. Several scenarios where case-augmented designs are used in econometrics are discussed by Cosslett (1981).

When the case and reference samples are both simple random samples, efficient semiparametric maximum likelihood procedures for both designs are derived in Lee et al. (2006) using the profile likelihood obtained by maximizing over the unknown distribution of $\boldsymbol{x}$. These procedures are relatively simple to implement when the logistic model (1) applies. For case-supplemented sampling (Design 1), the solution can be found by treating all observations from the whole-population sample as controls and running the data through a standard prospective binary regression program using the model p_1^* where, for the logistic model (1), $p_1^*(\boldsymbol{x}_i; \boldsymbol{\phi}) = \frac{e^{\rho + \boldsymbol{x}^T \boldsymbol{\beta}}}{1 + e^{\boldsymbol{x}^T \boldsymbol{\beta}} + e^{\rho + \boldsymbol{x}^T \boldsymbol{\beta}}}$ with ρ an additional nuisance

parameter. For case-enriched sampling (Design 2), the procedure is even simpler: we simply run the combined sample through a standard logistic regression program with a fixed offset, $\widehat{\rho} = \log(1 + n_2/n_{1c})$ where n_2 is the size of the case-only sample and n_{1c} is the number of cases in the whole-population sample.

In practice, it is relatively common to use a more complex survey design to draw the whole-population sample in particular. In conventional survey terms, Design 2 can be regarded as a dual frame survey, with one frame for the whole population and the other for the subpopulation of cases. The methods of Chapters 24 and 26 can be then used to estimate the census estimating equations, $\boldsymbol{S}(\boldsymbol{\beta}) = \mathbf{0}$, given in (2), and we can proceed just as in Section 2.2. Design 1 is more difficult to handle in a conventional survey framework because $\boldsymbol{S}(\boldsymbol{\beta})$ is not estimable from the resulting samples. The usual strategy, if cases are rare, is to pretend that all observations in the whole-population sample are controls and proceed as if we had a standard case–control sample.

The semiparametric maximum likelihood estimators are relatively simple to adapt to more complex designs. In both cases, the estimator satisfies equations of the form

$$\widehat{\boldsymbol{S}}_{\omega}(\boldsymbol{\beta}, \rho) = \omega_1 \frac{\sum_{\text{Sample 1}} \boldsymbol{U}_{1i}(\boldsymbol{\beta}, \rho)}{n_1} + \omega_2 \frac{\sum_{\text{Sample 2}} \boldsymbol{U}_{2i}(\boldsymbol{\beta}, \rho)}{n_2} = \mathbf{0}, \tag{14}$$

where $\omega_\ell = \frac{n_\ell}{n}$, for $\ell = 1, 2$ and the form of $\boldsymbol{U}_\ell$ depends on the design (see Lee et al., 2006 for details). For case-supplemented sampling (Design 1)

$$U_1 = \frac{\partial \log p_0^*(x, \boldsymbol{\phi})}{\partial \boldsymbol{\phi}} \quad \text{and} \quad U_2 = \frac{\partial \log p_1^*(x, \boldsymbol{\phi})}{\partial \boldsymbol{\phi}},$$

where $p_1^*(\boldsymbol{x}_i; \boldsymbol{\phi}) = \frac{e^\rho p_1(\boldsymbol{x}_i; \boldsymbol{\beta})}{1 + e^\rho p_1(\boldsymbol{x}_i; \boldsymbol{\beta})}$.

For case-enriched sampling (Design 2)

$$U_1 = \frac{\partial \left[(1 - y) \log p_0^{**}(\boldsymbol{x}, \boldsymbol{\phi}) + y \log p_1^{**}(\boldsymbol{x}, \boldsymbol{\phi}) - y\rho\right]}{\partial \boldsymbol{\phi}} \text{ and}$$

$$U_2 = \frac{\partial \left[\log p_1^{**}(\boldsymbol{x}, \boldsymbol{\phi}) + \log(e^\rho - 1) - \rho\right]}{\partial \boldsymbol{\phi}},$$

where p_1^{**} is logistic with logit $p_1^{**}(x_i; \boldsymbol{\phi}) = \rho + \beta_0 + \boldsymbol{x}_i^T \boldsymbol{\beta}_1$. If we write $\boldsymbol{\phi} = (\boldsymbol{\beta}^T, \rho)^T$ then, when model (1) is true, $\boldsymbol{\phi} = (\boldsymbol{\beta}^T, \rho)^T$ is the solution of

$$\boldsymbol{S}_{\omega}(\boldsymbol{\phi}) = \omega_1 E\left\{\boldsymbol{U}_1(\boldsymbol{\phi})\right\} + \omega_2 E_1\left\{\boldsymbol{U}_2(\boldsymbol{\phi})\right\} = \mathbf{0}, \tag{15}$$

where $E\{\cdot\}$ denotes the expectation over the joint distribution of $(Y, \boldsymbol{x})$, $E_1\{\cdot\}$ denotes the conditional expectation given that $Y = 1$. (The interpretation of ρ differs slightly between the two designs but is basically the log of the relative sampling rates of cases and controls. This means that its value depends upon the choice of ω_ℓ, $\ell = 1, 2$.)

Noting that (15) is a sum of population means, we can adapt (14) and (15) to more complex sampling schemes, just as we did with case–control sampling, using an estimator satisfying

$$\widehat{\boldsymbol{S}}_{\omega}(\boldsymbol{\phi}) = \omega_1 \widehat{\boldsymbol{\mu}}_1(\boldsymbol{\phi}) + \omega_2 \widehat{\boldsymbol{\mu}}_2(\boldsymbol{\phi}) = \mathbf{0}, \tag{16}$$

where $\widehat{\boldsymbol{\mu}}_1(\boldsymbol{\phi})$ is the sample estimate of the population mean $E\{\boldsymbol{U}_1(\boldsymbol{\phi})\}$ and $\widehat{\boldsymbol{\mu}}_2(\boldsymbol{\phi})$ is an estimate of the case-stratum mean $E_1\{\boldsymbol{U}_2(\boldsymbol{\phi})\}$. The estimated "sandwich" covariance matrix is given by

$$\widehat{\mathrm{Cov}}\left\{\widehat{\boldsymbol{\phi}}_\omega\right\} \approx \mathbf{J}_\omega(\widehat{\boldsymbol{\phi}}_\omega)^{-1}\widehat{\mathrm{Cov}}\left\{\widehat{\boldsymbol{S}}_\omega(\widehat{\boldsymbol{\phi}}_\lambda)\right\}\mathbf{J}_\omega(\widehat{\boldsymbol{\phi}}_\omega)^{-1},$$

with $\mathbf{J}_\omega(\boldsymbol{\phi}) = \left(-\frac{\partial \widehat{\boldsymbol{S}}_\omega(\boldsymbol{\phi})}{\partial \boldsymbol{\phi}^T}\right)$ and $\widehat{\mathrm{Cov}}\left\{\widehat{\boldsymbol{S}}_\omega(\boldsymbol{\phi})\right\} = \omega_1^2\widehat{\mathrm{Cov}}\{\widehat{\boldsymbol{\mu}}_1(\boldsymbol{\phi})\} + \omega_2^2\widehat{\mathrm{Cov}}\{\widehat{\boldsymbol{\mu}}_2(\boldsymbol{\phi})\}$ as the samples in Fig. 3 are taken independently. Here, $\widehat{\mathrm{Cov}}\{\widehat{\boldsymbol{\mu}}_\ell(\boldsymbol{\phi})\}$ denotes the usual survey estimate.

This is a relatively new area and much more work needs to be done on the relative performance of these procedures.

2.8. Practical considerations

Case–control studies are observational studies and, as such, are subject to the biases and difficulties of interpretation common to all observational studies. The historical nature of the data means that problems with measurement errors, selection biases, and missing data, especially when patterns of missingness are different for cases and controls, are all likely to be more than usually acute. The possibility of important unmeasured confounders means that we have to interpret any observed associations with great caution. Breslow (2005) has a very good account of the potential pitfalls and of steps that can be taken to minimize their effects. Although Breslow (2005) is written in the context of epidemiology, many of the considerations discussed there are critical in any case–control study in which we wish to model the effects of $\boldsymbol{x}$-variables on a subsequent binary response.

Obtaining controls and cases from surveys can introduce additional problems. If different frames are used for the two surveys then it is vital to make sure that the survey populations (as distinct from the target populations) are as similar as possible. For example, if cases are drawn from a register and controls are obtained by random digit dialing, then we would need to exclude cases that are not contactable by telephone. There are special problems if controls are obtained from a survey that was originally designed for some other purpose. Often the surveys will measure different sets of covariates and, even when the same covariate appears to be measured in both surveys, the definitions may differ significantly between the two. An excellent account of the special difficulties with survey data is given in Chapter 9 of Korn and Graubard (1999). On the positive side, survey statisticians are particularly sensitive to the problems of frame errors, nonresponse, measurement and other nonsampling errors and have developed reasonably effective methods for mitigating their effects. This means that many of the worst pitfalls would be avoided routinely.

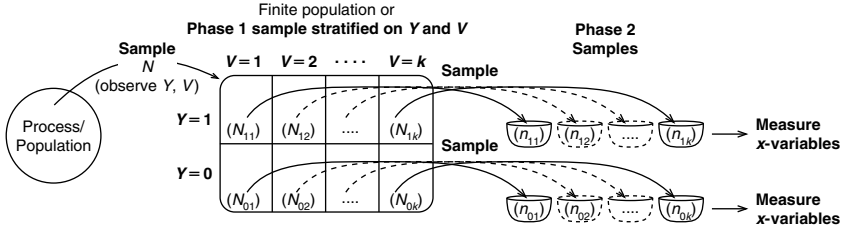
We note that there has been some waning of enthusiasm for traditional case–control studies among epidemiologists in recent years with the increasing interest in molecular markers for disease. If the disease itself affects the markers, then retrospective determination of their values is of limited usefulness. Instead, there has been more emphasis in storing prediagnostic samples from cohorts or population-based surveys and then using incidence-density sampling or case-cohort sampling. Breslow (2005) gives a good introduction to such studies. We do not discuss incidence-density sampling, which is a special case of matched sampling, but some aspects of case-cohort studies are covered in Section 2.7.

3. Two-phase case–control sampling

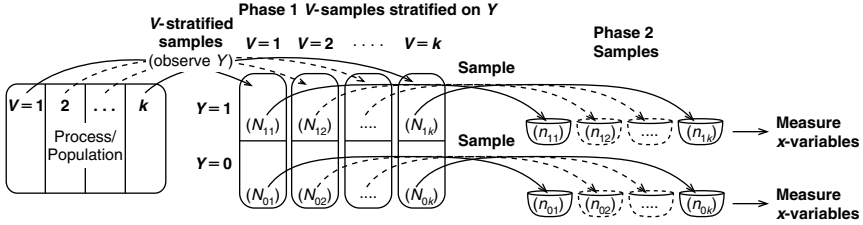
3.1. Motivation

Two-phase case–control sampling (more commonly called two-stage case–control sampling in the biostatistics literature) is used to describe any of the three sampling designs depicted in Fig. 4. The reweighting compromise suggested in Section 2.3 (i.e., use standard design weighting within subpopulations defined by case/control status but combine the sub-populations using sample proportions) seems to work reasonably well in practice in any of these situations, but we can do better if we take account of the special structures shown in Fig. 4. Before going on to discuss analysis, however, we will pause to motivate these designs and their growing importance in biostatistics. Virtually all of the theory to date has been developed for simple stratified random sampling at each phase.

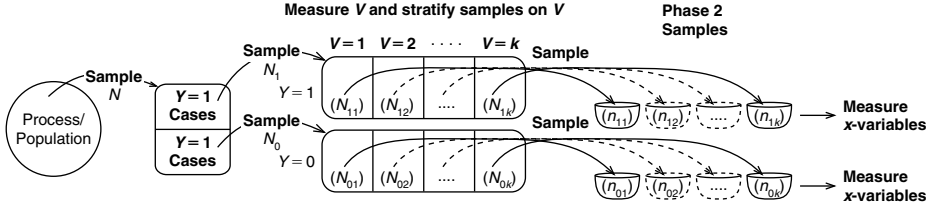
The two-phase (two-stage) case–control design was introduced by White (1982) as a design for studying an association between a binary response Y and a binary exposure



(a) Random first-phase



(b) *V*-stratified first-phase



(c) *Y*-stratified first-phase

Fig. 4. "Two-phase" case–control sampling.

variable V (in our notation) adjusted for discrete covariates. Motivated by considerations of cost-effectiveness, she proposed taking separate samples at phase two from the individuals in each of the four cells of the 2×2 cross-classification of Y and V, and determining covariate information only for the subsampled individuals. She proposed over-sampling small cells, for example, by taking equal sized subsamples from each of the four cells. She noted that the first-phase $Y \times V$-data could itself come either from case–control sampling (Fig. 4c), or be from a cohort or cross-sectional study (either could be of the form Fig. 4b, c).

By the end of the 1980s, following the work of Fears and Brown (1986), Breslow and Cain (1988), Cain and Breslow (1988), and Breslow and Zhao (1988), methods of analysis had been developed that could handle situations where $\boldsymbol{x}$ included continuous covariates and V was included as a linear term, as opposed to a set of categories, in the regression model. Indeed we could have a vector $\boldsymbol{V}$ of variables defining the V-strata, provided all the variables were discrete. Cain and Breslow (1988) outlined a number of situations where these methods can be useful. These include:

Efficiency: Cost savings can be obtained using a genuine two-phase design (e.g., Engels et al., 2005) and measuring covariates that are particularly expensive or particularly invasive only for comparatively small subsamples. Such studies are becoming increasingly useful, particularly as expensive new techniques for extracting genetic information become more and more widely available.

Secondary analysis: Adding a second-phase sample provides a cost-effective way of making an after-the-fact adjustment for a confounder that was not considered in the original single-phase study.

Incorporating "whole population" information: There may be administrative or other population $Y \times V$-data available for all individuals in the finite population(s) from which the cases and controls in a single-phase study were drawn. Efficiency can be increased by considering the finite-population data as the first-phase and the study data as the second phase.

Missing data: If, in a single-stage study, there are substantial numbers of missing values among the covariates and we are willing to assume that they are missing at random given Y- and the V-variables, then we can treat the data as coming from a two-phase study with those for which $\boldsymbol{x}$ is observed considered as subsamples within $Y \times V$-cells. This is more defensible than a complete-case analysis, especially when the "missingness" rates differ appreciably between the cells.

By making proper use of stratum-specific offsets, prospective logistic-regression programs can be used to obtain valid estimates of the parameters of a logistic regression (Fears and Brown, 1986) fitted to data from a two-phase study in the full generality described earlier. Substantial work is needed to correct the standard errors, however, and the procedure is not in general either maximum likelihood (Breslow and Cain, 1988; Breslow and Zhao, 1988) or efficient (Scott and Wild, 1991). Semiparametric maximum likelihood estimation for two-phase studies with a prospective first phase was developed for general models by Scott and Wild (1991, 1997, 2001b), whereas Breslow and Holubkov (1997) worked with logistic models and developed semiparametric maximum likelihood for a case–control first phase. They noted that the resulting estimator was the same as the Scott and Wild (1997) estimator. This means that, just as for simple case–control studies, when logistic models are fitted to two-phase data,

whether the first phase is prospective or case–control only affects the overall intercept β_0. Semiparametric efficiency was established by Breslow et al. (2003) for a prospective first phase and random sample-size subsampling mechanism, and more generally in Lee et al. (2007).

3.2. Analysis

As previously stated, the reweighting compromise suggested in Section 2.3 when we have complex sampling (i.e., use standard design weighting within sub-populations defined by case/control status but combine the sub-populations using sample proportions) seems to work reasonably well in practice but we can do better for two-phase case–control studies when simple random samples are drawn within strata using methods described in Scott and Wild (2001b) that are reasonably easy to implement.

We motivate the methods by starting with the situation in which our model has a separate coefficient (intercept) term for each level of the stratifying phase one variable V. In other words we have

$$\text{logit}\,\{\text{pr}(Y = 1|\boldsymbol{x}, \text{Stratum } h)\} = \beta_{0h} + \boldsymbol{x}^T\boldsymbol{\beta}_1, \tag{17}$$

and phase two sampling is of the form shown in Fig. 4a, b with simple stratified sampling, rather than a more complex design. (Fig. 4c differs very slightly, but in an easily corrected way.) Here fully efficient procedures are well-developed and easy to implement. Ordinary unweighted logistic regression (with a simple adjustment for the stratum intercepts if they are wanted) is the efficient semiparametric maximum likelihood procedure (Prentice and Pyke, 1979). The estimating equations can be written in the form

$$\sum_h \left(\omega_{1h} \frac{\sum_{\text{cases}} \boldsymbol{x}_i p_{0h}(\boldsymbol{x}_i; \boldsymbol{\beta})}{n_{1h}} - \omega_{0h} \frac{\sum_{\text{controls}} \boldsymbol{x}_i p_{1h}(\boldsymbol{x}_i; \boldsymbol{\beta})}{n_{0h}} \right) = \mathbf{0},$$

where $\omega_{ih} \propto n_{ih}$. The stratified equivalent of the estimating equation (7) is

$$\sum_h \left(\lambda_{1h} \frac{\sum_{\text{cases}} \boldsymbol{x}_i p_{0h}(\boldsymbol{x}_i; \boldsymbol{\beta})}{n_{1h}} - \lambda_{0h} \frac{\sum_{\text{controls}} \boldsymbol{x}_i p_{1h}(\boldsymbol{x}_i; \boldsymbol{\beta})}{n_{0h}} \right) = \mathbf{0}. \tag{18}$$

and it is straightforward to extend this to more general stratified designs. As $n_{0h}, n_{1h} \to \infty$, the solution of (18) converges almost surely to the solution of

$$\sum_h [\lambda_{1h} E_{1h}\{\boldsymbol{X} p_{0h}(\boldsymbol{X}; \boldsymbol{\beta})\} - \lambda_{0h} E_0\{\boldsymbol{X} p_{1h}(\boldsymbol{X}; \boldsymbol{\beta})\}] = \mathbf{0}, \tag{19}$$

with the obvious extension of the notation from the unstratified case. If model (17) is true, then equation (19) has solution $\boldsymbol{\beta}_1^* = \boldsymbol{\beta}_1$ and $\beta_{0h}^* = \beta_{0h} + b_{\lambda h}$ with $b_{\lambda h} = \log\left(\frac{\lambda_{1h} W_{0h}}{\lambda_{0h} W_{1h}}\right)$. Because equation (19) only involves stratum means, we can estimate them using survey data, for example by

$$\widehat{\boldsymbol{\mu}}_{\ell h}(\boldsymbol{\beta}) = \frac{\sum_{i \in S_{\ell h}} w_{ih} \boldsymbol{x}_{ih}\{y_{ih} - p_1(\boldsymbol{x}_{ih}; \boldsymbol{\beta})\}}{\sum_{i \in S_{\ell h}} w_{ih}}.$$

Substituting these estimators in place of the sample means in equation (18) leads to the estimating equation

$$\widehat{\boldsymbol{S}}_\lambda(\boldsymbol{\beta}) = \sum_h \sum_{i \in S_h} w^*_{ih} \boldsymbol{x}_i \{y_i - p_{1h}(\boldsymbol{x}_i; \boldsymbol{\beta})\} = \boldsymbol{0}, \tag{20}$$

with $w^*_{ih} \propto \lambda_{\ell h} w_{ih} / \sum_{i \in S_{\ell h}} w_{ih}$ for units in $S_{\ell h}$ ($\ell = 0, 1; h = 1, \ldots, H$). This can be fitted in any standard survey program by including these weights and the appropriate design information. Note that we need to be careful about how we include the so-called strata in the design specification. If primary sampling units are nested within the "strata," there is no problem and the strata should be included in the standard way. However, if the primary sampling units cut across the strata, like the age groups in our first example or the age × ethnic groups in our second example, then these are not strata in the usual survey sampling sense. They should not be included in the design specifications but simply handled through the weights.

Sometimes we want to model the contribution of the stratum variables using some smooth parametric curve rather than dummy variables. For example, age is a common stratifying variable in many studies and we might want to include a linear function of age in our model. The DW and PML methods suggested in Section 2 both apply directly and no new theory is needed. More efficient methods are not nearly so simple, however. Fully efficient methods have been developed in the case where simple random samples of cases and controls are drawn within each of the strata (see Breslow and Holubkov, 1997; Scott and Wild, 1997), but the resulting estimating equations are not linear combinations of stratum means and there is no obvious way of generalizing them to more complex sampling schemes. There is a slightly less efficient way that does extend easily, however. If we modify model (17) by including $b_{\lambda h} = \log\{\lambda_{1h} W_{0h} / (\lambda_{0h} W_{1h})\}$ as an offset (a known additive term), that is we set

$$\text{logit}\left\{\text{pr}^*(Y = 1 | \boldsymbol{x}, \text{Stratum } h)\right\} = b_{\lambda h} + \beta_{0h} + \boldsymbol{x}^T \boldsymbol{\beta}_1, \tag{21}$$

then equation (18) produces consistent, fully efficient, estimates of all the coefficients including β_{0h} ($h = 1, \ldots, H$). Including the same offsets in models where there is no β_{0h} term and the $\boldsymbol{x}$ vector includes functions of the stratifying variable produces consistent estimators of all the coefficients with high (although not full) efficiency (see Breslow and Cain, 1988; Fears and Brown, 1986). This generalizes to arbitrary designs immediately. We just use equation (20) with p_{1h} replaced by p^*_{1h} defined by setting $\text{logit}(p^*_{1h}) = b_{\lambda h} + \boldsymbol{x}^T \boldsymbol{\beta}$. Then any survey program that caters for offsets can be used to fit the model and provide estimated standard error, etc.

How much extra efficiency over and above that already gained using PML weighting do we get in this case? We have carried out a number of simulations, some of which are reported in Scott and Wild (2002). Without any clustering, the gain in efficiency from using the offset method (which is full maximum likelihood here) compared to the PML procedure was never more than 10%. The relative efficiencies stayed about the same as clustering that cut across strata was introduced. When clustering nested within strata was introduced, the gains disappeared progressively as the design effect increased and the PML procedure actually became more efficient than the offset method when the design effect reached about 1.5.

Table 2
Comparison of PML and offset (OS) estimates

Risk factor	PML(s.e.)	OS(s.e.)
No. of adults/room	2.30 (0.56)	1.93 (0.47)
Attends substantial social gatherings	0.37 (0.16)	0.38 (0.16)
No. of smokers in usual HH	0.20 (0.13)	0.25 (0.12)
Shares food, drink or pacifier	0.39 (0.26)	0.36 (0.22)
Respitory infection in HH member	0.41 (0.25)	0.37 (0.22)
Bed sharing	0.51 (0.26)	0.49 (0.24)

The meningitis study of Example 2 can be regarded as a two-phase study in which age and ethnicity are measured at Phase 1 and the remaining variables measured on a subsample stratified by these variables at Phase 2. Our base model has a linear term in age and no age $\times$ ethnic group interaction so it does not include a complete set of dummy variables for the first-phase strata. The effect of clustering was small here, with a design effect of about 1.1, and the control weights were more variable than those in our simulations, so we might expect somewhat larger improvements than we saw there. This turns out to be the case. Table 2 shows the results for the offset approach fitted to the model considered in Table 1, with the PML values repeated for comparison.

We see that the offset method does indeed work well here, giving a further 12% reduction in the estimated standard errors on average.

As we stated earlier, it is possible to produce fully efficient semiparametric estimators if we are willing to model the dependence structure within primary sampling units. We have begun to carry out some simulations using random effects model. The early results suggest that the extra work involved in the modeling will almost never be worth the effort if we are only interested in the parameters of the marginal model (1). Our tentative conclusion is that, the ad hoc partially weighted procedures (with PML weights) are simple to use and work well enough for most practical purposes in the range covered by our experience but this is another area where more empirical work is needed yet. We note, however, that there are some problems, like the case–control family design discussed in Section 4, where the within-cluster behavior is of interest in its own right. These require more sophisticated methods.

4. Case–control family studies

If we are primarily interested in the parameters of the marginal model (1), then the methods that we have discussed in previous sections are simple to implement and reasonably efficient. With cluster or multistage sampling, fully efficient methods require building parametric models for the within-cluster dependence and the extra effort that this would entail is rarely worthwhile. However, there are situations where the dependence structure is of interest in its own right. For example, it has become increasingly common for genetic epidemiologists to augment data from a standard case–control study with response and covariate information from family members, in an attempt to gain information on the role of genetics and environment. This can be regarded as a stratified

cluster sample, with families as clusters, and the intracluster structure is of the primary focus of attention here. The following example is fairly typical.

Wrensch et al. (1997) conducted a population-based case–control study of glioma, the most common type of malignant brain tumor, in the San Francisco Bay Area. They collected information on all cases of glioma that were diagnosed in a specified time interval and on a comparable sample of controls obtained through random digit dialing. They also collected brain tumor status and covariate information from family members of the participants in the original case–control sample. There were 476 brain cancer case families and 462 control families in the study.

We could use the methods that we have been discussing to fit a marginal model for the probability of becoming a glioma victim, but a major interest of the researchers was the estimation of within-family characteristics. One way of approaching this would be to fit a mixed logistic model with one or more random family effects.

Note that, strictly speaking, the original sampling scheme in this example is not included in our case–control set-up. The stratification here is related to the response variable but not completely determined by it. Stratum 1 contains the 476 families with a case diagnosed in a particular small time interval while Stratum 2 contains the remaining 1,942,490 families, some of which contain brain cancer victims.

Neuhaus et al. (2006) develop efficient semiparametric methods for stratified multistage sampling in situations where the stratification depends on the response, possibly in an unspecified way that has to be modeled, and observations within a primary sampling unit are related through some parametric model. The estimates require the solution of $p + H$ estimating equations, where p is the dimension of the parameter vector and H is the number of strata. The covariance matrix can also be estimated in a straightforward way using an analog of the inverse observed information matrix. The whole procedure can be implemented using a general maximization routine but this still requires some computing expertise.

We could also fit the same models using design weighted estimators, which has the big advantage of requiring no specialist software. In our example, case families would have weight 1 and control families would have weight $1{,}942{,}490/462 \approx 4200$. With such a huge disparity, we might expect the weighted estimates to be very inefficient. Unfortunately it turned out to be almost impossible to fit an interesting model for which the weighted estimates converged. One problem is that the weighted estimates are based almost entirely on the control sample and there is very little information about family effects in the control families. (Another problem is that we did not have information on age for family members and any model without age was grossly misspecified!) For this reason, we had to resort to simulation that is far from complete at this stage. It seems, however, that the efficiency of weighted estimates is less than 10% of the efficient semiparametric estimates here. More details are given in Neuhaus et al. (2002, 2006).

Although our simulations are at a very early stage, it is possible to draw a few tentative conclusions. The main one is that within-family quantities are very poorly estimated, even using fully-efficient procedures. Case–control family designs, where the information on family members is obtained as an add-on to a standard case–control design, simply do not contain enough information to estimate the parameters of interest to genetic epidemiologists unless the associations are extremely strong. More efficient variants are possible, however. For example, if we can identify families containing more than one case, then it is possible to get much greater efficiency by over-sampling

such families. In essence, we would be taking the family as the sampling unit, defining a "case family" as one containing multiple individual cases and then taking a case–control sample of families. This is an important area where a lot of work still needs to be done.

5. Conclusion

The subject of this chapter is one of those areas where practice has forged ahead of theory. One of the few books that discusses the topic in any depth is Korn and Graubard (1999). One aspect that has received a reasonable amount of theoretical attention in the literature is stratification. Efficient procedures for incorporating stratifying variables in the analysis have been developed by Scott and Wild (1997), Breslow and Holubkov (1997), and Lawless et al. (1999), among others, when the variables can take only a finite set of values. Breslow and Chatterjee (1999) have considered how best to use such information at the design stage. The extension of both analysis and design to situations where we have information on continuous variables such as age for all members of the population is an area that still needs work. Much less has been written on the effect of clustering, even though multistage sampling is in common use. Exceptions are Graubard et al. (1989), Fears and Gail (2000), Scott and Wild (2001a), and Scott (2006). There has also been some work on the choice of sampling design (see Waksberg, 1998), particularly on the relative merits of area sampling and random digit dialling (see Brogan et al., 2001; DiGaetano and Waksberg, 2002, for example), but much needs to be done. Because the essence of the problem boils down to estimating two population means [see equation (8)], it should be possible to transfer a lot of standard survey expertise about efficient design to this problem.

Sample Surveys: Inference and Analysis, Vol. 29B
ISSN: 0169-7161
© 2009 Elsevier B.V. All rights reserved
DOI: 10.1016/S0169-7161(09)00239-9

Inference under Informative Sampling

Danny Pfeffermann and Michail Sverchkov

1. Introduction

1.1. Selection bias

Survey data are frequently used for *analytic inference* about statistical models holding for the corresponding population data. By analytic inference we mean inference about the model parameters or functions of them like expectations, variances, regression coefficients, etc; and the prediction of unobserved data. The inference ordinarily takes the form of point estimation (and possibly prediction), confidence intervals, hypothesis testing, or posterior distributions. Familiar examples include the estimation of income elasticities from household surveys, the analysis of labor market dynamics from labor force surveys, and the study of the relationships between risk factors and disease incidence from health surveys. Sometimes, models are fitted to survey data for estimating known functions of the finite population values, such as means, proportions, or correlations. This is known as *descriptive inference*.

The data are usually collected for samples drawn by probability sampling. This induces a set of base weights for the sampled units reflecting unequal selection probabilities. Differential weighting can also result from a variety of adjustments, such as the accounting for unit nonresponse and calibration. Chapters 9 and 25 of this handbook discuss such adjustments in great depth. We refer to the final set of weights as the "sampling weights."

When the sampling weights are related to the values of the model outcome variable even after conditioning on the model covariates, the observed outcomes are no longer representative of the population outcomes due to the sampling or response process and the model holding for the sample data is then different from the model holding in the population. In symbols, $f(y_i|x_i, i \in s) \neq f(y_i|x_i)$, where s defines the sample with observations, and (y_i, x_i) are the measured values of the dependent and covariate variables for unit i. We say that the sampling and/or the response are *informative* in this case. Ignoring an informative sample may yield large biases and erroneous inference. The books edited by Kasprzyk et al. (1989) and Skinner et al. (1989) contain many examples illustrating the effects of ignoring an informative sampling scheme in the inference process. See also Pfeffermann (1993, 1996) and the more recent studies referenced in this chapter.

Example 1.1. Consider a finite population $U = \{1, \ldots, N\}$ and suppose that for unit $i \in U$, $Y_i \sim \text{Mult}(\{p_k\}, K)$, such that $\Pr(Y_i = k) = p_k, k = 1, \ldots, K; \sum_{k=1}^{K} p_k = 1$. Let $\Pr(i \in s | Y_i = k) = \pi_k$. Then, by Bayes rule, $P_s(Y_i = k) = \Pr(Y_i = k | i \in s) = \pi_k p_k / \sum_{j=1}^{K} \pi_j p_j = p_k^*$, such that $Y_i | i \in s \sim \text{Mult}(\{p_k^*\}, K)$. Ignoring the sample selection and estimating the population probabilities by the ordinary estimates $\hat{p}_k = n_k/n$, where n_k is the number of sampled units with observation $Y_i = k$ and $n = \sum_{k=1}^{K} n_k$, yields unbiased estimators for p_k^*, but biased estimators for p_k, unless $\pi_k = \text{const}$. Note that for known selection probabilities, one can construct the estimator $\tilde{p}_k = (\hat{p}_j/\pi_k)/\sum_{j=1}^{K} (\hat{p}_j/\pi_j)$, which is consistent for p_k under mild conditions but not strictly unbiased under the sample probability function $P_s(Y_i = k)$.

We consider the sample data as the outcome of three random processes. The first process generates a vector value of some random element for each of the N units in U. The second process selects n units at random from U to the sample (n can be random). The third process selects the responding units. This process is obviously not part of the original sample design controlled by the survey statistician and is often the result of "self selection," although it could be caused by many other reasons. See Chapters 9 and 10 of this handbook for further discussion. In the most part of this chapter, we assume full response but where appropriate, we discuss extensions to account for informative nonresponse.

Let $A_U = \{(y_1, x_1, z_1, l_1), \ldots, (y_N, x_N, z_N, l_N)\}$ define in general the N population realizations of the stochastic element (Y, X, Z, L), where Y is the model outcome variable of interest, X is a vector of model covariates, Z is a vector of design variables used for the sample selection and L is a vector of latent variables determining the response. The vector Z may contain some or all of the covariates X, and in special cases also the vector Y, such as in a case-control study, see Chapter 38. The population values $Z_U = \{z_1, \ldots, z_N\}$ are known to the sampler who designs the sampling scheme, but not necessarily to the analyst fitting the model. See Section 2. The vector L is seldom known, although it may contain elements of Y, X, and Z.

Let I_i be the sample indicator such that $I_i = 1$ if unit $i \in s$ and $I_i = 0$ otherwise, and denote by $\pi_i = \Pr(I_i = 1)$ the sample inclusion probability. In this chapter, we only consider probability sampling such that $\pi_i > 0$ for all i. The probabilities $\pi_1, \ldots, \pi_N$ are known to the sampler drawing the sample. We assume that they are known also to the analyst fitting the model for at least the sampled units. Let R_i be the response indicator such that $R_i = 1$ if unit $i \in s$ responds and $R_i = 0$ otherwise, and denote by $r \subseteq s$ the subset of respondents. The response probabilities $\Pr(R_i = 1 | I_i = 1)$ are generally unknown and can only be estimated under strict assumptions. Our interest in this chapter is in situations where the design or the latent variables underlying the sample selection and response probabilities are correlated with Y after conditioning on X, such that the sampling and/or the response process are informative when modeling $f(Y|X = x)$, the target of the inference process.

In what follows we use the abbreviation *pdf* to define the probability density function when Y is continuous and the probability function when Y is discrete. Following Pfeffermann et al. (1998a), the conditional *marginal* sample *pdf* $f_s(y_i|x_i)$ in the case of *full response* is defined as the conditional *pdf* of Y_i given that unit i is in the sample and

the covariates x_i. By Bayes theorem,

$$f_s(y_i|x_i) = f_U(y_i|x_i, I_i = 1) = \frac{\Pr(I_i = 1|x_i, y_i)\, f_U(y_i|x_i)}{\Pr(I_i = 1|x_i)}, \tag{1}$$

where $f_U(y_i|x_i)$ is the corresponding population *pdf*. The probabilities $\Pr(I_i = 1|x_i, y_i)$ are generally not the same as the sample inclusion probabilities π_i, which may depend on all the population values Z_U. However, the use of the marginal sample *pdf* only requires modeling $\Pr(I_i = 1|x_i, y_i)$. Note that $\Pr(I_i = 1|y_i, x_i) = E_U(\pi_i|y_i, x_i)$, where $E_U(\cdot)$ is the expectation under the population *pdf* (see Section 4.1).

It follows from (1) that unless $\Pr(I_i = 1|x_i, y_i) = \Pr(I_i = 1|x_i) \forall y_i$, the sample *pdf* is different from the population *pdf*, in which case the sampling design becomes informative and cannot be ignored at the inference process. In particular, it follows from (1) that under an informative sampling scheme, $E_s(Y_i|x_i) = E_U\left[\frac{\Pr(I_i=1|x_i,Y_i)Y_i}{\Pr(I_i=1|x_i)}\Big|x_i\right] \neq E_U(Y_i|x_i)$, where $E_s(\cdot)$ is the expectation under the sample *pdf* $f_s(y_i|x_i)$, illustrating that ignoring an informative sampling scheme can bias the inference.

Example 1.2. Suppose that in the population $Y_i|x_i \sim N(\beta_0 + x_i'\beta, \sigma^2)$ and that $\Pr(I_i = 1|y_i, x_i) = E(\pi_i|y_i, x_i) = \exp[A_1 y_i + A_2 y_i^2 + g(x_i)]$, where A_1 and $A_2 < 0$ are constants and $g(x_i)$ is some deterministic function of the covariates. Simple algebra using (1) shows that in this case, $f_s(y_i|x_i) = N[(\beta_0 + A_1\sigma^2 + x_i'\beta)/C, \sigma^2/C]$, where $C = (1 - 2\sigma^2 A_2)$. Thus, although the sample values have again a normal distribution, $E_s(Y_i|x_i) = (\beta_0 + A_1\sigma^2 + x_i'\beta)/C \neq \beta_0 + x_i'\beta = E_U(Y_i|x_i)$, and the variance of the residual terms is also changed. In the special case where $A_2 = 0$, the sample model slope coefficients (but not the intercept) and the residual variance are the same as under the population model. If $A_1 = 0$ as well, the inclusion in the sample is noninformative and the population and sample models coincide.

REMARK 1.1. *The definition of the sample pdf (1) can be extended to account for nonresponse by distinguishing between the sample selection and the response. Using the response indicators R_i, the respondents pdf is defined as*

$$\begin{aligned} f_r(y_i|x_i) &= f(y_i|x_i, I_i = 1, R_i = 1) \\ &= \frac{\Pr(R_i = 1|y_i, x_i, I_i = 1)\Pr(I_i = 1|y_i, x_i)\, f_U(y_i|x_i)}{\Pr(R_i = 1|x_i, I_i = 1)\Pr(I_i = 1|x_i)}. \end{aligned} \tag{2}$$

REMARK 1.2. *The pdfs (1) and (2) refer to the marginal distribution of the measurement y_i. This definition generalizes very naturally to the joint pdf of two or more measurements associated with different units. More generally, define for every subset $s \subset U$ the sample indicator A_s, such that $A_s = 1$ if the selected sample is s, and $A_s = 0$ otherwise, and assume for convenience full response. Denote the data associated with s by (y_s, x_s). The joint sample pdf of $Y_s|x_s$ is then*

$$f_s(y_s|x_s) = f_U(y_s|x_s, A_s = 1) = \frac{\Pr(A_s = 1|y_s, x_s)\, f_U(y_s|x_s)}{\Pr(A_s = 1|x_s)}. \tag{3}$$

The *pdf* $f_U(y_s|x_s)$ can be general, allowing in particular for correlated measurements, but modeling the probability $\Pr(A_s = 1|y_s, x_s)$ is practically only feasible if the sample can be decomposed into exclusive and exhausting subsets s_k such that $\Pr(A_s = 1|y_s, x_s) \prec \prod_k \Pr(A_{s_k} = 1|y_{s_k}, x_{s_k})$ and $\Pr(A_{s_k} = 1|y_{s_k}, x_{s_k})$ satisfies the same model in all the subsets (see Example 1.3). In particular, if the population outcomes are independent given the covariates and $\Pr(A_s = 1|y_s, x_s) \prec \prod_{i\in s} \Pr(I_i = 1|y_i, x_i)$, (3) takes the form,

$$f_s(y_s|x_s) = \prod_{i\in s} \frac{\Pr(I_i = 1|y_i, x_i) f_U(y_i|x_i)}{\Pr(I_i = 1|x_i)} = \prod_{i\in s} f_s(y_i|x_i), \tag{4}$$

such that the sample outcomes are likewise independent.

Example 1.3. Consider the case of a clustered population $U = \cup_l U_l$, with independent measurements between clusters, such that $f_U(y_U|x_U) = \prod_l f_U(y_{U_l}|x_{U_l})$, where (y_U, x_U) defines the population measurements. Suppose that the sample is drawn by a single-stage cluster sampling design with the clusters selected independently with probabilities $\pi_l = r(y_{U_l}, x_{U_l})$ for some function r. Then, $\Pr(A_s = 1|y_U, x_U) = \prod_{k\in s} r(y_{U_k}, x_{U_k}) \times \prod_{j\notin s} [1 - r(y_{U_j}, x_{U_j})]$. Since $(y_{U_k}, x_{U_k}) = (y_{s_k}, x_{s_k})$ for $k \in s$, it follows that $\Pr(A_s = 1|y_s, x_s) = C \times \prod_{k\in s} r(y_{s_k}, x_{s_k}) = C \times \prod_{k\in s} \Pr(A_{s_k} = 1|y_{s_k}, x_{s_k})$, where for given covariates x_{U_j}, $j \notin s$, C is a constant satisfying, $C = \int \prod_{j\notin s} [1 - r(y_{U_j}, x_{U_j})] f_U(y_{U_j}|x_{U_j}) dy_{U_j}$. The case of a nonclustered population with independent measurements and Poisson sampling of individual units is a special case where each cluster consists of a single element, giving rise to the sample *pdf* (4).

1.2. Distinction between the sample distribution and the randomization distribution

The sample distribution refers to the distribution of the sample measurements, as defined by the population model and the sampling design, with the realized sample of respondents held fixed. This implies, for example, that the sampled clusters under two-stage sampling and the observed covariates, $\{x_i, i \in s\}$, are fixed under this distribution. The same is true for the distribution of the responding units, which depends also on the response process. The randomization distribution, on the other hand, conditions on the population values $\{y_i, x_i, i \in U\}$, which are treated as fixed values, and the only stochastic element used for inference is the random selection of the sample (or the respondents). The use of this distribution does not permit therefore conditioning on the sampled clusters or the observed covariates.

The use of the sample (respondents) distribution requires modeling $\Pr(I_i = 1|x_i, y_i)$ (Eq. 1) and $\Pr(R_i = 1|y_i, x_i, I_i = 1)$ in the case of nonresponse (Eq. 2), but it permits the computation of the conditional *pdf* of the sample measurements given the covariates, and hence the use of classical inference tools, if the sample selection and response are approximately independent between the units. As discussed and illustrated in Section 4.2, the latter requirement is often not binding. This is not possible under the randomization distribution because the values of the outcome variable (and often also the auxiliary variables) are unknown for units outside the sample, such that the randomization distribution of any given statistic over all possible sample selections is generally unknown. Consequently, the use of this distribution for inference is

restricted mostly to estimation problems, with probabilistic statements like confidence intervals generally requiring asymptotic normality assumptions. Note also that the randomization distribution cannot be used for predicting the outcome value of a unit outside the sample given the values of the auxiliary variables, even under noninformative sampling.

The sample distribution is also different from the UD distribution (often referred to as the ξp distribution), defined as the combined distribution over all possible realizations of the finite population measurements (the population U distribution) and all possible sample values for given population values (the randomization D distribution). See Chapter 24. The UD distribution is often used for comparing the precision of design-based estimators in situations where direct comparison of the randomization variances is not feasible. The obvious difference between the UD distribution and the sample distribution is that the sample distribution conditions on the observed sample of units, (and the observed values of the auxiliary variables), whereas the UD distribution accounts for all possible sample selections. In the case of nonresponse, the UD distribution should be replaced by the UDR distribution; the combined distribution over all possible realizations of the finite population measurements, all possible samples and all possible samples of respondents, given the selected sample.

2. Informative and ignorable sampling

In this section, we assume for convenience full response but extension of the results to the case of nonresponse follows through. Thus far, we suppressed for convenience from the notation the parameters underlying the population *pdf* and the sampling process. Consider the sample *pdf* in (3). With added parameter notation, it can be written as

$$f_s(y_s|x_s;\theta,\gamma) = \frac{\Pr(A_s = 1|y_s, x_s;\gamma)\, f_U(y_s|x_s;\theta)}{\Pr(A_s = 1|x_s;\theta,\gamma)}. \tag{5}$$

Thus, the population and sample *pdf*s of $Y_s|X_s = x_s$ are the same when

$$\Pr(A_s = 1|y_s, x_s;\gamma) = \Pr(A_s = 1|x_s;\theta,\gamma)\ \forall y_s. \tag{6}$$

In this case, inference on θ can be achieved by fitting the population distribution to the sample data, ignoring the sample selection. Note, that this conclusion refers to the selected sample defined by the event $A_s = 1$.

The condition (6) is a strong condition. In a fundamental article on missing values, Rubin (1976) establishes conditions under which the sampling process can be ignored for likelihood, Bayesian, or sampling distribution (repeated sampling from a model) inference, that is, conditions under which the population *pdf* $f_U(y_s|x_s;\theta)$ can be used for inference. Let $Y_U' = (Y_s', Y_{\tilde{s}}')$ represent the random population outcomes with $Y_{\tilde{s}}$ denoting the outcomes for the nonsampled units. Suppressing for convenience the conditioning on the covariates, denote by γ the parameters governing $\Pr(A_s = 1|Y_U = y_U;\gamma)$ and let the symbol $\perp$ define independence. Two fundamental conditions established by Rubin (1976) are

(I) Data *missing at random*; for each γ, $A_s \perp Y_{\tilde{s}}|Y_s$ for all possible $Y_{\tilde{s}}$,
(II) Data *observed at random*; for each γ and $Y_{\tilde{s}}$, $A_s \perp Y_s|Y_{\tilde{s}}$ for all possible Y_s.

Rubin shows that the sample selection process can be ignored for Bayesian- and likelihood-based inferences when condition I holds and the parameters θ indexing the population *pdf* are distinct from γ. However, for sampling distribution inference, both the conditions I and II are required for ignoring the sample selection. The latter inference mode should be interpreted in this case as being conditional on the realized sample $(A_s = 1)$.

Little (1982) extended Rubin's results by distinguishing between the original sample selection and the response process. Another important distinction is that Little conditions on the population values Z_U of the design variables used for the sample selection, such that the emphasis is on the conditional distribution of Y_s given Z_U, with Y_s defining all the fully and partially observed data in the case of item nonresponse. Inference on the target population model $f_U(y;\theta)$ (or more generally $f_U(y|x_U;\theta)$) requires therefore integrating the conditional *pdf* of $Y_s|Z_U$ over the distribution of Z_U. See Section 3.

Sugden and Smith (1984) established conditions under which a sampling process that depends on design variables Z is ignorable given partial information on the design. Let $d_s = D_s(Z_U = z_U)$ contain all the available design information from knowledge of the selection scheme, the sample inclusion probabilities and any known values or functions of z_U. Using previous notation, a key condition for ignorability of the sampling process given the design information is that $A_s \perp Z_U|d_s$, implying $\Pr(A_s = 1|Z_U = z_U) = \Pr(A_s = 1|d_s)$ for all z_U for which $d_s = D_s(Z_U = z_U)$. The authors show that under this condition and if the parameters θ^* governing the *pdf* $f_U(y_U|z_U)$ are distinct from the parameters θ_z governing the *pdf* $g_U(z_U)$, the sample selection can be ignored for Bayesian- and likelihood-based inferences that condition on d_s. The condition $A_s \perp Z_U|d_s$ is sufficient also for predictive inference in the sense that the prediction of $Y_{\bar{s}}$ can be based in this case on $f_U(y_{\bar{s}}|y_s, A_s, d_s)$. However, for sampling distribution inference, the stronger condition $Y_s \perp Z_U|D_s; \theta^*$ for all θ^* needs to be satisfied.

Discussion

For large-scale multi-stage sample surveys with possibly many design variables, it is generally difficult and often impractical to check directly the conditions that permit ignoring the sample selection or nonresponse before conducting the inference. On the other hand, even when the sample *pdf* is different from the population *pdf*, it does not necessarily imply that the inference under consideration is wrong. For a simple illustration consider the special case of Example 1.2 in Section 1, where $A_2 = 0$. In this case, the sample *pdf* is normal with the same slope coefficients and residual variance as under the population *pdf*. Thus, for inference about the slope coefficients one can ignore the sampling process even though the sample model intercept is different from the population model intercept. A similar phenomenon is obtained for logistic models when the sample selection depends on y but not on x. See Pfeffermann et al. (1998a). Keeping this in mind, we review in Section 7 several test statistics proposed in the literature for assessing whether ignoring the sample selection is justified for the inference under consideration.

The aim of this chapter is to discuss ways of making valid inference when the sample selection or the response mechanism cannot be ignored, as concluded either by checking

directly the ignorability conditions or by application of test procedures. Many survey analysts prefer basing the inference on methods that do not require sampling ignorability even when the sample and response processes are deemed ignorable. In Sections 3–6, we review and discuss the main approaches to inference that do not require ignorability conditions.

3. Overview of approaches that account for informative sampling and nonresponse

3.1. Including the design variables among the covariates

As implied by (6), the population model (*pdf*), $f_U(y_s|x_s)$ and the sample model $f_s(y_s|x_s)$ are the same when $\Pr(A_s = 1|y_s, x_s) = \Pr(A_s = 1|x_s)\forall y_s$. By (2), the response process can be ignored for inference when $\Pr(R_i = 1|y_i, x_i, I_i = 1) = \Pr(R_i = 1|x_i, I_i = 1)\forall y_i$. Thus, a possible way to account for the sampling and response effects is to include among the model covariates all the variables and interactions determining the sample and response probabilities. Denoting these variables by $J = Z \cup L$ with population values J_U, the use of this approach requires modeling,

$$f_U(y_s|x_s, J_U = j_U) = \int f_U(y_s, y_{\bar{s}}|x_U, j_U)dy_{\bar{s}}, \tag{7}$$

assuming $f_U(y_s|x_U, j_U) = f_U(y_s|x_s, j_U)$. Variants of this approach are considered by DeMets and Halperin (1977), Holt et al. (1980), Nathan and Holt (1980), Jowell (1985), Chambers (2003), and Gelman (2007).

Example 3.1. Suppose that one is interested in estimating the average effect of education on wages and that this effect is different across different ethnicity groups. Under a balanced sample for which the sampled (responding) proportions in the various ethnicity groups match the corresponding population proportions, the education effect can be estimated by modeling the wages as a function of education. Suppose, however, that because of the sampling design some of the ethnicity groups are misrepresented in the sample. Ignoring the misrepresentation of the sample will clearly bias the estimator of the education effect in this case. Accounting for the sampling design can be achieved by including the ethnicity effects among the model covariates. A simple example is the model,

$$Y_{ijk} = \beta_0 + \beta_1 Ed_j + \beta_{2jk}\big[Ed_j \times Et_k\big] + e_{ijk}, \tag{8}$$

where the index i defines the individual, j the education level, and k the ethnicity group (one of which serving as the "baseline" and hence dropped from the equation). As pointed out by Gelman (2007), the interactions of education and ethnicity define poststratification cells within which the sample selection can be ignored. Estimating the "average" effect of education on wages over the various ethnicities requires integrating out the ethnicity effects by "regressing" the right-hand side of (8) against education using the population model relationship, that is, $E_U(Y|Ed) = E_{U,Et|Ed}[E_U(Y|Ed, Et)]$. More generally,

$$f_U(y_s|x_s) = \int f_U(y_s|x_s, j_U) f_U(j_U|x_s)dj_U. \tag{9}$$

Example 3.2. Suppose that a sample of size n is selected with probabilities defined by the population values of a design variable Z, and that all the sampled units respond. Let the population distribution of Y, X, Z be trivariate normal. The data available to the analyst consists of the observed values of Y and X for the sampled units and the population values of Z. Using properties of the multivariate normal distribution, $E_U(Y|X) = \beta_0 + \beta_{yx}X$, but the ordinary least squares (OLS) estimator of β_{yx} is biased in this case because the sampling probabilities depend on Z, which is correlated with Y and X. The maximum likelihood estimator (*mle*) of β_{yx} based on the observations (y_s, x_s) and the population values Z_U is (DeMets and Halperin, 1977),

$$\hat{\beta}_{yx} = \left\{ s_{xy} + \frac{s_{yz}s_{xz}}{s_z^2}\left(\frac{\hat{\sigma}_z^2}{s_z^2} - 1\right)\right\} \Big/ \left\{ s_x^2 + \frac{s_{xz}^2}{s_z^2}\left(\frac{\hat{\sigma}_z^2}{s_z^2} - 1\right)\right\}, \tag{10}$$

where $s_{uv} = n^{-1}\sum_{i=1}^{n}(u_i - \overline{u}_s)(v_i - \overline{v}_s)$ and $\hat{\sigma}_z^2 = N^{-1}\sum_{i=1}^{N}(z_i - \overline{z}_U)^2$, with $(\overline{u}_s, \overline{v}_s, \overline{z}_U)$ defining the corresponding sample and population means. The estimator $\hat{\beta}_{yx}$ reduces to the OLS estimator when $\hat{\sigma}_z^2 \cong n^{-1}\sum_{i=1}^{n}(z_i - \overline{z}_s)^2$ (e.g., the sample is selected by simple random sampling) or when $s_{xz} = 0$. Holt et al. (1980) extended this result to the case where Y, X, Z are vector variables. Nathan and Holt (1980) established conditions under which $\hat{\beta}_{yx}$ is consistent without the multivariate normality assumption. Pfeffermann and Holmes (1985) studied the robustness of the estimator to certain model misspecifications.

REMARK 3.1. *The mle of the population mean of Y is* $\hat{\mu}_y = \overline{y}_s + \hat{\beta}_{yz}^{\mathrm{OLS}}(\overline{z}_U - \overline{z}_s)$, *where* $\hat{\beta}_{yz}^{\mathrm{OLS}}$ *is the OLS estimator when regressing Y against Z in the sample. Clearly, the sample mean,* $\overline{y}_s$ *can be severely biased as an estimator of* μ_y *when the selection depends on Z and* $Corr(Y, Z) \neq 0$.

Discussion

The use of this approach seems appealing, and it has the advantage of allowing classical model-based inference methods once the design and response variables are included in the model. Unfortunately, it is very limited for the following reasons:

(I) It requires knowledge of the population values of all the variables determining the sample selection and response. Although the population values of the design variables used for the sample selection are known to the sampler drawing the sample, they may not be known to the analyst fitting the model because of confidentiality restrictions or other reasons. In the case of nonresponse, both the sampler and the analyst may have very limited information on which variables explain the nonresponse. Note also that knowledge of the population values of the design variables is imperative when predicting unobserved values of the outcome variables under the extended model with the added design variables.

(II) In large-scale surveys, including in the model all the geographic and operational variables used for the sampling design may be formidable (Alexander, 1987).

(III) Including the variables determining the sample inclusion among the model covariates may increase the prediction power of the model and thus be useful for prediction purposes, but the resulting model may no longer have a scientific

interpretation, requiring integrating them out from the model (Eq. 9). A simple example taken from Holt et al. (1980) illustrates the problem. Suppose that it is required to model the relationship between income and education based on a fully responding sample selected with probabilities proportional to the taxes paid on a previous year. Clearly, regressing income against education and the tax value has no scientific meaning and the coefficient of the education variable will most likely be highly insignificant, with all the variation in the income variable explained by the tax variable. In this example, it may be relatively simple to integrate out the tax variable, but in practice there may be many covariates and many design variables, and modeling the relationship between the design variables and the covariates to integrate out the effect of the design variables can be very complicated.

(IV) The approach is not operational when the inclusion in the sample depends also on the outcome values, that is, $Z = \{Y, Z^*\}$ and $\Pr(A_s = 1|Y_U, X_U, Z^*_U) \neq \Pr(A_s = 1|X_U, Z^*_U)$. See Pfeffermann (1996). A similar situation arises when the nonresponse is not missing at random.

3.2. Using the sampling weights as surrogates for the design variables

For situations where there are too many design variables determining the sample selection to include them all in the model, or when some or all of these variables are unknown to the analyst, it is often advocated to include in the model the sampling weights instead of the design variables. Examples of the use of this approach can be found in DuMouchel and Duncan (1983), Särndal and Wright (1984), Rubin (1985), Chambers et al. (1998), and Wu and Fuller (2006).

Rubin (1985) defines the vector $a = (a_1, \ldots, a_N)' = a(Z_U)$ to be an adequate summary of Z if $\Pr(A_s = 1|Z_U) = \Pr(A_s = 1|a)$. The author shows that the vector $\pi_U = (\pi_1, \ldots, \pi_N)$ of the sample inclusion probabilities is the coarsest possible adequate summary of Z. It follows from Section 2 that for sampling designs such that $\Pr(A_s = 1|Y_U, Z_U) = \Pr(A_s = 1|Z_U)$, if π_U is an adequate summary, the sample selection can be ignored for inference on the distribution $f_U(y_s|x_s, \pi_U)$.

Discussion

The use of this approach may require knowledge of the sample inclusion probabilities for all the population units. Here again, this information may not be available to the analyst fitting the model. Extension of the approach to the case of nonresponse is particularly problematic since the response probabilities are generally unknown and can at best be estimated. Another major problem with this approach is that for general sampling designs, the vector π_U may not be an adequate summary of Z. Indeed, it is hard to conceive that for large-scale multi-stage surveys, a single vector can summarize adequately all the information entailed in many geographic and operational design variables. Sugden and Smith (1984) and Smith (1988) studied the necessary design information other than the vector π_U required to warrant sampling ignorability.

REMARK 3.2. *Although as just noted it is not generally true that by conditioning on π_U the sample inclusion process can be ignored in the sense that $f_U(y_s|x_s, \pi_U, A_s = 1) =$*

$f_U(y_s|x_s, \pi_U)$, it is nonetheless true that the marginal distributions are the same, that is, $f_s(y_i|x_i, \pi_i) = f_U(y_i|x_i, \pi_i, I_i = 1) = f_U(y_i|x_i, \pi_i)$. See Skinner (1994) and Remark 4.1 in this chapter.

3.3. Methods based on probability weighting

In the previous subsections, we considered methods requiring knowledge of the variables J determining the sample inclusion and response probabilities, or at least adequate summary of them. The methods considered below only require knowledge of the sample inclusion probabilities or the sampling weights for the responding sampled units. As such, they are restricted to situations of full response, or when the response probabilities can be estimated sufficiently accurately, in which case the sampling weight for a responding unit is the inverse of the product of the unit's selection probability and its estimated response probability. Often the sampling weights are slightly adjusted. One common adjustment forces the probability-weighted estimators of the population totals of some of the measured variables to equal the corresponding known population totals. This adjustment, known as "calibration," can also be used to estimate the parameters governing the model for the response mechanism. It is discussed in detail in Chapter 25.

We introduce the technique of probability weighting with a simple example. Suppose that the population Y-values are distributed with mean $E_U(Y_i) = \theta$. A sample is selected with probabilities $\pi_i = \Pr(i \in s)$. Assume first full response and no weight adjustments. It is desired to estimate θ. Consider the Horvitz and Thompson (1952) estimator $\hat{\overline{Y}}_{\mathrm{HT}} = N^{-1} \sum_{i\in s} Y_i/\pi_i = N^{-1} \sum_{i\in s} w_i y_i$, where $w_i = 1/\pi_i$. As is well known, $\hat{\overline{Y}}_{\mathrm{HT}}$ is randomization (design) unbiased for the population mean $\overline{Y}_U = N^{-1} \sum_{j=1}^{N} Y_j$, that is, $E_D(\hat{\overline{Y}}_{\mathrm{HT}}|Y_U = y_U) = \overline{y}_U$, where the randomization distribution is over all possible sample selections with the population y-values held fixed (see Chapter 2). Under general conditions on the sampling design, $\hat{\overline{Y}}_{\mathrm{HT}}$ is also randomization consistent for $\overline{y}_U$ in the sense that $\operatorname*{plim}_{n\to\infty, N\to\infty} |\hat{\overline{Y}}_{\mathrm{HT}} - \overline{y}_U| = 0$, where "*plim*" defines the "limit in probability" under the randomization distribution. See Chapter 40 for the concept of consistency in finite population sampling.

Now, under very general conditions, the random population mean $\overline{Y}_U$ is model unbiased and consistent for θ. We conclude, therefore, that $\hat{\overline{Y}}_{\mathrm{HT}}$ is consistent for θ. More precisely, if $(\hat{\overline{Y}}_{\mathrm{HT}} - \overline{Y}_U)$ is $O_p(n^{-0.5})$ under the randomization distribution and $(\overline{Y}_U - \theta)$ is $O_p(N^{-0.5})$ under the population model, then $(\hat{\overline{Y}}_{\mathrm{HT}} - \theta) = (\hat{\overline{Y}}_{\mathrm{HT}} - \overline{Y}_U) + (\overline{Y}_U - \theta) = O_p(n^{-0.5})$ under the DU distribution, the distribution over all possible samples for a given realization of the population values, and over all possible realizations of the population values. We may decompose the DU variance of $\hat{\overline{Y}}_{\mathrm{HT}}$ around θ as,

$$\mathrm{Var}_{DU}\left(\hat{\overline{Y}}_{\mathrm{HT}}\right) = E_U\left[\mathrm{Var}_D\left(\hat{\overline{Y}}_{\mathrm{HT}}|Y_U\right)\right] + \mathrm{Var}_U\left[E_D\left(\hat{\overline{Y}}_{\mathrm{HT}}|Y_U\right)\right]. \tag{11}$$

For single-stage sampling and when n is much smaller than N, as is usually the case, the second term on the right-hand side of (11) is negligible compared with the first term, and $\mathrm{Var}_{DU}(\hat{\overline{Y}}_{\mathrm{HT}})$ can be estimated by the randomization variance estimator $\hat{\mathrm{Var}}_D(\hat{\overline{Y}}_{\mathrm{HT}}|Y_U)$. This result does not necessarily hold, however, for cluster sampling since in this

case $\mathrm{Var}_D(\hat{\overline{Y}}_{\mathrm{HT}}|Y_U)$ is typically of order $O(1/m)$, where m is the number of sampled clusters, and under a suitable model, $\mathrm{Var}_U[E_D(\hat{\overline{Y}}_{\mathrm{HT}}|Y_U)]$ is $O(1/M)$, where M is the number of population clusters. For $\hat{\mathrm{Var}}_D(\hat{\overline{Y}}_{\mathrm{HT}}|Y_U)$ to be a proper estimator for $\mathrm{Var}_{DU}(\hat{\overline{Y}}_{\mathrm{HT}})$ in this case, m must be much smaller than M. See Pfeffermann (1993) and Graubard and Korn (2002).

As noted above, the probability-weighting procedure can also be applied when not all the sampled units respond, provided that the response probabilities can be estimated with sufficient accuracy. In this case, the inclusion probability π_i is the product of the unit's original sample-selection probability and its (estimated) probability of response. See Chapter 9. When the sampling weights w_i comprise the estimates of the unit response probabilities or encompass other adjustments (such as calibration), the weighted estimator $N^{-1}\sum_{i\in s} w_i y_i$ remains nearly design unbiased and randomization consistent for the population mean $\overline{Y}$ under mild conditions. As a result, using such weights is still called "probability weighting."

The probability-weighting procedure easily extends to the estimation of other model parameters. Denoting $\sigma^2 = \mathrm{Var}_U(Y_i) = E_U(Y_i^2) - \theta^2$ in the example above, a *DU* consistent estimator for σ^2 is $N^{-1}\sum_{i\in s} y_i^2/\pi_i - (\hat{\overline{Y}}_{\mathrm{HT}})^2$. (One can add $\hat{\mathrm{Var}}_D(\hat{\overline{Y}}_{\mathrm{HT}}|Y_U = y_U)$ to correct for the randomization bias of $(\hat{\overline{Y}}_{\mathrm{HT}})^2$ as an estimator of θ^2, but this will increase the variance of the resulting estimator of the variance.)

> **Example 3.3.** Consider again the estimation of the average education effect on wages considered in Example 3.1, but suppose now that the ethnicity affiliation is unknown, making it impossible to fit the model (8). Denoting for convenience by y_i the wage of sampled unit i and by x_i the corresponding education level, the average education effect could possibly be estimated as, $\hat{\beta}_w = \sum_{i\in s} w_i(y_i - \hat{\overline{Y}}_{\mathrm{HT}})(x_i - \hat{\overline{X}}_{\mathrm{HT}})\big/\sum_{i\in s} w_i(x_i - \hat{\overline{X}}_{\mathrm{HT}})^2$. This estimator, however, is *DU* consistent for the true education effect only if the relationship in the population between wages and education is linear, which is not the case under the model (8). In the latter case, the estimator $\hat{\beta}_w$ will estimate the best linear approximation under a quadratic loss function for the true relationship between wages and education in the population under consideration. See the discussion at the end of this subsection.

REMARK 3.3. *Beaumont (2008) proposed replacing the base weights $w_i = 1/\pi_i$ by their estimated conditional expectations $\hat{w}_i = \hat{E}_s(w_i|y_i)$, where the subscript s signifies that the expectation is taken with respect to the model holding for the weights in the sample. The author shows that under correct model specification, the "smoothed" H-T estimator $\hat{\overline{Y}}_{\mathrm{HTS}} = N^{-1}\sum_{i\in s} \hat{w}_i y_i$ is randomization consistent for $\overline{y}_U$ and has a smaller randomization variance than the standard H-T estimator that uses the original sampling weights. The smoothed weights can be used in principle for enhancing other design-based estimators in common use.*

The methods reviewed thus far assume a known explicit form for the estimator under consideration. Often, however, the explicit form of the estimator is unknown even under ignorable sampling. For example, the *mle* of the regression coefficients in logistic regression is the solution of a set of estimating equations (EE) that can only be solved

iteratively (see Example 3.4 below). How can probability weighting be used in such cases? One idea is to estimate the EE that would be obtained if all the population values had been observed (hereafter, the "census" equations) by design-consistent estimators and then solve the resulting EE. Clearly, the census equations are free of any sampling effects.

Example 3.4. The logistic model with covariates x_i assumes $p_i(x_i) = P_U(Y_i = 1|x_i) = \exp(x_i'\beta)/[1 + \exp(x_i'\beta)]$. The corresponding census likelihood equations are $\sum_{k=1}^{N} [y_k - p_k(x_k)]x_k = \sum_{k=1}^{N} u_k(x_k) = 0$. A design-unbiased estimator for these equations is

$$\sum_{i\in s} w_i u_i\,(x_i) = \sum_{i\in s} w_i\,[y_i - p_i\,(x_i)]\,x_i = 0. \tag{12}$$

The "probability-weighted" estimator for the vector coefficient β is obtained by solving the equations $\sum_{i\in s} w_i u_i(x_i) = 0$.

REMARK 3.4. *When the census equations are defined by the likelihood equations as in Example 3.4, the estimator obtained by solving the probability-weighted EE is known in the sampling literature as the "pseudo maximum likelihood estimator (pmle)." See Binder (1983), Skinner et al. (1989), and Pfeffermann (1993) for discussion with many examples.*

Example 3.5. Consider the population two-level (random intercept) model:

$$\begin{aligned} &\text{At level 1 (say pupils)}, \quad Y_{ij} = \beta_{0i} + x_{ij}'\beta + \varepsilon_{ij},\ \varepsilon_{ij} \sim N\left(0, \sigma_\varepsilon^2\right),\ j = 1, \ldots, M_i; \\ &\text{At level 2 (say schools)},\ \beta_{0i} = l_i'\alpha + v_i, \qquad\qquad v_i \sim N\left(0, \sigma_v^2\right),\ i = 1, \ldots, N, \end{aligned} \tag{13}$$

where x_{ij} and l_i are known covariates and ε_{ij} and v_i are independent for all i and j. The unknown parameters are the vectors of coefficients $\vartheta' = (\beta', \alpha')$ and the variances $\tau' = (\sigma_\varepsilon^2, \sigma_v^2)$. Assume full response. Under ignorable sampling of second- and first-level units, the *mle* of (ϑ', τ') are computed most conveniently by iterating between the estimation of ϑ for "known" τ, and the estimation of τ for "known" ϑ, with the "known" values defined by the estimators obtained in the previous iteration. The two sets of estimators on the r-th iteration are obtained as the solutions of the equations, $P^{(r)}\vartheta = q^{(r)}$; $R^{(r)}\tau = s^{(r)}$, with appropriate definition of the matrices $(P^{(r)}, R^{(r)})$ and the vectors $(q^{(r)}, s^{(r)})$, $r = 1, 2, \ldots$, (Goldstein, 1986). If applied to all the population values, these equations define the corresponding census equations.

Suppose now that second-level units are sampled with probabilities π_i that are possibly related to the intercepts (random effects) β_{0i}, and first-level units are sampled with probabilities $\pi_{j|i}$ that are possibly related to the outcomes Y_{ij}. The *pmle* for this model can be obtained by first expressing the elements of the matrices $(P^{(r)}, R^{(r)})$ and the vectors $(q^{(r)}, s^{(r)})$ as sums over second- and first-level units, and then estimating each population total of the form $\sum_{i=1}^{N} t_i$ by the H-T estimator $\sum_{i\in s} (t_i/\pi_i)$, and each population total of the form $\sum_{j=1}^{M_i} t_{ij}$ by the H-T estimator $\sum_{j\in s_i} (t_{ij}/\pi_{j|i})$. See Pfeffermann et al. (1998b). We return to multilevel modeling of survey data in Section 6.2.

The use of probability-weighted EE is not restricted to normal likelihood equations. In general, if the left-hand side of the census equations has the form $\sum_{k=1}^{N} u(y_k, x_k; \theta)$, then it can be estimated under the randomization distribution by $\sum_{i\in s} w_i u(y_i, x_i; \theta)$. The probability-weighted estimator, $\hat{\theta}_{PW}$ of θ is the solution of $\sum_{i\in s} w_i u(y_i, x_i; \theta) = 0$. For example, if $u(y_k; \theta) = [\Delta(\theta - y_k) - F_U(\theta)]$ where $F_U(\theta)$ is the cumulative population distribution at θ and $\Delta(a) = 1(0)$ when $a \geq 0 (a < 0)$, then the probability-weighted estimator for $F_U(\theta)$ is obtained by solving $\sum_{i\in s} w_i u(y_i; \theta) = 0$, yielding $\hat{F}_{U,\mathrm{PW}}(\theta) = \sum_{i\in s} w_i \Delta(\theta - y_i) \big/ \sum_{i\in s} w_i$, which is the familiar Hajek (1971a) estimator. Godambe and Thompson (1986a) established optimality properties of estimators that solve EE of the form $\sum_{i\in s} w_i u(y_i, x_i; \theta) = 0$. See Chapter 36 for the estimation of distribution functions from complex survey data. The use of EE for sample survey inference is considered in Chapter 26.

Discussion

Methods based on probability weighting are in broad use both for the estimation of finite-population quantities (often referred to as "descriptive inference") and for "analytic inference" on population model parameters. The main attraction of these methods is that they are deemed to be "model free," except perhaps when estimating the response probabilities which are often based on models (sometimes implicitly), see Chapter 9. It is often argued, therefore, that probability weighting is more robust to possible model misspecification than direct model-based inference, but this argument should be cautioned.

Probability-weighted estimators are randomization consistent for the corresponding descriptive population quantities (*cdpq*), defined as the (hypothetical) solutions of the EE for the model parameters if all the population values had been observed (Pfeffermann, 1993). However, if the population model is misspecified, the target *cdpq* could be the wrong estimand. For example, the estimator $\hat{\beta}_w = \sum_{i\in s} w_i (y_i - \hat{\overline{Y}}_{\mathrm{HT}})(x_i - \hat{\overline{X}}_{\mathrm{HT}}) \big/ \sum_{i\in s} w_i (x_i - \hat{\overline{X}}_{\mathrm{HT}})^2$ estimates the *cdpq* $B = \sum_{k=1}^{N} (y_k - \overline{y}_U)(x_k - \overline{x}_U) \big/ \sum_{k=1}^{N} (x_k - \overline{x}_U)^2$, which is generally consistent for the regression coefficient β under the simple regression model $E_U(Y_k|x_k) = \alpha + \beta x_k$, but if the true population regression relationship is actually polynomial and contains x^2 as a second covariate, the use of the estimator $\hat{\beta}_w$ may yield erroneous inference. See Pfeffermann (1993) and Chapter 24 for further discussion and other examples.

Estimating the randomization variance of probability-weighted estimators is generally simple, using available techniques in finite population sampling considered in many chapters of this handbook. Binder (1983) developed a general approach for estimating the randomization variance of estimators obtained as the solution of probability-weighted EE, see Chapters 24 and 26. Fuller (1975), Binder (1983), Chambles and Boyle (1985), and Francisco and Fuller (1991) developed central limit theorems applicable to probability-weighted estimators. In spite of these desirable properties of probability-weighting, the method has some severe limitations (see also Chapter 24):

(1) It is restricted mostly to point estimation. Probabilistic inference like confidence intervals or hypothesis testing requires large sample normality assumptions since the randomization distribution of weighted statistics depends on the sampling design and it is generally unknown for small samples. Consequently,

the randomization distribution does not lend itself to the use of classical inference methods, such as likelihood-based inference or Bayesian statistics.

(2) The variances of probability-weighted estimators are computed with respect to the randomization distribution and the use of this approach does not permit conditioning on the selected sample, for example, conditioning on the observed covariates or the selected clusters in a multilevel model.

(3) As often illustrated in the literature, probability-weighted estimators generally have larger variances than model-based estimators, notably for small samples and large dispersion of the sampling weights. See the references in Pfeffermann (1993, 1996) and the more recent references in the next section.

(4) The use of the randomization distribution does not lend itself to prediction problems, such as the prediction of the dependent variable for given nonsampled covariates under a regression model, or the prediction of small area means for areas with no samples, in a small area estimation problem. This is true even for simple random sampling. Small area estimation under informative sampling of areas and within the selected areas is considered in Section 6.3.

4. Use of the sample distribution for inference

4.1. Definition and relationship to the population distribution

Basing the inference on the sample distribution of the sample outcomes overcomes many of the problems underlying the approaches reviewed in Section 3. In particular, it does not require knowledge of the design and latent variables determining the sample selection and response probabilities and allows modeling directly the population or sample *pdf* of $Y|x$. However, it requires modeling the sample and response probabilities as functions of the observed data.

Consider first the case of full response. The marginal sample *pdf* is defined then by (1). Note that $\Pr(I_i = 1|x_i, y_i) = \int \Pr(I_i = 1|\pi_i, x_i, y_i) f_U(\pi_i|x_i, y_i) d\pi_i = \int \pi_i f_U(\pi_i|x_i, y_i) d\pi_i = E_U(\pi_i|x_i, y_i)$. Pfeffermann and Sverchkov (1999) showed that for a general pair of vector random variables (v_1, v_2) measured for unit $i \in U$:

$$E_U(v_{1i}|v_{2i}) = E_s(w_i v_{1i}|v_{2i}) / E_s(w_i|v_{2i}); \tag{14}$$

$$E_U(\pi_i|v_{2i}) = 1/E_s(w_i|v_{2i}). \tag{15}$$

Adding parameter notation, it follows that the sample *pdf* can be written alternatively as,

$$\begin{aligned} f_s(y_i|x_i; \theta, \gamma) &= \frac{E_U(\pi_i|y_i, x_i; \gamma)\, f_U(y_i|x_i; \theta)}{E_U(\pi_i|x_i; \theta, \gamma)} \\ &= \frac{E_s(w_i|x_i; \theta, \gamma)\, f_U(y_i|x_i; \theta)}{E_s(w_i|y_i, x_i; \gamma)}. \end{aligned} \tag{16}$$

Rearranging yields,

$$\begin{aligned} f_U(y_i|x_i; \theta) &= \frac{E_s(w_i|y_i, x_i; \gamma)\, f_s(y_i|x_i; \theta, \gamma)}{E_s(w_i|x_i; \theta, \gamma)} \\ &= \frac{E_s(w_i|y_i, x_i; \gamma)\, f_s(y_i|x_i; \theta^*)}{E_s(w_i|x_i; \theta^*, \gamma)}. \end{aligned} \tag{17}$$

The expectations in the right-hand side of (16) and in (17) are with respect to the sample *pdf* of the base weight $w_i = (1/\pi_i)$. Thus, when the weights are known for the sampled units, which is usually the case under full response, the expectation $E_s(w_i|y_i, x_i; \gamma)$ can be modeled and estimated by regressing w_i against (y_i, x_i), using classical model fitting procedures (see below). Similarly, the sample *pdf* $f_s(y_i|x_i; \theta^*)$ can be identified and estimated using classical procedures applied to the observed data. It follows, therefore, from (17) that the population *pdf* $f_U(y_i|x_i; \theta)$ can be estimated based on only the sample outcomes and weights, without knowledge of the design variables values in Z_U or an adequate summary of them. Moreover, for given (estimated) expectations $E_s(w_i|y_i, x_i; \gamma)$, the goodness of fit of the population model can be evaluated by testing the goodness of fit of the sample model using classical techniques, since the sample model relates to the observed outcomes. See Krieger and Pfeffermann (1997) for illustrations. By definition, the sample *pdf* conditions on the selected sample and the observed covariates, unlike probability weighting.

REMARK 4.1. *Skinner (1994) proposed to extract the population model in (17) by using the following two identities: i)* $f_U(y_i|x_i, w_i) = f_s(y_i|x_i, w_i)$*, (follows from Eq. 1, see also Remark 3.2), and ii)* $f_U(w_i|x_i) = w_i f_s(w_i|x_i)/E_s(w_i|x_i)$ *(follows from Eq. 17). Use of the two identities yields,* $f_U(y_i|x_i) = \int f_U(y_i|x_i, w_i) f_U(w_i|x_i) dw_i = \int f_s(y_i|x_i, w_i) \frac{w_i f_s(w_i|x_i)}{E_s(w_i|x_i)} dw_i$*. The last expression is the same as (17). Application of this approach requires modeling* $f_s(y_i|w_i, x_i)$ *and* $f_s(w_i|x_i)$*, rather than* $f_s(y_i|x_i)$ *and* $E_s(w_i|y_i, x_i)$ *in (17).*

The relationships (16) and (17) can be extended to the case of nonresponse in two ways. If the response probabilities can be estimated with sufficient accuracy, the inclusion probability in the respondents sample can be estimated by the product of the original sample-selection probability and the estimated response probability. Estimating the sampling weight by the inverse of the estimated inclusion probability in the respondents sample yields then the Eqs. (16) and (17) as the marginal respondent *pdf*s (with the index s replaced by r), and Remark 4.1 applies to this case as well. Alternatively, the case of nonresponse can be treated by noting that (2) can be written as, $f_r(y_i|x_i) = \Pr(R_i = 1|y_i, x_i, I_i = 1) f_s(y_i|x_i)/\Pr(R_i = 1|x_i, I_i = 1)$, with $f_s(y_i|x_i)$ defined by (16). This representation requires modeling $\Pr(R_i = 1|y_i, x_i, I_i = 1)$, but the modeling process cannot be carried out by regressing R_i against (y_i, x_i) because the outcome values (and possibly the covariates) are only known for the respondents $(R_i = 1)$. Nonetheless, for a given hypothesized model for the response probabilities, the goodness of fit of the resulting respondent *pdf* can be tested by classical procedures since it refers to the observed outcomes.

Following we assume that the weights w_i are known. Suppose first that they are continuous such as in probability proportional to size (PPS) sampling with a continuous size variable. If the form of the population model is known, the expectations $E_s(w_i|y_i, x_i)$ and $E_s(w_i|x_i)$ needed for estimating the sample *pdf* (Eq. 16) can be estimated by a three-step procedure:

(1) Identify and estimate $\hat{E}_s(w_i|y_i, x_i) = E_s(w_i|y_i, x_i; \hat{\gamma})$, using the sample data.
(2) Integrate $\int [1/E_s(w_i|y, x_i; \hat{\gamma})] f_U(y|x_i; \theta) dy$ to obtain $E_U(\pi_i|x_i; \theta; \hat{\gamma})$ as a function of θ (follows from 14).
(3) Compute $\hat{E}_s(w_i|x_i; \theta, \hat{\gamma}) = 1/E_U(\pi_i|x_i; \theta, \hat{\gamma})$ (follows from 15).

On the other hand, if one is interested in estimating the population *pdf* using an estimate of the sample *pdf* $f_s(y_i|x_i;\theta^*)$ (Eq. 17) obtained by fitting a model to the sample data, then the expectation $E_s(w_i|x_i;\theta^*,\gamma)$ can be estimated by integrating $\hat{E}_s(w_i|y_i,x_i;\gamma)$ in Step 1 over the *pdf* $f_s(y_i|x_i;\hat{\theta}^*)$. See Pfeffermann and Sverchkov (1999, 2003) and Pfeffermann et al. (2006) for examples and further discussion.

Thus far, we treated the case where the sample inclusion probabilities are continuous. Estimation of the expectations $E_s(w_i|y_i,x_i;\gamma)$ and $E_s(w_i|x_i;\theta,\gamma)$ in the case of discrete inclusion probabilities is similar.

Example 4.1. Consider the case of multinomial-logistic regression with a discrete covariate x and M possible values for the outcome variable Y. Assuming that $E_s(w_i|Y_i=m,x_i=k)$ is not a function of the model parameters, it can be estimated by $\overline{w}_{mk}$, the mean of the weights in the cell (an application of the method of moments), and $\hat{\pi}_{mk}=\hat{\Pr}_U(i\in s|Y_i=m,x_i=k)=(1/\overline{w}_{mk})$. Hence,

$$\Pr_s(Y_i=m|x_i=k)=\frac{\Pr(i\in s|Y_i=m,x_i=k)\Pr_U(Y_i=m|x_i=k)}{\Pr(i\in s|x_i=k)}$$

$$\cong\frac{[\Pr_U(Y_i=m|x_i=k)/\overline{w}_{mk}]}{\sum_{m^*=1}^{M}[\Pr_U(Y_i=m^*|x_i=k)/\overline{w}_{m^*k}]}. \tag{18}$$

The sampling weights feature in the sample model, but this is not an application of classical probability weighting. Clearly, when $\overline{w}_{mk}=\text{const}$, (18) is the same as the population model. See Pfeffermann and Sverchkov (2003) for the fitting of this model. For an example of the evaluation of the expectations $E_s(w_i|x_i)$ with discrete selection probabilities but continuous outcome and explanatory variables, see Pfeffermann and Sverchkov (1999).

4.2. Independence under the sample distribution

In subsequent sections, we use an independence result established in Pfeffermann et al. (1998a). By this result, under some general regularity conditions and for many commonly used sampling schemes for selection with unequal probabilities, if the population measurements are independent, the sample measurements are "asymptotically independent" with respect to the sample model. The asymptotic framework requires that the population size increases but the sample size is held fixed.

The restriction to independent population measurements is not as restrictive as it may seem. To see why, consider again the two-level model of Example 3.5,

$$\text{Level 1:}\quad Y_{ij}=\beta_{0i}+x'_{ij}\beta+\varepsilon_{ij},\ \varepsilon_{ij}\sim N\left(0,\sigma_\varepsilon^2\right);\qquad j=1,\ldots,M_i$$

$$\text{Level 2:}\quad \beta_{0i}=l'_i\alpha+v_i,\ v_i\sim N\left(0,\sigma_v^2\right);\qquad i=1,\ldots,N, \tag{19}$$

where ε_{ij} and v_i are independent for all i and j. Suppose that n second-level units are sampled with probabilities π_i, and then m_i first-level units are sampled with probabilities $\pi_{j|i}$ from selected second-level unit i. Denote by $\lambda=(\alpha',\sigma_v^2)$ and $\theta_i=(\beta_{0i},\beta,\sigma_\varepsilon^2)$ the second- and first-level parameters, respectively. By (16), the corresponding sample

models are,

$$\text{Level 1:} \quad f_{s_i}\left(y_{ij}|x_{ij};\theta_i,\gamma_1\right) = \frac{E\left(\pi_{j|i}|y_{ij},x_{ij};\gamma_1\right) f_U\left(y_{ij}|x_{ij};\theta_i\right)}{E\left(\pi_{j|i}|x_{ij};\theta_i,\gamma_1\right)}$$

$$\text{Level 2:} \quad f_s(\beta_{0i}|l_i;\lambda,\gamma_2) = \frac{E\left(\pi_i|\beta_{0i},l_i;\gamma_2\right) f_U\left(\beta_{0i}|l_i;\lambda\right)}{E\left(\pi_i|l_i;\lambda,\gamma_2\right)}. \tag{20}$$

By the independence result, if $Y_{ij}|\theta_i$ are independent under the population model, then they are asymptotically independent under the sample model. Similarly, if the random intercepts β_{0i} are independent under the population model, then they are asymptotically independent under the sample model. Thus, the sample model defined by (20) is again a genuine two-level model, although with different distributions and more parameters. Fitting two-level models to survey data is considered in Section 6.2.

4.3. Estimation of model parameters

In what follows we assume that the sample outcomes are independent. By (16), the sample model contains the parameters θ governing the population model and the parameters γ governing the sampling weight expectation $E_s(w_i|y_i,x_i;\gamma)$. As argued previously, the latter expectation can usually be estimated by regressing w_i against (y_i,x_i), provided, of course, that the sampling weights are known. This would generally be the case if all the sample units respond, or when the response probabilities can be estimated sufficiently accurately. The estimates $\hat{\gamma}$ obtained that way can be held fixed when estimating θ. Alternatively, γ can be estimated jointly with θ, but it is important to ascertain that the sample model is identifiable. Pfeffermann et al. (1998a) showed that in Example 1.2 of this chapter not all the sample model parameters are identifiable. In situations where the response mechanism is informative, the parameters γ^* governing the response probabilities $\Pr(R_i = 1|y_i,x_i,I_i = 1;\gamma^*)$ (Eq. 2) must be estimated along with θ. In this case, the identifiability of the model can be more problematic, although it is often resolved by not having the same covariates in the population model and the model for the response probabilities, or by adding calibration constraints. In what follows we assume full response and that γ is known, and review methods for estimating the parameters $\theta = (\theta_0,\theta_1,\ldots,\theta_k)'$ governing the population model $f_U(y_i|x_i;\theta)$. We consider the case of single-stage sampling and assume that the sample outcomes are independent. Under mild conditions, θ is the unique solution of the equations,

$$W_U(\theta) = \sum_{j\in U} E_U\left(\delta_j|x_j\right) = 0, \tag{21}$$

where $\delta_j = (\delta_{j,0},\delta_{j,1},\ldots,\delta_{j,k})' = \partial \log f_U(y_j|x_j;\theta)/\partial\theta$ is the j-th score. Pfeffermann and Sverchkov (2003) considered three different approaches for estimating θ. The common feature of these approaches is that the only data used for estimation are the observations $\{(y_i,x_i,w_i), i\in s\}$. In Section 4.4, we consider the use of the "full likelihood" that assumes knowledge of the covariates $\{x_j, j\in U\}$ and possibly also additional design information.

The first approach redefines the parameter equations with respect to the sample model. Assuming that $E_s(w_i|x_i;\theta,\gamma)$ is differentiable with respect to θ, the sample model

parameter equations are

$$W_{1s}(\theta) = \sum_{i\in s} E_s\{[\partial \log f_s(y_i|x_i;\theta,\gamma)/\partial\theta]\,|x_i\} = \sum_{i\in s} E_s\{[\delta_i + \partial \log E_s(w_i|x_i;\theta,\gamma)/\partial\theta]\,|x_i\} = 0. \tag{22}$$

The vector θ is estimated under this approach by solving the equations,

$$W_{1s,e}(\theta) = \sum_{i\in s} [\delta_i + \partial \log E_s(w_i|x_i;\theta,\gamma)/\partial\theta] = 0. \tag{23}$$

Note that (23) defines the likelihood equations based on the sample model, which we refer to as the *sample likelihood*.

The second approach of estimating θ applies the relationship (14) to the parameter equations (21). For a random sample from the sample model, the parameter equations are then,

$$W_{2s}(\theta) = \sum_{i\in s} E_s(q_i\delta_i|x_i) = 0, \tag{24}$$

where $q_i = w_i/E_s(w_i|x_i)$. The vector θ is estimated under this approach by solving the equations,

$$W_{2s,e}(\theta) = \sum_{i\in s} q_i\delta_i = 0. \tag{25}$$

The third approach uses the property that if θ solves (21), then it solves also the equations, $\tilde{W}_U(\theta) = \sum_{j\in U} E_U(\delta_j) = E_X\Big[\sum_{j\in U} E_U(\delta_j|x_j)\Big] = 0$, where $E_X(\cdot)$ is the expectation of X with respect to the population distribution (viewed as random). Hence, by (14), for a random sample from the sample model, the parameter equations are $W_{3s}(\beta) = \sum_{i\in s} E_s(w_i\delta_i) = 0$, with EE,

$$W_{3s,e}(\beta) = \sum_{i\in s} w_i\delta_i = 0. \tag{26}$$

Note that the Eqs. (26) are the *pseudolikelihood* equations (Remark 3.4).

REMARK 4.2. *The difference between the estimating Eqs. (25) and (26) is that the latter use the sampling weights w_i, where as the former use the adjusted weights $q_i = w_i/E_s(w_i|x_i)$. When the sample selection probabilities depend on the covariates in x, but not on the outcome variable Y, the sampling is ignorable. Hence, it is generally only necessary to account for the net sampling effects on the target conditional pdf of $Y_i|X_i = x_i$. This is achieved by using the weights q_i. In contrast, the sampling weights w_i account for the sampling effects on the joint distribution of (Y_i, X_i). As a result, they tend to be more variable than the weights q_i. Note, in particular, that when w is a deterministic function of x (e.g., when all the variables determining the sample selection and response probabilities are included among the covariates, see Section 3.1), $w_i = E_s(w_i|x_i)$, $q_i = 1$, and the Eqs. (25) reduce to the ordinary unweighted likelihood equations $\sum_{i\in s} \delta_i = 0$. Pfeffermann and Sverchkov (1999, 2003) illustrated that estimating θ by solving the Eqs. (25) yields estimators with lower randomization variance than estimating θ by solving the Eqs. (26).*

Example 4.2. Let the population model be, $Y_j = x_j'\beta + \varepsilon_j$; $E_U(\varepsilon_j|x_j) = 0$, $E_U(\varepsilon_j^2|x_j) = \sigma_\varepsilon^2$. Solving (25) yields, $\hat{\beta}_q = \left[\sum_{i\in s} q_i x_i x_i'\right]^{-1} \sum_{i\in s} q_i x_i y_i$. Solving (26) yields, $\hat{\beta}_w = [\sum_{i\in s} w_i x_i x_i']^{-1} \sum_{i\in s} w_i x_i y_i$, which is the familiar "probability-weighted" estimator. As easily verified, the use of the weights q_i yields randomization consistent estimators for the census regression coefficients $\tilde{B} = \left\{\sum_{j\in U}\left[x_j x_j'/E_s(w_j|x_j)\right]\right\}^{-1} \sum_{j\in U}[x_j y_j/E_s(w_j|x_j)]$, and hence consistent estimators for β under the *DU* distribution, even when $E_s(w_i|x_i)$ is misspecified.

Pfeffermann and Sverchkov (2003) studied parametric and resampling estimators for the variances of the estimators obtained under the three approaches considered above. For the likelihood Eq. (23) and under some regularity conditions, the variances can be estimated by the inverse information matrix, using familiar properties of *mle*s. For the estimating Eqs. (25) and (26), the variances can be estimated under similar regularity conditions as,

$$\hat{V}_s\left(\hat{\theta}\right) = \left[\dot{W}_{ks,e}\left(\hat{\theta}\right)\right]^{-1} \left\{\sum_{i\in s}\left[a_i\delta_i\left(\hat{\theta}\right)\right]\left[a_i\delta_i\left(\hat{\theta}\right)\right]'\right\} \left[\dot{W}_{ks,e}\left(\hat{\theta}\right)\right]^{-1}, \tag{27}$$

where $\dot{W}_{ks,e}(\hat{\theta}) = [\partial W_{ks,e}(\theta)/\partial\theta]_{\theta=\hat{\theta}}$, with $k = 2$ and $a_i = q_i$ in the case of (25), and $k = 3$ and $a_i = w_i$ in the case of (26). Alternatively, the variances can be estimated by resampling methods, such as the bootstrap or the jackknife, since by the independence result of Section 4.2 the sample observations are at least approximately independent.

REMARK 4.3. *The use of the adjusted weights q_i can be justified by least squares estimation. Consider the population model,*

$$y_j = g_\theta\left(x_j\right) + \varepsilon_j,\ E_U\left(\varepsilon_j|x_j\right) = 0,\ E_U\left(\varepsilon_j^2|x_j\right) = \sigma^2 v\left(x_j\right), \tag{28}$$

where $g_\theta(\cdot)$ has a known form and the function $v(x) > 0$ is known. By (14), for any $\tilde{\theta}$ in the parameter space Θ, $\frac{1}{n}\sum_{i\in s} E_s\left\{q_i \frac{[Y_i - g_{\tilde{\theta}}(x_i)]^2}{v(x_i)}|x_i\right\} = \frac{1}{n}\sum_{i\in s} E_s\left\{q_i \frac{[Y_i - g_\theta(x_i) + g_\theta(x_i) - g_{\tilde{\theta}}(x_i)]^2}{v(x_i)}|x_i\right\} = \sigma^2 + \frac{1}{n}\sum_{i\in s}\left\{\frac{[g_\theta(x_i) - g_{\tilde{\theta}}(x_i)]^2}{v(x_i)}\right\}$, with the expectation $E_s(\cdot)$ taken over the joint sample distribution of $(w, Y)|x$. It follows that,

$$\theta = \underset{\tilde{\theta}}{\operatorname{argmin}} \frac{1}{n}\sum_{i\in s} E_s\left\{q_i \frac{\left[Y_i - g_{\tilde{\theta}}\left(x_i\right)\right]^2}{v\left(x_i\right)}\right\}. \tag{29}$$

The vector θ can be estimated, therefore, by the generalized least square (GLS) estimator,

$$\hat{\theta}_{GLS} = \underset{\tilde{\theta}}{\operatorname{argmin}} \frac{1}{n}\sum_{i\in s}\left\{\hat{q}_i \frac{\left[y_i - g_{\tilde{\theta}}\left(x_i\right)\right]^2}{v\left(x_i\right)}\right\}, \text{where } \hat{q}_i = \frac{w_i}{\hat{E}_s\left(w_i|x_i\right)}. \tag{30}$$

Under mild regularity conditions, the estimator $\hat{\theta}_{GLS}$ is consistent for θ even when the expectation $\hat{E}_s(w_i|x_i)$ in $\hat{q}_i$ is misspecified.

Magee (1998) considered the case where $g_\theta(x)$ in (28) is linear. The author showed that in this case and under certain moment assumptions, any estimator $\hat{\theta}_{h,\alpha}$ satisfying, $\hat{\theta}_{h,\alpha} = \arg\min_{\tilde{\theta}} \sum_{i\in s}\{w_i[y_i - g_{\tilde{\theta}}(x_i)]^2/h(x_i,\alpha)\}$ with positive h is consistent for θ. The

weights $h(x_i, \alpha)$ belong to a parameterized family of functions with the vector parameter α chosen to minimize a scalar variance criterion, such as the determinant or the trace of the asymptotic variance estimator $\hat{V}[\hat{\theta}_{h,\alpha}]$. The resulting "quasi-Aitken" estimator is shown to have asymptotically a lower variance under the sample distribution than the probability-weighted estimator $\hat{\beta}_w$ defined in Example 4.2. See Eq. 13 of Magee (1998) for the asymptotic variance estimator $\hat{V}[\hat{\theta}_{h,\alpha}]$ used by the author. Generalization of these results to nonlinear models, however, is not straightforward. See the discussion in Kott (2007).

4.4. The full likelihood

Theoretically, a more efficient procedure of estimating the unknown parameters is to base the likelihood on the joint *pdf* of the sample data and the sample membership indicators, that is,

$$f(I_U, y_s|x_s, x_{\tilde{s}}; \theta, \gamma) = \prod_{i\in s} \Pr(i \in s|y_i, x_i; \gamma)\, f_U(y_i|x_i; \theta) \times \prod_{j\notin s} \big[1 - \Pr\big(j \in s|x_j; \theta, \gamma\big)\big], \tag{31}$$

where $\mathrm{I}_U = \{I_1, \ldots, I_N\}$ is the vector of sample indicators and $\Pr(j \in s|x_j) = \int \Pr(j \in s|y_j, x_j) f_U(y_j|x_j) dy_j$ is the "propensity score" for unit j. The *pdf* (31) assumes that $\Pr(\mathrm{I}_U|y_U, x_U) = \prod_{k\in U} \Pr(I_k|y_k, x_k)$ (Poisson sampling). The "full likelihood" based on (31) has the advantage of accounting for the sampling probabilities of units outside the sample, but it requires knowledge of the covariates of all the population units. See, for example, Gelman et al. (2003), Pfeffermann and Sverchkov (2003), and Little (2004). Modeling the joint distribution of the covariates for units outside the sample and integrating them out of the likelihood can be very complicated and is formidable when there are many of them. Pfeffermann et al. (2006) compared empirically the use of the sample likelihood with the use of the full likelihood for multilevel models in a Bayesian framework. The two approaches yielded similar results, but this of course may not be the case in other applications.

Another way of defining the full likelihood is by application of the missing information principle (MIP, Ceppellini et al., 1955; Orchard and Woodbury, 1972). The basic idea is to express the sample score function as the conditional expectation of the population score function, given the sample data. Following Chambers (2003), define the *full-sample* likelihood as, $L_s(\lambda) = f(y_s, x_s, I_U, z_U; \lambda)$, where (y_s, x_s) represent the observed outcomes and covariates and z_U is the known matrix of the population values underlying the sample selection (see Section 2). The corresponding *full-population* likelihood is $L_U(\lambda) = f(y_U, x_U, I_U, z_U; \lambda)$, where $y_U = (y_s, y_{\tilde{s}})$ and $x_U = (x_s, x_{\tilde{s}})$. The MIP principle states that

$$sc_s(\lambda) = (\partial/\partial\lambda) \log[L_s(\lambda)] = E_U[(\partial/\partial\lambda) \log L_U(\lambda)\,|y_s, x_s, I_U, z_U]. \tag{32}$$

A similar identity defines the relationship between the population information matrix and the sample information matrix.

Breckling et al. (1994) and Chambers et al. (1998) considered applications of the MIP to complex survey data. In particular, Chambers et al. (1998) studied the use of the MIP when only limited design information is available rather than the full information

entailed in Z_U. The authors showed examples where the use of the MIP is more efficient than the use of the sample likelihood (22), which does not use any design information other than the weights $\{w_i, i \in s\}$.

REMARK 4.4. *Sections 4.1 and 4.3 assume full response. As noted before, in the case of nonignorable nonresponse, the parameters governing the response process need to be estimated along with the parameters governing the population pdf, and possibly also the parameters governing the sample-selection process (Eq. 2). An important advantage of the use of "sample likelihood" or the "full likelihood" is that they do not require in principle knowledge of the sample selection and response probabilities, although they require in this case modeling the two selection processes as functions of all the available data.*

5. Prediction under informative sampling

5.1. The sample-complement distribution

In this section, we consider the prediction of outcome values for units outside the sample with application to the prediction of finite population totals. Prediction of small area means for sampled and nonsampled areas is considered in Section 6.3. Notice that under informative sampling, the sample-complement, $\tilde{s} = U - s$, is likewise an informative sample and the model holding for units $j \in \tilde{s}$ is different in this case from the sample model, and also from the population model, although as shown in Remark 5.1, the latter difference can often be ignored when the population is large and the sampling fraction is small. In what follows we assume single-stage sampling and full response, but the definitions and results of this section can be extended to the case of nonresponse, similarly to the extensions considered in previous sections.

Sverchkov and Pfeffermann (2004) defined the conditional marginal sample-complement *pdf* for units outside the sample as,

$$f_{\tilde{s}}(y_i|x_i) \stackrel{\text{def}}{=} f_U(y_i|x_i, I_i = 0) = \frac{\Pr(I_i = 0|y_i, x_i)\, f_U(y_i|x_i)}{\Pr(I_i = 0|x_i)}. \tag{33}$$

Using the relationships between the population *pdf* and the sample *pdf* in Section 4.1 and the equality $\Pr(I_i = 0|y_i, x_i) = 1 - \Pr(I_i = 1|y_i, x_i) = 1 - E_U(\pi_i|y_i, x_i)$, the marginal sample-complement *pdf* and its expectation can be written as,

$$f_{\tilde{s}}(y_i|x_i) = \frac{E_U[(1-\pi_i)|y_i, x_i]\, f_U(y_i|x_i)}{E_U[(1-\pi_i)|x_i]} = \frac{E_s[(w_i - 1)|y_i, x_i]\, f_s(y_i|x_i)}{E_s[(w_i - 1)|x_i]}, \tag{34}$$

$$E_{\tilde{s}}(Y_i|x_i) = \frac{E_U[(1-\pi_i) Y_i|x_i]}{E_U[(1-\pi_i)|x_i]} = \frac{E_s[(w_i - 1) Y_i|x_i]}{E_s[(w_i - 1)|x_i]}. \tag{35}$$

REMARK 5.1. *As with the population pdf, the sample-complement pdf is determined by the sample pdf and the expectation $E_s(w_i|y_i, x_i)$, both of which can be modeled and estimated from the sample using classical inference techniques. Note that unless $E_U(\pi_i|y_i, x_i) = E_U(\pi_i|x_i)$, the population, sample and sample-complement pdfs are all*

different. Nonetheless, for small sampling fractions such that $\pi_i < \delta$ for all $i \in U$ with probability 1,

$$f_{\tilde{s}}(y_i|x_i) = f_U(y_i|x_i) + \frac{E_U\{[E_U(\pi_i|x_i) - \pi_i]\,|y_i, x_i\} f_U(y_i|x_i)}{E_U[(1-\pi_i)\,|x_i]}$$
$$= f_U(y_i|x_i)\,(1+\Delta)\,, \tag{36}$$

where $-\delta < \Delta < \delta/(1-\delta)$. It follows from (36) that in the common case where δ is very small, the difference between the population pdf and the sample-complement pdf is also small, which of course is not surprising. This, however, is not always the case and in small area estimation, for example, the number of selected areas may actually exceed the number of unselected areas, implying that in this case predicting the area means for nonsampled areas based on the population model can yield large biases. See Section 6.3. Another example is informative nonresponse where the probability to respond is ordinarily high and hence the distribution of the outcomes for the nonrespondents can be very different from the population distribution. The imputation of the missing values has to be based in this case on the distribution of the outcomes for nonresponding units.

REMARK 5.2. *The definition of the sample-complement pdf generalizes to the joint pdf of two or more measurements associated with nonsampled units. In particular, defining similarly to Section 1 by $\tilde{A}_{\tilde{s}}$ the sample-complement indicator such that $\tilde{A}_{\tilde{s}} = 1$ if the sample-complement is $\tilde{s}$ and $\tilde{A}_{\tilde{s}} = 0$ otherwise, and denoting the data associated with $\tilde{s}$ by $(Y_{\tilde{s}}, X_{\tilde{s}})$, the joint sample-complement pdf of $Y_{\tilde{s}}$ for given $X_{\tilde{s}} = x_{\tilde{s}}$ is then,*

$$f_{\tilde{s}}(y_{\tilde{s}}|x_{\tilde{s}}) = f_U\Big(y_{\tilde{s}}|x_{\tilde{s}}, \tilde{A}_{\tilde{s}} = 1\Big) = \frac{\Pr\Big(\tilde{A}_{\tilde{s}} = 1|y_{\tilde{s}}, x_{\tilde{s}}\Big)\, f_U(y_{\tilde{s}}|x_{\tilde{s}})}{\Pr\Big(\tilde{A}_{\tilde{s}} = 1|x_{\tilde{s}}\Big)}. \tag{37}$$

5.2. Prediction of finite population totals under informative sampling

Model-based prediction of finite population means or totals under noninformative sampling is treated in Chapter 23. Let $T = \sum_{j=1}^{N} Y_j$ be the population total. Denote the "design information" available for prediction by $D_s = \{(y_i, w_i), i \in s; (x_j, I_j), j = 1 \ldots N\}$, and let $\hat{T} = \hat{T}(D_s)$ be a predictor of T. The mean square error (MSE) of $\hat{T}|D_s$ with respect to the population *pdf* is

$$\mathrm{MSE}\Big(\hat{T}|D_s\Big) = E_U\left[\Big(\hat{T} - T\Big)^2 |D_s\right] = E_U\left\{\Big[\hat{T} - E_U(T|D_s)\Big]^2 |D_s\right\} + V_U(T|D_s)$$
$$= \Big[\hat{T} - E_U(T|D_s)\Big]^2 + V_U(T|D_s)\,. \tag{38}$$

It follows from (38) that $\mathrm{MSE}(\hat{T}|D_s)$ is minimized when $\hat{T} = E_U(T|D_s)$. Now,

$$E_U(T|D_s) = \sum_{i\in s} E_U(Y_i|D_s, I_i = 1) + \sum_{j\notin s} E_U\big(Y_j|D_s, I_j = 0\big)$$
$$= \sum_{i\in s} y_i + \sum_{j\notin s} E_{\tilde{s}}\big(Y_j|x_j\big), \tag{39}$$

where the last equality assumes that Y_j for $j \in \tilde{s}$ and D_s are uncorrelated given x_j. The prediction problem reduces, therefore, to the estimation of the expectations $E_{\tilde{s}}(Y_j|x_j)$, which can be assessed by use of (35).

Let the sample-complement model take the general form,

$$Y_j = C_\beta\left(x_j\right) + \epsilon_j, \quad E_{\tilde{s}}\left(\epsilon_j|x_j\right) = 0,$$
$$E_{\tilde{s}}\left(\epsilon_j^2|x_j\right) = \sigma^2 v\left(x_j\right), \quad E_{\tilde{s}}\left(\epsilon_k\epsilon_j|x_k, x_j\right) = 0, \quad k \neq j, \tag{40}$$

where $C_\beta(x)$ is a known (possibly nonlinear) function of x that is governed by an unknown vector parameter β. The variance function $v(x) > 0$ is assumed known. Under mild conditions, the vector β satisfies, $\beta = \arg\min_{\tilde{\beta}} \sum_{i\in s} E_{\tilde{s}}\left\{\frac{[Y_i - C_{\tilde{\beta}}(x_i)]^2}{v(x_i)}\right\}$; where the expectation is with respect to the sample-complement *distribution* of Y. By (35), $\beta = \arg\min_{\tilde{\beta}} \sum_{i\in s} E_s\left\{\tilde{r}_i \frac{[Y_i - C_{\tilde{\beta}}(x_i)]^2}{v(x_i)}\right\}$; with $\tilde{r}_i = \frac{w_i - 1}{E_s(w_i) - 1}$ and the expectation taken with respect to the joint sample distribution of (w, Y). Noting that $E_s(w_j) = \text{const}$, β can be estimated as, $\hat{\beta} = \arg\min_{\tilde{\beta}} \sum_{i\in s} (w_i - 1)\frac{[y_i - C_{\tilde{\beta}}(x_i)]^2}{v(x_i)}$. The predictor of the population total T is then, $\hat{T} = \sum_{i\in s} y_i + \sum_{j\notin s} C_{\hat{\beta}}(x_j)$, where $C_{\hat{\beta}}(x_j)$ is obtained from $C_\beta(x_j)$ by substituting $\hat{\beta}$ for β.

Example 5.1. An important special case of the predictor $\hat{T}$ occurs when $C_\beta(x_j)$ in (40) is linear with an intercept and $v(x_j) = \text{const}$. Let $x_j' = (1, \tilde{x}_j')$ and $\beta' = (\beta_{\tilde{s}0}, \beta_{\tilde{s}}')$. Denoting $T_{\tilde{s}}(\tilde{x}) = \sum_{i\notin s} \tilde{x}_i$ and $[\hat{T}_{\tilde{s}}, \hat{T}_{\tilde{s}}(\tilde{x})] = [(N - n)/\sum_{i\in s}(w_i - 1)]$ $[\sum_{i\in s}(w_i - 1)(y_i, \tilde{x}_i)]$, the predictor takes in this case the form, $\hat{T}_{\text{Reg}} = \sum_{i\in s} y_i + \hat{T}_{\tilde{s}} + \tilde{B}_{\tilde{s}}'[T_{\tilde{s}}(\tilde{x}) - \hat{T}_{\tilde{s}}(\tilde{x})]$, where $\tilde{B}_{\tilde{s}}$ is a probability-weighted estimator for the vector coefficient of $\beta_{\tilde{s}}$, but with weights $(w_i - 1)$ instead of the base weights w_i. As is easily verified, $\hat{T}_{\text{Reg}}$ is randomization consistent for the realized population total T. This predictor can be obtained also as a special case of the *cosmetic* predictors proposed by Brewer (1999b), and it only requires knowledge of the sampled covariates and their population mean, but notice that the development of the cosmetic predictors and their MSE assumes noninformative sampling.

REMARK 5.3. *Under the sample-complement model (40) with some added mild conditions, β satisfies also, $\beta = \arg\min_{\tilde{\beta}} \frac{1}{n}\sum_{i\in s} E_s\left\{r_i \frac{[Y_i - C_{\tilde{\beta}}(x_i)]^2}{v(x_i)}\middle| x_i\right\}$; where $r_i = \frac{(w_i - 1)}{E_s(w_i|x_i) - 1}$ and the expectation is taken with respect to the joint sample distribution of $(w, Y)|x$ (follows similarly to the derivation of Eq. 29). A more efficient estimator of β is therefore, $\hat{\beta} = \arg\min_{\tilde{\beta}} \sum_{i\in s}\left(\hat{r}_i \frac{[y_i - C_{\tilde{\beta}}(x_i)]^2}{v(x_i)}\right)$; with $\hat{r}_i = \frac{(w_i - 1)}{\hat{E}_s(w_i|x_i) - 1}$. This estimator, however, requires modeling and estimating the expectation $E_s(w_j|x_j)$. See the discussion in Remark 4.2. Sverchkov and Pfeffermann (2004) considered other models and corresponding predictors.*

Next consider the case of no auxiliary variables ($x_k = 1$ for all k). Then, by (39) and (35), $\hat{T} = \sum_{i\in s} y_i + (N - n)\hat{E}_s\left[\frac{w_j - 1}{\hat{E}_s(w_j) - 1} Y_j\right]$. Estimating the two sample expectations in the second term by the respective sample means (an application of the method of moments) yields the predictor, $\hat{T} = \sum_{i\in s} y_i + \frac{(N-n)}{\sum_{i\in s}(w_i - 1)} \sum_{i\in s}(w_i - 1) y_i$. For sampling

designs such that $\sum_{i\in s} w_i = N$ for all s, or when estimating $\hat{E}_s(w_i) = N/n$, the latter predictor reduces to the H-T estimator $\hat{T}_{H\text{-}T} = \sum_{i\in s} w_i y_i$.

Example 5.2. For small sampling fractions, it is often sensible to predict all the population y-values by their expectation under the population distribution, rather than just predict the y-values for the nonsampled units by their sample-complement expectations. Consider again the case of no auxiliary variables. By (14), $E_U(Y_i) = E_s(w_i Y_i)/E_s(w_i)$. Estimating the two expectations by their sample means yields the Hajek estimator, $\hat{T}_{\text{Hajek}} = \sum_{k=1}^{N} \hat{E}_p(Y_i) = N\sum_{i\in s} w_i y_i / \sum_{i\in s} w_i$.

5.3. Mean square error estimation

Estimating $\text{MSE}(\hat{T}|D_s) = E_U[(\hat{T} - T)^2|D_s]$ for a predictor $\hat{T}$ requires strict model assumptions that may be hard to validate. To deal with this problem, Sverchkov and Pfeffermann (2004) proposed replacing the model expectation by the randomization expectation, that is, estimating instead $\text{MSE}_{D_s}(\hat{T}) = E_{D_s}[(\hat{T} - T)^2|Y_U = y_U, x_U]$, where E_{D_s} defines the randomization expectation over all possible realizations of D_s, that is, over all possible sampled y-values and their corresponding x-values and base weights, with the population values of x and y held fixed. Estimation of the randomization MSE of the various predictors has the additional advantage of allowing their use under the design-based approach.

Sverchkov and Pfeffermann (2004) considered the following two-step MSE estimator (assuming, as before, single-stage sampling).

Step 1. Generate a single "pseudo population" by selecting *with replacement* N units from the original sample with probabilities proportional to $w_i = 1/\pi_i$, where N is the population size. The justification for this step is given below. Denote by T_{pp} the sum of the y-values in the pseudo population.

Step 2. Select independently a large number B of samples from the pseudo population generated in Step 1, using the same sampling scheme as used for the selection of the original sample and reestimate the population total.

Let $\hat{T}$ represent any of the predictors and $\hat{T}_{\text{pp}}^b$ the predictor obtained for sample b. Estimate,

$$\hat{\text{MSE}}_{D_s}\left(\hat{T}\right) = \frac{1}{B}\sum_{b=1}^{B}\left(\hat{T}_{\text{pp}}^b - T_{\text{pp}}\right)^2. \tag{41}$$

The performance of the estimator (41) in estimating the randomization MSE depends on the "closeness" of the pseudo population generated in Step 1 to the actual population from which the original sample was drawn. The closeness of the two populations can be verified in part by noting that the marginal distribution of $Y_i|x_i$ in the pseudo population on any given draw is the same as in the original population. To see this, note that the pseudo population generated in Step 1 is a sample with replacement from the original sample with selection probabilities Cw_i on each draw, where $C = 1/\sum_{j=1}^{n} w_j$. Denoting by $f_{\text{pp}}(y_i|x_i)$ the marginal pseudo population *pdf* on a given draw, we obtain by (16) and (15),

$$f_{\text{pp}}(y_i|x_i) = \frac{E_s(Cw_i|y_i, x_i)\, f_s(y_i|x_i)}{E_s(Cw_i|x_i)} = \frac{E_U(\pi_i|x_i)\, f_s(y_i|x_i)}{E_U(\pi_i|y_i, x_i)} = f_U(y_i|x_i). \tag{42}$$

REMARK 5.4. *As with the standard bootstrap method for variance estimation, a successful application of this procedure requires that the original sample size is sufficiently large and that the sample measurements are approximately independent (see Section 4.2).*

REMARK 5.5. *Step 1 is similar and asymptotically equivalent to duplicating sample unit i, w_i times. Notice, however, that the use of the duplication procedure does not yield pseudo populations of size N unless $\sum_{i=1}^{n} w_i = N$. It is also not clear how to establish the relationship (42) when using this procedure.*

6. Other applications of the sample distribution

In this section, we discuss more elaborate applications of the sample distribution, all of which assuming full response. Where appropriate, we first review corresponding probability weighting approaches. Another application, the analysis of longitudinal data, is discussed in Chapter 32.

6.1. Nonparametric estimation under informative sampling

Nonparametric and semiparametric estimation with complex survey data are discussed in Chapter 27. Chambers et al. (2003) considered two alternative classes of nonparametric estimators for $g(x) = E_U(Y_i|X_i = x)$ under informative sampling. The first class is applicable for sampling designs where for each inclusion probability π_i corresponds a sizeable subsample of units, all selected with the same probability. Classical examples are stratified or cluster sampling schemes with all the units in the same stratum or cluster being selected with equal probabilities.

By (14) and repeated application of Bayes theorem,

$$g(x) = \frac{E_s[w_i E_s(Y_i|X_i = x, w_i)\, f_s(x|w_i)]}{E_s[w_i f_s(x|w_i)]}. \tag{43}$$

Thus, for given $E_s(Y_i|X_i = x, w_i)$ and $f_s(x|w_i)$, $g(x)$ can be expressed as the ratio of two sample-based expectations, where both expectations (the external expectation in the numerator and the expectation in the denominator) are with respect to the sample *pdf* $f_s(w_i)$. The expectation $E_s(Y_i|X_i = x, w_i)$ is with respect to the sample distribution and thus can be estimated from the sample data. The same is true for $f_s(x|w_i)$ when for each sampling weight w_i corresponds a sizeable number of sampled units. For example, in the case of stratified sampling, one can use a smooth nonparametric estimate $\hat{f}_s(x|w_i)$ within each stratum. Estimating the external expectation in the numerator and the expectation in the denominator by the corresponding sample means yields then the estimator

$$\hat{g}(x) = \frac{\sum_{i\in s} w_i \hat{E}_s(Y_i|X_i = x, w_i)\, \hat{f}_s(x|w_i)}{\sum_{i\in s} w_i \hat{f}_s(x|w_i)}. \tag{44}$$

The second class of estimators considered by Chambers et al. (2003) is obtained by writing, using (14),

$$E_{U,e} = E_U[Y_i - g(x_i)] = E_s\{w_i\,[Y_i - g(x_i)]\}/E_s(w_i) = 0, \tag{45}$$

implying $E_s\{w_i[Y_i - g(x_i)]\} = 0$ since $E_s(w_i) = \text{const}$. Estimating $E_s\{w_i[Y_i - g(x_i)]\}$ by a kernel-based estimate with kernel $K(\cdot)$ and bandwidth $b(x)$ defines the EE,

$$E_{se} = \sum_{i\in s} w_i K\left[\frac{x - x_i}{b(x)}\right][y_i - g(x)] = 0, \tag{46}$$

which in turn yields the nonparametric estimate,

$$\hat{g}_w(x) = \sum_{i\in s} w_i y_i K\left[\frac{x - x_i}{b(x)}\right] \Big/ \sum_{i\in s} w_i K\left[\frac{x - x_i}{b(x)}\right]. \tag{47}$$

REMARK 6.1. *An alternative estimator to (47) is obtained by writing, using again (14),*

$$\tilde{E}_{U,e|x} = E_U\{[Y_i - g(x_i)]\,|x_i\} = E_s\{q_i[Y_i - g(x_i)]\,|x_i\} = 0, \tag{48}$$

where, as before, $q_i = w_i/E_s(w_i|x_i)$. Following the same steps as above yields the alternative nonparametric estimator,

$$\hat{g}_q(x) = \sum_{i\in s} q_i y_i K\left[\frac{x - x_i}{b(x)}\right] \Big/ \sum_{i\in s} q_i K\left[\frac{x - x_i}{b(x)}\right]. \tag{49}$$

Simulation results show that the estimators (47) and (49) perform similarly, which can be explained by the fact that the sample totals $\sum_{i\in s} w_i y_i K\left[\frac{x-x_i}{b(x)}\right]$ and $\sum_{i\in s} w_i K\left[\frac{x-x_i}{b(x)}\right]$ can be viewed as kernel-based estimates for $E_s[w_i Y_i|x_i]$ and $E_s[w_i|x_i]$, respectively.

6.2. *Multilevel modeling under informative sampling*

Consider the following general two-level population model:

$$\begin{aligned} &\text{Level 1:} \quad Y_{ij}|\left(\beta_{0i}, x_{ij}\right) \sim f_U\left(y_{ij}|x_{ij}, \beta_{0i}; \theta_1\right), \quad j = 1, \ldots, M_i \\ &\text{Level 2:} \quad \beta_{0i}|l_i \sim \varphi_U(\beta_{0i}|l_i; \theta_2), \quad i = 1, \ldots, N, \end{aligned} \tag{50}$$

where f_U and φ_U denote the first- and second-level *pdf*s with known covariates x_{ij} and l_i, and unknown hyperparameters θ_1 and θ_2, respectively. The model defined by (19) is a special case of (50) by which f_U and φ_U are normal densities with $\theta_1 = (\beta, \sigma_\varepsilon^2)$ and $\theta_2 = (\alpha, \sigma_v^2)$. The problem is to estimate $\theta = (\theta_1, \theta_2)$ under informative sampling of first- and/or second-level units. Specifically, consider the following two-stage sampling process. In the first stage, a sample s of $n < N$ second-level units (say, schools) is selected with probabilities $\pi_i = \Pr(i \in s)$ that may be correlated with the random effects β_{0i} after conditioning on the covariates l_i. In the second stage, a subsample s_i of $m_i < M_i$ first-level units (say, pupils) is sampled from each selected second-level unit i with probabilities $\pi_{j|i} = \Pr(j \in s_i | i \in s)$ that may be correlated with the outcomes Y_{ij} after conditioning on the covariates x_{ij}.

In example 3.5, we described the computation of the *pmle* estimators proposed by Pfeffermann et al. (1998a) for the model defined by (19). Grilli and Pratesi (2004) applied a *pmle* approach to multilevel models of ordinal and binary outcomes.

Both studies propose scaling the sampling weights to reduce the bias in the case of small first-level sample sizes. More generally, Asparouhov (2006) proposed estimating the model hyperparameters in (50) by maximizing the probability-weighted likelihood,

$$l(\theta_1, \theta_2) = \prod_{i=1}^{n} \left\{ \int \left[\prod_{j=1}^{m_i} f_U\left(y_{ij}|x_{ij}, \beta_{0i}; \theta_1\right)^{w_{j|i}c_{1i}} \right] \varphi_U(\beta_{0i}|l_i; \theta_2) d\beta_{0i} \right\}^{w_i c_{2i}}, \qquad (51)$$

where $w_{j|i} = 1/\pi_{j|i}$ and $w_i = 1/\pi_i$ are the base weights and (c_{1i}, c_{2i}) are first- and second-level scaling constants. The author compares several methods of weight scaling. The estimators proposed by Pfeffermann et al. (1998a) and Grilli and Pratesi (2004) can be viewed as special cases of the family of estimators obtained from maximizing the likelihood (51) with specific choices of the scaling factors, but they applied different maximization algorithms. Rabe-Hesketh and Skrondal (2006) extended the likelihood (51) to more than two levels using adaptive quadrature for approximating the integrals in the pseudolog likelihood. As in the previously cited articles, the authors studied the bias of the *pmle* and proposed weight scaling for bias reduction. Korn and Graubard (2003) considered the one-way random effects model $Y_{ij} = \beta + \beta_{0i} + \varepsilon_{ij}$, obtained from (19) by dropping the covariates at both levels. The authors proposed estimating the variances σ_v^2 and σ_ε^2 by method of moments type estimators that require knowledge of the pairwise selection probabilities $\pi_{jk|i} = \Pr(j, k \in s_i | i \in s)$ within the selected second-level units.

We now turn to the use of the sample model for multilevel modeling under informative sampling. The basic idea has already been outlined for the linear two-level model (19). Rather than fitting the population model using probability weighting, one derives the sample model at each level, using (20). The latter equations apply to any two-level model with $f_U(y_{ij}|x_{ij}; \theta_i)$ and $f_U(\beta_{0i}|l_i; \lambda)$ defining the first- and second-level models, respectively. As noted in Section 4.2, the model defined by (20) is a genuine two-level model with approximately independent random effects under mild conditions, but it generally has a more complicated structure and contains more unknown parameters than its population counterpart.

Fitting the model (20) requires modeling $E(\pi_{j|i}|y_{ij}, x_{ij}; \gamma_1)$ and $E(\pi_i|\beta_{0i}, l_i; \gamma_2)$. Pfeffermann et al. (2006) considered the model (19) and illustrated the modeling of the two expectations for a particular informative sampling design. The authors estimated the model hyperparameters, including the parameters $\gamma = (\gamma_1, \gamma_2)$ using a Bayesian framework with Markov Chain Monte Carlo (MCMC) simulations. They compared the Bayesian estimators with the *pmle* estimators of Pfeffermann et al. (1998a) by way of a simulation study. The general conclusion of this study is that both procedures yield approximately unbiased estimators for most of the parameters over repeated sampling from the *UD* distribution (see Section 1.2), the exception being σ_v^2. For a small number of second-level units, both methods produce biased estimators for this variance component. However, the bias when using the Bayesian methodology applied to the sample model tends to be much smaller. Moreover, using the latter procedure generally yields better credibility intervals than the conventional *t*-based confidence intervals obtained by use of the *pmle* estimators and their standard errors. Recall, however, that the computation of the *pmle* only requires specifying the population model.

6.3. Small area estimation under informative sampling of areas and within areas

Model-based small area estimation uses multilevel models for the prediction of area means or other quantities of interest. The use of a model overcomes the problem of having small samples in at least some of the areas (and possibly no samples in other areas) by linking the data in the various areas via the model equations. This allows borrowing strength across the sampled areas and enables predicting the target quantities for areas with no samples. See Chapter 32 for a comprehensive account of small area models and inference methods in common use. Our interest in this section is in situations where the selection of the sampled areas and/or the sampling schemes within the selected areas are informative.

A possible way of handling informative sampling within the selected areas is by modeling the direct, design-based probability-weighted estimators for the quantities of interest instead of the individual (first-level) observations. For example, modeling the randomization unbiased H-T estimators $\hat{\bar{Y}}_{\mathrm{HT},i} = \sum_{j\in s_i} w_{j|i} y_{ij}/M_i$, where s_i defines the sample from area i, M_i is the area size and $w_{j|i}$ are the unit level sampling weights. Kott (1989), Arora and Lahiri (1997) and Prasad and Rao (1999) model probability-weighted direct estimators using the random effects model $\tilde{\theta}_i = \theta_i + e_i$; $\theta_i = x_i'\beta + \beta_{0i}$. In this model $\tilde{\theta}_i$ is the probability-weighted estimator of the true area mean θ_i, which is linked to a vector of area level covariates x_i with an added error (random effect) β_{0i}. This model is a special case of the two-level model (19) considered in previous sections. See Chapter 32 for further discussion and uses of this model for small area estimation.

Restricting to direct design-based estimators as the input data results in loss of efficiency if first-level individual observations (y_{ij}, x_{ij}) are available. Malec et al. (1999) considered informative sampling of unit-level observations and used a "probability-weighted" marginal likelihood in a Bayesian framework for inference. The authors used data from the U.S. health survey NHANES III (http://www.cdc.gov/nchs/about/major/nhanes/nh3data.htm), for estimating overweight prevalence in counties. This is implemented by fitting logistic models with fixed age/race/gender effects and correlated normal random race/gender effects. To account for informative sampling within the selected counties, they estimated the sampling probabilities using the sampling weights and used them for constructing the probability-weighted likelihood. The county prevalence estimates are obtained by combining this likelihood with appropriate prior distributions and applying the Bayesian methodology with the aid of MCMC simulations for computing the county posterior proportions of overweights.

None of the preceding studies considers informative selection of areas. How can the sample and sample-complement models be used for small area estimation under informative sampling of areas and within the selected areas? Suppose that the population is divided into N areas of which n areas are selected with probabilities $\pi_i = \Pr(i \in s)$, and that from each selected area i of size M_i, a sample of m_i units is drawn with probabilities $\pi_{j|i} = \Pr(j \in s_i | i \in s)$. Denote by $w_i = 1/\pi_i$ and $w_{j|i} = 1/\pi_{j|i}$ the corresponding base weights and let the population model have the general form (50). The target population parameters are the true small area means, $\bar{Y}_i = \sum_{j=1}^{M_i} Y_{ij}/M_i$ for $i = 1 \ldots N$, (the means in sampled and nonsampled areas). We denote by $D_s = \{(y_{ij}, w_{j|i}, w_i), (i, j) \in s; x_{lm}, (l, m) \in U\}$ the known data used for estimation. The MSE of a predictor $\hat{\bar{Y}}_i$ with respect to the population *pdf*, given D_s and I_i ($I_i = 1$ if area i is sampled, $I_i = 0$

otherwise) is

$$\begin{aligned}\mathrm{MSE}\left(\hat{\overline{Y}}_i|D_s, I_i\right) &= E_U\left[\left(\hat{\overline{Y}}_i - \overline{Y}_i\right)^2 |D_s, I_i\right] \\ &= \left[\hat{\overline{Y}}_i - E_U\left(\overline{Y}_i|D_s, I_i\right)\right]^2 + V_U\left(\overline{Y}_i|D_s, I_i\right), \end{aligned} \tag{52}$$

implying that the MSE is minimized when $\hat{\overline{Y}}_i = E_U(\overline{Y}_i|D_s, I_i)$.

Denote by $f_s(\beta_{0i}|\cdot)$, $[(f_{\tilde{s}}(\beta_{0i}|\cdot))]$ the conditional sample (sample-complement) *pdf* of the random effects β_{0i}, with expectation operator $E_s(E_{\tilde{s}})$, and by $f_{si}(y_{ij}|\cdot)$, $[f_{\tilde{s}i}(y_{ij}|\cdot)]$. The conditional sample (sample-complement) *pdf* of the outcomes y_{ij}, with expectation operator $E_{si}(E_{\tilde{s}i})$. Assume the mild condition $f_{\tilde{s}i}(y_{il}|D_s, \beta_{0i}, I_i = 1) = f_{\tilde{s}i}(y_{il}|x_{il}, \beta_{0i}, I_i = 1)$, (unobserved outcomes in a sampled area are independent of the observed outcomes and their sampling weights when conditioning on the area random effect and the covariates). Using (14) and (35), Pfeffermann and Sverchkov (2007) showed that for area i in the sample:

$$\begin{aligned}E_U\left(\overline{Y}_i|D_s, I_i = 1\right) &= \frac{1}{M_i}\left\{\sum_{j\in s_i} y_{ij} + \sum_{l\notin s_i} E_s\left[E_{\tilde{s}i}(Y_{il}|x_{il}, \beta_{0i}, I_i = 1)\,|D_s\right]\right\} \\ &= \frac{1}{M_i}\left\{\sum_{j\in s_i} y_{ij} + \sum_{l\notin s_i} E_s\left[\left.\frac{E_{si}\left[\left(w_{l|i} - 1\right) Y_{il}|x_{il}, \beta_{0i}, I_i = 1\right]}{E_{si}\left[\left(w_{l|i} - 1\right)|x_{il}, \beta_{0i}, I_i = 1\right]}\right| D_s\right]\right\}. \end{aligned} \tag{53}$$

For area i not in the sample:

$$\begin{aligned}E_U\left(\overline{Y}_i|D_s, I_i = 0\right) &= \frac{1}{M_i}\sum_{k=1}^{M_i} E_{\tilde{s}}[E_U\left(Y_{ik}|x_{ik}, \beta_{0i}, I_i = 1\right)|D_s] \\ &= \frac{1}{M_i}\sum_{k=1}^{M_i} \frac{E_s\left[\left(w_i - 1\right)\dfrac{E_{si}\left(w_{k|i}Y_{ik}|x_{ik}, \beta_{0i}, I_i = 1\right)}{E_{si}\left(w_{k|i}|x_{ik}, \beta_{0i}, I_i = 1\right)}|D_s\right]}{E_s[(w_i - 1)|D_s]}. \end{aligned} \tag{54}$$

It follows from (53) and (54) that the computation of the predictors for sampled and non-sampled areas requires modeling $f_{si}(y_{ij}|x_{ij}, \beta_{0i}, I_i = 1)$, $E_{si}(w_{j|i}|y_{ij}, x_{ij}, \beta_{0i}, I_i = 1)$, and $f_s(\beta_{0i}|D_s)$, and estimating the unknown parameters featuring in these models. All these densities and expectations refer to the sample data such that the modeling and estimation can be carried out using classical model fitting techniques. The expectations $E_s(\cdot)$ can be estimated by simple averaging over the selected areas. Informative selection of areas occurs when the area sampling weights w_i are correlated with the area random effects β_{0i}, but no model is assumed relating the two terms.

Pfeffermann and Sverchkov (2007) illustrated the computation of the predictors (53) and (54) assuming that the sample model identified for the observed outcomes is $Y_{ij} = x'_{ij}\beta + \beta_{0i} + e_{ij}$; $\beta_{0i}|I_i = 1 \overset{\text{ind}}{\sim} N\left(0, \sigma^2_\beta\right)$, $e_{ij}|I_{ij} = 1 \overset{\text{ind}}{\sim} N\left(0, \sigma^2_e\right)$, and $E_{si}\left(w_{j|i}|y_{ij}, x_{ij}, \beta_{0i}, I_i = 1\right) = k_i \exp\left(a'x_{ij} + by_{ij}\right)$. The predictors for sampled and

nonsampled areas in this case are

$$\hat{E}_U\big(\overline{Y}_i|D_s, I_i = 1\big) = \frac{1}{M_i}\left\{(M_i - m_i)\,\hat{\theta}_i + m_i\left[\overline{y}_i + \big(\overline{X}_i - \overline{x}_i\big)'\,\hat{\beta}\right] + (M_i - m_i)\,\hat{b}\hat{\sigma}_e^2\right\}, \tag{55}$$

$$\hat{E}_U\big(\overline{Y}_i|D_s, I_i = 0\big) = \overline{X}_i'\hat{\beta} + \hat{b}\hat{\sigma}_e^2 + \left[\sum_{i\in s}(w_i - 1)\,\hat{\beta}_{0i}\Big/\sum_{i\in s}(w_i - 1)\right], \tag{56}$$

where $(\overline{y}_i, \overline{x}_i)$ are the sample means of y and x in area i, $\overline{X}_i$ is the true area mean of x, $\hat{\beta}_{0i} = \hat{\gamma}_i\left[\overline{y}_i - \overline{x}_i'\hat{\beta}\right]$; $\hat{\gamma}_i = \hat{\sigma}_\beta^2\big/\big[\hat{\sigma}_\beta^2 + \hat{\sigma}_e^2/m_i\big]$, and $\hat{\theta}_i = \hat{\beta}_{0i} + \overline{X}_i\hat{\beta}$ is the empirical best linear unbiased predictor (EBLUP) of $\theta_i = \overline{X}_i'\beta + \beta_{0i} = E_{si}(\overline{Y}_i|X_i, \beta_{0i})$. The authors estimated b by fitting the model $E_{si}(w_{j|i}|x_{ij}, y_{ij}, I_i = 1) = k_i\exp(a'x_{ij} + by_{ij})$ to the sample data, $(\hat{\sigma}_\beta^2, \hat{\sigma}_e^2)$ by the method of moments and $\hat{\beta}$ by GLS with the unknown variances replaced by their estimators.

The terms $\{(M_i - m_i)\hat{b}\hat{\sigma}_e^2\}$ in (55) and $\hat{b}\hat{\sigma}_e^2$ in (56) correct for sampling effects within the selected areas, that is, the difference between the sample-complement expectation and the sample expectation. Notice that for $b = 0$, the sampling within the selected areas is ignorable, and the predictor (55) reduces to the EBLUP predictor under noninformative sampling (Chapter 32). The last term in (56) corrects for the fact that under informative selection of areas, the mean of the random effects for areas outside the sample is different from zero.

Pfeffermann and Sverchkov (2007) developed bootstrap MSE estimators for the predictors (55) and (56). They also outlined the basic steps in computing the predictors under a general two-level sample model fitted to the sample data, with continuous or discrete outcomes and fixed and random effects, allowing for a general model relating the unit sampling weights to the outcomes.

7. Tests of sampling ignorability

The methods described in this chapter involve the use of the sampling weights or other design and response information with various degrees of complexity. It is clear, therefore, that when the sample selection and response are in fact noninformative, the estimators obtained by these methods are more entangled and often more variable than classical model-dependent estimators that ignore the sample selection and nonresponse. For the complex sampling designs in common use, it is generally difficult and often impractical to check directly the conditions described in Section 2 that permit ignoring the sampling process (the sample selection and response). This suggests the need for test procedures that can guide the analyst whether the sampling process is ignorable for the type of inference intended.

Several test procedures have been proposed in the literature, mostly in connection to linear regression analysis. A common feature of these tests is that they compare probability-weighted estimators for the target parameters with the ordinary (equally weighted) estimators that ignore the sampling process. For instance, when estimating the population regression model as in Example 4.2, one can use the test statistic,

$$\lambda = \left(\hat{\beta}_{\mathrm{OLS}} - \hat{\beta}_w\right)'\,\hat{\mathrm{Var}}\left(\hat{\beta}_{\mathrm{OLS}} - \hat{\beta}_w\right)^{-1}\left(\hat{\beta}_{\mathrm{OLS}} - \hat{\beta}_w\right), \tag{57}$$

where $\hat{\beta}_{\text{OLS}}$ and $\hat{\beta}_w$ are the OLS and the probability-weighted estimators, respectively. The null hypothesis that the sampling process can be ignored for estimating β can be formulated as $H_0 : E_s(Y_i|X_i = x_i) = x_i'\beta = E_U(Y_i|X_i = x_i)$. Note that both estimators are unbiased for β under H_0, but whereas $\hat{\beta}_w$ is consistent for β even under informative sampling (follows from (26), see also Example 4.2), $\hat{\beta}_{\text{OLS}}$ is generally only consistent for β under noninformative sampling. For sufficiently large samples, the distribution of the test statistic (57) under H_0 can be approximated by the chi-square distribution with $f = \dim(\beta)$ degrees of freedom. See, for example, Fuller (1984) for an application of this test in the case of a stratified multistage sample with $\hat{\text{Var}}(\hat{\beta}_{\text{OLS}} - \hat{\beta}_w)$ computed under the randomization distribution. DuMouchel and Duncan (1983) proposed testing the ignorability of the sampling process for estimating β by i) augmenting the design matrix X_s of the regression model by the columns W_sX_s, where $W_s = \text{Diag}(w_1, \ldots, w_n)$, ii) fitting the unweighted regression $\hat{Y}_s = X_s\hat{\beta}_{\text{OLS}} + W_sX_s\hat{\gamma}_{\text{OLS}}$ and, iii) testing $H_0 : \gamma = 0$ using the conventional F-statistic. Note that the use of this test assumes homoscedastic, uncorrelated residuals. Nordberg (1989) extended the DuMouchel-Duncan test to generalized linear models. Pfeffermann (1993) extended the use of the test statistic of the form (57) to other models with unknown vector parameters. Chambers et al. (2003) introduced similar tests for nonparametric estimation. Pfeffermann and Nathan (1985) proposed simple test statistics based on cross-validation techniques.

REMARK 7.1. *Like with any other test procedures, nonrejection of the null hypothesis could be a result of a type 2 error. Moreover, even if the sample selection and nonresponse can be ignored for the estimation of a given vector parameter, it does not necessarily imply that the sampling process has no effect on other features of the population model, such as the distribution of the residuals. On the other hand, the null hypothesis can be rejected even if the sampling process is ignorable, either because of a type 1 error or because the population model is misspecified.*

Next, we consider the use of the sample model for testing sampling ignorability. Suppose first that we want to test whether the population and sample expectations of $Y|x$ are the same, that is, $E_U(Y_i|x_i) = E_s(Y_i|x_i)$ or equivalently, $\text{Cov}_s(Y_i, w_i|x_i) = 0$ (the equivalence follows from Eq. 14). The equality of the two expectations can be assessed, therefore, by testing $\text{Corr}_s(Y_i, w_i|x_i) = 0$. This can be implemented most conveniently by regressing $w_i = \delta_0 + h(x_i, \delta_x) + \delta_y y_i + \eta_i$ with some appropriate function $h(x, \delta_x)$ and unknown coefficients $\delta = (\delta_0, \delta_x, \delta_y)$, and then testing $\delta_y = 0$ using the conventional t-statistic. The test refers to the sample distribution, such that the unknown coefficients can be estimated using standard techniques like OLS.

REMARK 7.2. *By (16), $f_U(y_i|x_i) = f_s(y_i|x_i)$ iff $E_s(w_i|y_i, x_i) = E_s(w_i|x_i)$, implying that if $E_s(w_i|y_i, x_i)$ is correctly specified, testing $E_s(w_i|y_i, x_i) = E_s(w_i|x_i)$ actually tests the more general hypothesis that the marginal population and sample pdfs are the same.*

Pfeffermann and Sverchkov (2007) applied this test for testing the informativeness of the sample selection in a small-area estimation context. Similar tests can be applied for testing the equality of moments of the distribution of the population model residuals and the corresponding moments of the distribution of the sample model residuals. This

enables testing more closely whether the distribution of the population model residuals is unaffected by the sampling process (see Remark 7.1). Note, however, that these tests refer to the marginal distribution of the residuals. Pfeffermann and Sverchkov (1999) applied such test procedures in the context of regression analysis. See Chambers et al. (2003) for other similar tests.

Eideh and Nathan (2006) proposed testing the hypothesis $E_s(w_i|y_i, x_i) = E_s(w_i|x_i)$ by using the Kullback-Leibler information measure. They applied the test for the case where $E_s(w_i|y_i, x_i) = \exp(a_0 + a_x x_i + a_y y_i)$. The Kullback-Leibler measure compares two density functions. In classical applications, samples are available from the two densities, but when applied for testing sampling ignorability, no data are directly observed from the population density. Extension of this test to more general situations should be investigated.

Another, more general test procedure uses the EE discussed in Section 4.3, comparing the EE that ignore the sampling process with the EE that account for it. The results of this testing procedure can guide the researcher which EE to use in practice. Consider the population model parameter equations, $W_U(\theta) = \sum_{j\in U} E_U[\delta_j|x_j] = 0$, where $\delta_j = (\delta_{j,0}, \delta_{j,1}, \ldots, \delta_{j,k})' = \partial \log f_U(y_j|x_j; \theta)/\partial\theta$ is the jth score [Eq. (21)]. A plausible method of accounting for the sampling process discussed in Section 4.3 uses instead the sample-based analog, $\sum_{i\in s} E_s(q_i\delta_i|x_i) = 0$, where $q_i = w_i/E_s(w_i|q_i)$ [Eq. (24)]. Thus, the sampling process can be ignored for the estimation of θ if $R_n = n^{-1}\sum_{i\in s} R(x_i) = 0$, where $R(x_i) = E_s(\delta_i|x_i) - E_s(q_i\delta_i|x_i)$. Pfeffermann and Sverchkov (2003) proposed testing $H_0 : R_n = 0$ using the Hotelling statistic,

$$H(\hat{R}) = \frac{n-(k+1)}{k+1}\hat{R}_n'\hat{S}_n^{-1}\hat{R}_n, \tag{58}$$

with $\hat{R}_n = n^{-1}\sum_{i\in s}\hat{R}(x_i)$, $\hat{R}(x_i) = (\hat{\delta}_i - \hat{q}_i\hat{\delta}_i)$, $S_n = n^{-1}\sum_{i\in s}(\hat{R}(x_i) - \hat{R}_n)(\hat{R}(x_i) - \hat{R}_n)'$, $\hat{\delta}_i = \partial \log f_U(y_i|x_i; \theta)/\partial\theta_{|\theta=\hat{\theta}}$ and $\hat{q}_i = w_i/\hat{E}_s(w_i|x_i)$. Under the null hypothesis and for sufficiently large samples, $H(\hat{R}) \sim F_{k+1,n-(k+1)}$. The estimator $\hat{\theta}$ needed for the evaluation of the score can be estimated under H_0 by assuming sampling ignorability. Pfeffermann and Sverchkov (2003) applied the statistic (58) for testing the ignorability of the sampling process when fitting multinomial-logistic models. Another test along these lines is proposed by Wu and Fuller (2005) for the case of linear regression. The authors propose testing $E_U(Y_i|x_i) = E_s(Y_i|x_i)$ using the relationship $E_U(Y_i|x_i) = E_s(q_iY_i|x_i)$, (follows from (14)). They showed that in this case $E_s(Y_i|x_i) = x_i'\beta + q_ix_i'\gamma$ and hence $E_U(Y_i|x_i) = E_s(Y_i|x_i)$ iff $\gamma = 0$. The sampling ignorability can be assessed, therefore, by testing the significance of $\hat{\gamma}_{\mathrm{OLS}}$.

8. Brief summary

In this chapter, we discuss the approaches proposed in the literature to deal with informative probability sampling, commenting also on how they can possibly be extended to account for informative nonresponse. The approaches differ in the kind of data required for their application (knowledge of the population values of the design variables or adequate summary of them, or just the sample observations and the sampling weights), the level at which the model has to be specified (population model or the sample model),

the extra modeling steps required, the use of the randomization distribution as part of the inference process, the type of inference accommodated by them (point estimation, prediction…), and of course, statistical efficiency. The different approaches differ also in computation demands and intensity, and the skills and knowledge required from the analysts applying them.

We find the use of the sample model (Sections 4–7) to be very flexible in terms of data requirements (it only requires in principle knowledge of the sample data) and inference possibilities, but it does require modeling the sampling and response probabilities as functions of the observed data. Including the design variables among the covariates allows using classical model-based methods, but the population values of these variables are often unknown to the analyst fitting the model. The use of probability weighting (Section 3.3), likewise only requires knowledge of the observed data, and it does not require modeling the sample selection probabilities, but it is limited mostly to point estimation of population model parameters. The use of this approach does not avoid modeling the response probabilities. We strongly recommend experimenting with the use of these approaches, applying them to the same data sets, to further study their properties and advantages.

Acknowledgements

We are grateful to Fred Smith and Phil Kott for thoroughly reading an early version of this chapter and making many constructive comments and suggestions.

Sample Surveys: Inference and Analysis, Vol. 29B
ISSN: 0169-7161
© 2009 Elsevier B.V. All rights reserved
DOI: 10.1016/S0169-7161(09)00240-5

Asymptotics in Finite Population Sampling

Zuzana Prášková and Pranab Kumar Sen

1. Introduction

In *finite population sampling* (FPS), the theory of objective or probabilistic sampling plays a fundamental role in statistical inference. In practice, a set or collection of a *finite* number (say, N) of objects or *units* comprise a *population,* and on the basis of an objectively chosen subset of these units, called a *sample*, the task is to draw valid statistical conclusions on various characteristics of the population. The population size N, though finite, may typically be large, and the sample size, say n, though presumably less than N, may or may not be small compared with N (i.e., the *sampling fraction* (SF) n/N may not be very small). In *survey sampling*, the *sampling frame* defines the units and the size of the survey population from which the sample is taken unambiguously. It also reconstructs a population having an uncountable (or infinite) number of natural units in terms of a finite population by redefining suitable *sampling units.* Thus, given the sampling frame and units, one may like to draw inference on the population through an objective sampling scheme. In some other cases, though the units are clearly defined, the size of the population may not be known and needs to be estimated along with its other characteristics. In either case, sampling is usually made *without replacement* (WOR), generally, leading to relatively smaller margins of sampling fluctuations, though, for sampling *with replacement* (WR), the theory is relatively simpler in form. Again, in either scheme, the different units in the population may all have a common probability for inclusion in the sample, leading to *equal probability or simple random sampling* (SRS), or they may be stratified into some subsets, for each of which SRS may be applied. In the extreme case, the units in the population, depending on their sizes or some other characteristics, may possibly have different probabilities for inclusion in the sample (i.e., *varying probability sampling)*. There are other variations, such as *cluster sampling* (possibly within strata), *double sampling, inter-penetrating sampling, successive sampling,* etc., which are all characterized by an objective procedure defined by a probability law governing the drawing of the sample units. (See Part 1 of this handbook for several alternative sampling methods.)

For small N, sampling distributions of statistics can be studied by direct enumeration of all possible cases. However, for large N, as is typically the case in survey sampling, this enumerational process generally becomes prohibitively laborious. Even

if N is large, n/N may not be very small, and in such cases, there may be a profound need to examine appropriate large sample approximations for sampling distributions and related probability inequalities, incorporating them in applications. There may be another scenario wherein N is large but not n, where the asymptotics could be different from the case where the SF is not close to 0. These asymptotics in FPS constitute our main objective. The asymptotic theory depends on the sampling design, and for diverse schemes, diverse techniques have been used to achieve the general goals. It is intended to provide here a general account of these asymptotics.

This chapter is a revisited and updated version of Sen (1988), published in *Handbook of Statistics*, Volume 6. During the past 20 years, new developments have occurred in this field that will be incorporated here. As such, we plan to present some results of the earlier sections of Sen (1988) more succinctly, adding the later developments and new sections. In Section 2, we start with the asymptotics in *simple random sampling without replacement* (SRSWOR) schemes. For U-statistics, containing the sample mean and variances as special cases, asymptotic normality and related results are shortly presented there. Asymptotics in stratified sampling WOR and some results from *simple random sampling with replacement* (SRSWR) are also included. Some asymptotics on probability inequalities in SRS are briefly appended to Section 2. In Section 3, enlarged and updated, we deal with asymptotics in *resampling* methods in FPS. *Capture-mark-release-recapture* (CMRR) techniques and asymptotic results on the estimation of the size of a finite population are briefly presented in Section 4. Asymptotic results on *sampling with varying probabilities* are treated in greater detail. Limit theorems along with the allied *coupon collector problem* are kept in Section 5 that is completed with latter results. Section 6 deals with asymptotics in maximum entropy, especially in *rejective* (conditional Poisson) and *order*, especially Pareto, sampling schemes. Section 7 is devoted to asymptotics arising in *successive subsampling with varying probabilities without replacement* (SSVPWOR). In the concluding section, we shortly discuss results that are not reviewed in the previous sections.

2. Asymptotics in SRS

SRSWOR is characterized by a population set $\boldsymbol{A}_N = (a_1, \ldots, a_N)$ with N units and a sample $\boldsymbol{X}_n = (X_1, \ldots, X_n)$ of size n drawn WOR so that $(X_1, \ldots, X_n)$ is a (random) subset of $\boldsymbol{A}_N$, governed by the probability law

$$P\{X_1 = a_{i_1}, \ldots, X_n = a_{i_n}\} = N^{-[n]} \tag{1}$$

for every $1 \le i_1 \ne \cdots \ne i_n \le N$, where $N^{-[n]} = (N^{[n]})^{-1}$ and $N^{[n]} = N \cdots (N-n+1)$ for $n \le N$ ($N^{[0]} = 1$). Based on $\boldsymbol{X}_n$, we may be interested in the estimation of the *population mean* and *population variance*

$$\overline{a}_N = N^{-1} \sum_{i=1}^{N} a_i \quad \text{and} \quad \sigma_N^2 = (N-1)^{-1} \sum_{i=1}^{N} [a_i - \overline{a}_N]^2, \tag{2}$$

among other characteristics of the population. The unbiased minimum variance estimators based on $\boldsymbol{X}_n$ (viz., Nandi and Sen, 1963) are given by

$$\overline{X}_n = n^{-1}\sum_{i=1}^{n} X_i \quad \text{and} \quad s_n^2 = (n-1)^{-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2, \tag{3}$$

respectively, both being special cases of U-statistics, introduced as follows. For a function $g(X_1, \ldots, X_m)$ symmetric in its m arguments, $m \geq 1$, we define a (population) *parameter*

$$\theta_N = \theta(\boldsymbol{A}_N) = N^{-[m]} \sum_{1 \leq i_1 \neq \cdots \neq i_m \leq N} g(a_{i_1}, \ldots, a_{i_m}). \tag{4}$$

The corresponding sample function, viz.,

$$U_n = n^{-[m]} \sum_{1 \leq i_1 \neq \cdots \neq i_m \leq n} g(X_{i_1}, \ldots, X_{i_m}) \tag{5}$$

termed a *U-statistic (with the kernel g of degree m)* is an unbiased minimum variance estimator of θ_N (Nandi and Sen, 1963). In fact, as $g(\cdot)$ is assumed to be symmetric in its $m \geq 1$ arguments, in (4) and (5), we may take $1 \leq i_1 < \cdots < i_m \leq N$ (and $1 \leq i_1 < \cdots < i_m \leq n$) and replace $N^{-[m]}$ (and $n^{[-m]}$) by $\binom{N}{m}^{-1}$ (and $\binom{n}{m}^{-1}$). Note that the X_i are not independent random variables (r.v.). Nevertheless, they are symmetric dependent r.v. For known $\boldsymbol{A}_N$, the exact sampling distribution of U_n may be obtained by direct enumeration of all possible $\binom{N}{n}$ samples of size n from $\boldsymbol{A}_N$. Obviously, this process becomes prohibitively laborious for large N (and n). As such, there is a genuine need to provide suitable approximations to the large sample distribution when N and n are both large, though n/N (the SF) needs not be very small. In this context, the permutational central limit theorems (PCLT) play a vital role. For the particular case of $\overline{X}_n$, Madow (1948) initiated the use of PCLT in FPS, and since then, this has been an active area of fruitful research. This has an immediate connection with linear rank statistics, for which the most general PCLT is due to Hájek (1961). For a systematic review of Hájek asymptotics in FPS and connection to the theory of linear rank test statistics, we may refer to Sen (1995) or Prášková and Sen (1998). For general U_n, the asymptotic normality result for SRS has been studied by Nandi and Sen (1963), with further generalizations due to Sen (1972), Krewski (1978), Majumdar and Sen (1978), Zhao and Chen (1990), and others.

Though theoretically the asymptotic theory is justified for N indefinitely large, in practice the asymptotic approximations work out quite well for N even moderately large. It has been reported in detail that under very general regularity conditions (Nandi and Sen, 1963), as $N \to \infty, n \to \infty$, and $n/N \to \alpha$, $\sqrt{[n/(1-\alpha)]}(U_n - \theta_N)$ has closely a normal distribution with zero mean and variance which is itself a function of the population elements. Actually, more general invariance principles have been established, permitting the asymptotic normality even for random sample sizes. These asymptotic results extend readily to (i) several U-statistics (vector case) and (ii) functions of several U-statistics. In that generality, the case of stratified sampling plans (Bickel and Freedman, 1984; Krewski and Rao, 1981), as well as the so-called ratio and regression estimators (Fuller, 1975; Scott and Wu, 1981), studied as separate problems, is covered by the asymptotic theory developed for U-statistics.

We mention *stratified sampling without replacement* in more detail. Suppose that a population $\boldsymbol{A}_N$ of size N is divided into H disjoint strata $\boldsymbol{A}_{N_1}, \ldots, \boldsymbol{A}_{N_H}$ of sizes $N_1, \ldots, N_H$, $N = N_1 + \cdots + N_H$. Let $\boldsymbol{X}_{n_h} = (X_{h1}, \ldots, X_{hn_h})$ be a sample of size n_h selected from the population $\boldsymbol{A}_{N_h}$ by SRSWOR. Samples $\boldsymbol{X}_{n_h}$ are selected independently from each stratum. The total sample size is $n = n_1 + \cdots + n_H$. Let $\overline{a}_{N_h}$ and $\sigma^2_{N_h}$ be the mean and the variance in population $\boldsymbol{A}_{N_h}$ according to (2), and let $\overline{X}_{n_h}$ and $s^2_{n_h}$ be their sample counterparts as given in (3).

The stratified population mean $\theta = \frac{1}{N}\sum_{h=1}^H N_h \overline{a}_{n_h} = \sum_{h=1}^H w_h \overline{a}_{n_h}$, where $w_h = \frac{N_h}{N}$, can be estimated by $\widehat{\theta} = \sum_{h=1}^H w_h \overline{X}_{n_h}$. The variance of this unbiased estimator is

$$\tau^2 = \mathsf{Var}\,\widehat{\theta} = \sum_{h=1}^H w_h^2\, \mathsf{Var}\,(\overline{X}_{n_h}) = \sum_{h=1}^H w_h^2 \frac{\sigma^2_{N_h}}{n_h} \cdot \frac{N_h - n_h}{N_h}, \tag{6}$$

and an unbiased estimator of τ^2 is

$$\widehat{\tau}^2 = \sum_{h=1}^H w_h^2 \frac{s^2_{n_h}}{n_h} \cdot \frac{N_h - n_h}{N_h}. \tag{7}$$

We can easily see that both $\widehat{\theta}$ and $\widehat{\tau}^2$ are linear combinations of independent U-statistics and thus the asymptotic theory developed for U-statistics could be applied. Bickel and Freedman (1984) proceed in a rather direct way. They supposed that the number of strata H and the population and sample sizes N_h and n_h depend on an index ν such that $n(\nu) = n_1(\nu) + \cdots + n_H(\nu) \to \infty$ as $\nu \to \infty$ in any way whatsoever, for example, many small samples or a few larger samples or some combinations of both are possible. Assuming that $2 \le n_h \le N_h - 1$ for all h, they proved that both the standardized stratified sample mean $(\widehat{\theta} - \theta)/\tau$ and the studentized version $(\widehat{\theta} - \theta)/\widehat{\tau}$ converge to the standard normal distribution under a generalized Feller-Lindeberg condition

$$\frac{\sum_{h=1}^H \alpha_h \sum_{j \in B_{h\epsilon}} \left(a_{hj} - \overline{a}^2_{N_h}\right)}{\sum_{h=1}^H \alpha_h \sum_{j=1}^{N_h} (a_{hj} - \overline{a}_{N_h})^2} \to 0 \tag{8}$$

as $\nu \to \infty$, where

$$B_{h\epsilon} = \left\{1 \le j \le N_h : |a_{hj} - \overline{a}_{N_h}| \ge \epsilon \frac{\tau n_h}{w_h}\right\} \tag{9}$$

and $\alpha_h = N_h(N_h - n_h)/n_h(N_h - 1)$. (We suppressed the dependence on ν for simplicity.) Bickel and Freedman (1984) obtained the result for any linear combination of strata means. Case $H = 1$ (one stratum) coincides with earlier asymptotic results for SRSWOR by Erdös and Rényi (1959) and Hájek (1960). It may be also remarked that in connection with resampling methods, second-order asymptotics for the distribution of the stratified sample mean were considered by Chen and Sitter (1993) as $n \to \infty$ and (i) $H^2 \to \infty$ faster than n_h for each h or (ii) H bounded or $H^2 \to \infty$ but not faster than n_h. Bloznelis (2007) established a similar result for general U-statistics based on stratified sampling WOR where H remains bounded.

We consider now the case of SRSWR where we have $X_1, \ldots, X_n$ independent and identically distributed (i. i. d.) r.v., and

$$P\{X_1 = a_i\} = N^{-1} \quad \text{for } i = 1, \ldots, N. \tag{10}$$

As such, the classical central limit theorems and weak convergence results for U-statistics (see Hoeffding, 1948; Miller and Sen, 1972) remain applicable. However, in SRSWR, the classical estimators of the population mean and variance, considered in (3), are not optimal in the sense that there are other estimators (based on the distinct units in the sample) which have smaller variance (or risk with a convex loss function). In a SRSWR(N, n), let ν_n be the number of distinct units (so that $1 \le \nu_n \le n$), and let $\overline{X}_{(\nu_n)}$ and $s^2_{(\nu_n)}$ be the sample mean and variance based on these distinct units. Then, it is known that $\overline{X}_{(\nu_n)}$ has a smaller risk than $\overline{X}_n$ and a similar result holds for the variance estimators, although $s^2_{(\nu_n)}$ is not an unbiased estimator of σ^2 (but it can be made unbiased by introducing a multiplicative factor $c(n, \nu_n)$). The better performance characteristics of these estimators are mainly due to the Basu (1958) sufficiency of the number of distinct units in SRSWR. We refer to Sinha and Sen (1989) for some deeper results in this direction.

Further, we will discuss the asymptotics in SRSWR based on the distinct units in the sample. First, we may note that in a SRSWR(N, n),

$$P\{\nu_n = k\} = N^{-n}\binom{N}{k}(\Delta^k 0^n) \quad \text{for } k = 1, \ldots, n, \tag{11}$$

where $\Delta^k a^q = (a+k)^q - \binom{k}{1}(a+k-1)^q + \cdots + (-1)^k\binom{k}{k}a^q$ for $a \ge 0, q \ge 0$ and $k \ge 1$. As such, it is easy to verify that for every $n \ge 1$,

$$E(\nu_n) = N\{1 - (1 - N^{-1})^n\}, \tag{12}$$

$$E(\nu_n^{-1}) = N^{-n}\sum_{k=1}^{N}(N-k+1)^{n-1}. \tag{13}$$

Thus, if $\lim_{N\to\infty}(n/N) = \alpha,\ 0 < \alpha < \infty,$ then

$$\lim_{N\to\infty}\{N^{-1}E(\nu_n)\} = 1 - e^{-\alpha}, \qquad \lim_{N\to\infty}\{NE(\nu_n^{-1})\} = (1 - e^{-\alpha})^{-1}. \tag{14}$$

If we put $Z_n = \nu_n/E(\nu_n)$, then from (12)–(14) we obtain that both EZ_n and $E(Z_n^{-1})$ converge to 1 as n increases. On the other hand, Z_n is a positive valued r.v., and we know that for every positive x, $(x + x^{-1})/2 \ge 1$, where the strict equality holds only for $x = 1$. Thus, noting that $[EZ_n + EZ_n^{-1}]/2 \to 1$ as n increases, we immediately conclude from the above inequality that Z_n converges to 1 in probability, as $n \to \infty$. Consequently, we have for a SRSWR(N, n), as $n/N \to \alpha,\ 0 < \alpha < \infty$,

$$n^{-1}\nu_n \xrightarrow{P} (1 - e^{-\alpha})/\alpha. \tag{15}$$

In this context, we may note that

$$(1 - e^{-\alpha})/\alpha < 1 \text{ for every } \alpha > 0, \tag{16}$$

where for small values of α, the left-hand side of (16) is close to 1. Further, we may note that the distribution of ν_n is independent of the population units $\{a_1, \ldots, a_N\}$ and hence, given $\nu_n = k \ge 1$, the probability distribution of the distinct units (say, $X_1', \ldots, X_k'$) is the same as in the SRSWOR(N, k). Thus, given $\nu_n = k$, we are in a position to adapt all the asymptotic results for the classical situation in SRSWOR(N, k). Finally, (15)

ensures that $n^{-1}\nu_n \xrightarrow{P} (1-e^{-\alpha})/\alpha$ as n (or N) increases with $n/N \to \alpha > 0$ so that we can again use the central limit theorem for random sample sizes (and its ramifications for U-statistics, discussed in Miller and Sen, 1972) and conclude that the asymptotic results on the distribution of U-statistics for SRSWOR(N, n) all extend smoothly for the distinct sample unit-based U-statistics, provided that we replace $n/N(\simeq \alpha)$ by $1-(1-N^{-1})^n(\simeq 1-e^{-\alpha})$. In view of (16), we conclude that unless α is very small, the use of the distinct units in SRSWR(N, n) generally leads to some increase in the efficiency of the estimators.

One important application of the asymptotics in FPS discussed so far is in the area of probability and moment inequalities for SRSWOR. These are discussed in detail in Sen (1988), and hence, we present only a few important ones. Reverse martingale property for SRSWOR (Sen, 1970) allows to derive the Hájek-Rényi-Chow inequality for U-statistics in FPS without any extra condition. This enables us to prove asymptotic results like strong laws of large numbers and establishes strong consistency of sample statistics (see, e.g., Sen and Singer, 1993, Chapter 2.4). Further, coupled with invariance principles, it provides sharper asymptotic inequalities that are useful for asymptotic analysis in other context, too. In particular, let us present the following: let $\{d_{Ni};\, 1 \le i \le N,\, N \ge 1\}$ be a triangular array of real numbers satisfying the normalizing constraints

$$\sum_{i=1}^{N} d_{Ni} = 0 \text{ and } \sum_{i=1}^{N} d_{Ni}^2 = 1. \tag{17}$$

Also, let $q = \{q(t) : 0 \le t \le 1\}$ be a continuous, nonnegative, U-shaped, and square integrable function inside $I = [0, 1]$. Finally, let $\boldsymbol{Q} = (Q_1, \ldots, Q_N)$ take on each permutation of $(1, \ldots, N)$ with the common probability $(N!)^{-1}$. Then

$$P\left\{\max_{1\le k\le N} q(k/N)\Big|\sum_{i=1}^{k} d_{NQ_i}\Big| \ge 1\right\} \le \int_0^1 q^2(t)\mathrm{d}t. \tag{18}$$

Clearly, in SRSWOR, $\sum_{i=1}^{k} d_{NQ_i}$ is a standardized sample sum of $d_{Ni} = (a_i - \bar{a}_N)/\sum_{i=1}^{N}(a_i - \bar{a}_N)^2$, and (18) may be used to provide a simultaneous (in $k, 1 \le k \le N$) confidence band for the population mean by choosing q in an appropriate way. For a related inequality (exploiting the fourth moment but not the inherent martingale structure), we may refer to Hájek and Šidák (1967, p. 185):

$$P\left\{\max_{1\le k\le n}\Big|\sum_{i=1}^{k} d_{NQ_i}\Big| \ge t\right\} \le \frac{n}{N}\left[\max_{1\le i\le N} d_{Ni}^2 + 3\frac{n}{N}\right]\left(1-\frac{n}{N}\right)^{-3} t^{-4}(1+\varepsilon_N), \tag{19}$$

where $\varepsilon_N \to 0$ as $N \to \infty$.

Generally, we may obtain better bounds by exploiting the weak convergence results in Section 2 of Sen (1988) for large values of N. As in Sen (1972), we consider the case of general U-statistics so that the case of sample means (or sums) can be obtained as a particular one. Note that by virtue of the weak convergence result, stated after formula (13) there, we have, for every $t > 0$ and $n,\ n/N \to \alpha,\ 0 < \alpha \le 1$,

$$\lim_{N\to\infty} P\left\{\max_{m\le k\le n} k|U_k - \theta_N| \ge tm[N\bar{\zeta}_{1,N}]^{1/2}\right\} = P\left\{\sup_{0\le u\le\alpha} |W^o(u)| \ge t\right\}, \tag{20}$$

where $W^o = \{W^o(t), 0 \le t \le 1\}$ is a Brownian bridge and $\overline{\zeta}_{1,N}$ is an estimable parameter, explicitly defined in (2.7)–(2.8) in Sen (1988). We note that for $m = 1$ and $g(x) = x, \overline{\zeta}_{1,N} = (1 - n^{-1})\sigma_N^2$. In general, $\overline{\zeta}_{1,N}$ can be consistently estimated by jackknifing from U_n (Nandi and Sen, 1963). Noting that $W^o(s/(s+1)) = (s+1)^{-1}W(s)$, $s \ge 0$, where $W = \{W(t), t \ge 0\}$ is a standard Brownian motion process on $[0, \infty)$, we may rewrite the right-hand side of (20) as

$$P\left\{\sup_{0\le u\le \alpha/(1-\alpha)} |(u+1)^{-1}W(u)| \ge t\right\}. \tag{21}$$

An upper bound for (21) is given by

$$P\left\{\sup_{0\le u<\infty} |(u+1)^{-1}W(u)| \ge t\right\} = 2\sum_{k=1}^{\infty}(-1)^{k+1}\exp(-2k^2t^2). \tag{22}$$

For small values of α as is usually the case encountered in practice, we may get a better bound of (21):

$$\begin{aligned} P\left\{\sup_{0\le u\le \alpha/(1-\alpha)} |(u+1)^{-1}W(u)| \ge t\right\} &\le P\left\{\sup_{0\le u\le \alpha/(1-\alpha)} |W(u)| \le t\right\} \\ &\le 4P\{W(\alpha/(1-\alpha)) \ge t\} = 4[1 - \Phi(t(1/\alpha - 1)^{1/2})], \end{aligned} \tag{23}$$

where $\Phi(\cdot)$ is the standard normal d.f. In particular, for kernels of degree 1, (23) may be compared with (19), and, as $1 - \Phi(x)$ converges to 0 exponentially, as $x \to \infty$, usually (23) performs much better than (19). We conclude this section with the remark that for moderate values of N, the equality sign in (20) may generally be replaced by a less than or equality sign so that we may have a conservative property for small values of N.

3. Resampling in FPS: Asymptotics

Resampling methods, including the jackknife and the bootstrap, can provide standard error estimates and nonparametric confidence intervals for the parameters of interest or approximate sampling distributions of statistics. Though originally developed for i. i. d. r.v., they have been extensively studied in complex sample surveys.

In SRS or other sampling plans, regression or other estimators, *jackknife* was mainly introduced to serve a dual purpose: to reduce the bias of estimators (which are typically of the nonlinear form) and to provide an efficient and asymptotically normally distributed estimator of the sampling variance of the (jackknifed) estimator. In the same setup as in Section 2, for a general estimator $T_n = T_n(X_1, \ldots, X_n)$ (containing U-statistics U_n as a special case), we may define the *pseudovalues* $T_{n,i} = nT_n - (n-1)T_{n-1}^{(i)}$, $i = 1, \ldots, n$, where $T_{n-1}^{(i)} = T_{n-1}(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)$. Then the jackknifed estimator of a parameter θ_N is defined by

$$T_n^* = n^{-1}(T_{n,1} + \cdots + T_{n,n}), \tag{24}$$

and the Tukey (1958) form of the jackknifed variance estimator is given by

$$S_n^2 = (n-1)^{-1}\sum_{i=1}^{n}(T_{n,i} - T_n^*)^2. \tag{25}$$

To motivate the jackknifed estimator, we may start with a possibly biased estimator T_n for which we may have

$$ET_n = \theta_N + n^{-1}a_1(N) + n^{-2}a_2(N) + \cdots, \tag{26}$$

where the $a_j(N)$ are real numbers depending possibly on the population size N and the set A_N. Using $n-1$ for n in (26) for each $T_{n-1}^{(i)}$ and (24), we obtain that under (26),

$$ET_n^* = \theta_N - a_2(N)/n(n-1) + \cdots = \theta_N + \mathcal{O}(n^{-2}). \tag{27}$$

Thus, the bias of T_n is reduced from $\mathcal{O}(n^{-1})$ to that of $\mathcal{O}(n^{-2})$ for T_n^*. In addition to this important feature of "bias reduction," the variance estimator S_n^2 also plays a very important role in drawing statistical conclusions on θ_N. Since in this chapter, we are primarily concerned with the *asymptotics* in FPS, we shall mainly restrict ourselves to the discussion of the large sample properties of T_n^* and S_n^2.

Keeping in mind the ratio, regression, and other estimators (which are all expressible as functions of some U-statistics), we conceive a general estimator T_n of the form $T_n = h(\boldsymbol{U}_n)$ where $h(\cdot)$ is a smooth function and $\boldsymbol{U}_n$ is a vector of U-statistics, defined as in Section 2. Also, we keep in mind the conditional (permutational) distribution generated by the $n!$ equally likely permutations of $X_1, \ldots, X_n$ among themselves and define $\mathcal{F}_n$ as the sigma-field generated by the order collection of $X_1, \ldots, X_n$. Then it follows from the basic results in Majumdar and Sen (1978) that

$$T_n^* = T_n + (n-1)E[(T_n - T_{n-1})|\mathcal{F}_n] \quad \forall n > m, \tag{28}$$

$$S_n^2 = n(n-1)\mathrm{Var}[(T_n - T_{n-1})|\mathcal{F}_n] \quad \forall n > m. \tag{29}$$

Thus, for both the jackknifed estimator T_n^* and the variance estimator S_n^2, the inherent permutational distributional structure provides the access for the necessary modifications. This theoretical justification for jackknifing has been elaborately studied in Sen (1977). To fix the notations, we let $\boldsymbol{\mu}_N = E\boldsymbol{U}_n$, and (in a matrix setup) we assume that

$$\sup_N E\|\boldsymbol{g}(X_1, \ldots, X_m)\|^4 < \infty, \quad m = \max(m_1, \ldots, m_p), \tag{30}$$

where $\boldsymbol{g}(\cdot)$ stands for the vector of kernels of degrees $m_1, \ldots, m_p$. Further, we assume that $h(\boldsymbol{u})$ has bounded second-order partial derivatives with respect to $\boldsymbol{u}$ in some neighborhood $\boldsymbol{\mu}_N$, and $h(\boldsymbol{\mu}_N)$ is finite. Finally, let us define

$$\sigma_{Nn}^2 = E[(T_n^* - \theta_N)^2],\ n \geq n_0, \quad \text{where } n_0 \geq m \text{ is finite}, \tag{31}$$

and assume that there exists a sequence $\{\sigma_N^2\}$ of positive numbers such that

$$n\sigma_{Nn}^2 - [(N-n)/(N-1)]\sigma_N^2 \to 0 \text{ as } n \to \infty, \quad \underline{\lim}\sigma_N^2 > 0. \tag{32}$$

Then, it can be proved (see Majumdar and Sen, 1978) that $S_n^2 - \sigma_N^2$ strongly converges to 0 as n increases; this strong convergence is in the sense that for every $\varepsilon > 0$ and $\delta > 0$, there exists a positive integer $n_0 = n_0(\varepsilon, \delta)$ such that

$$P\left\{\max_{n_0 \leq n \leq N} |S_n^2 - \sigma_N^2| > \varepsilon\right\} < \delta, \quad n \geq n_0.$$

Further, it follows that for any (fixed) $\alpha, 0 < \alpha \leq 1$, if $n/N \to \alpha$ as $N \to \infty$ then, as $n \to \infty$,

$$n^{1/2}(T_n^* - \theta_N)/S_n \text{ is asymptotically normal } (0, 1-\alpha). \tag{33}$$

This is very useful in setting up a confidence interval for the parameter θ_N or to test a null hypothesis $H_0 : \theta_N = \theta_0$ (specified).

As a simple illustration, consider a typical *ratio-estimator* of the form

$$T_n = U_n^{(1)}/U_n^{(2)}, \quad U_n^{(j)} = n^{-1}\sum_{i=1}^{n} g_j(X_i), \quad j = 1, 2, \tag{34}$$

where the functions $g_1(\cdot)$ and $g_2(\cdot)$ may be of quite general form. In fact, we may even consider some U-statistics for $U_n^{(1)}$ and $U_n^{(2)}$. In such a setting, T_n is not generally an unbiased estimator of the population parameter $\theta_N = \mu_N^{(1)}/\mu_N^{(2)}$, though the $U_n^{(j)}$ may unbiasedly estimate the $\mu_N^{(j)}$, $j = 1, 2$. Typically, the bias of T_n is of the form in (26), and hence, jackknifing reduces the bias to the order n^{-2}. Further, here $h(a, b) = a/b$ so that

$$\frac{\partial^2}{\partial a^2}h(a,b) = 0, \quad \frac{\partial^2}{\partial a \partial b}h(a,b) = -b^{-2}, \quad \frac{\partial^2}{\partial b^2}h(a,b) = 2b^{-2}h(a,b).$$

Consequently, whenever $\mu_N^{(2)}$ is strictly positive and finite, for finite θ_N, the regularity conditions are all satisfied; for some specific cases, we may refer to Majumdar and Sen (1978) and Krewski (1978).

The basic advantage of using these results is that these asymptotics are readily adaptable for FPS sequential testing and estimation procedures. Further, the asymptotic inequalities discussed in the preceding section also remain applicable for the jackknifed estimators. In particular, (20) through (23) also hold when we replace the U-statistics and their variances by the T_k^* and the Tukey estimator of their variances. Finally, the results can be easily extended to the case where the T_n are q-vectors, for some $q \geq 1$. In that case, we would have a *tied-down Brownian sheet* approximation (in law) and strong convergence of the matrix of jackknifed variance-covariances. For the vector case, we would have an analogous result involving a multivariate normal distribution.

In the discussions so far, we have mainly confined ourselves to SRS (WR or WOR). Jackknifing is usable in other sampling schemes as well, see, for example, Krewski and Rao (1981), Rao and Wu (1985), or Shao (1996) among others for asymptotic results on jackknifing in stratified multistage sampling and comparison with asymptotic properties of variance estimators obtained by other resampling techniques like balanced repeated replication (BRR) or by linearization. An orthogonal decomposition by Hoeffding (1948) can be also used to establish asymptotic properties of jackknife variance estimators of nonlinear statistics in stratified samples (Bloznelis, 2003). The reverse martingale structure underlying the jackknifed versions may not generally hold in unequal probability sampling schemes. Recently, however, Berger and Skinner (2005) proposed a jackknife variance estimator for a smooth function of population mean which can be applicable to a general class of unequal probability sampling designs. The approach is based on an analogy between the jackknifing and linearization method. The consistency and asymptotic normality are then obtained by using asymptotic theory developed for

Horvitz–Thompson estimator and in accordance with the asymptotic theory explained here and in the next sections of this chapter. Using Hájek (1964) variance approximation, Berger (2007) proposed a new jackknife variance estimator of a smooth function of the population mean that is consistent under some regularity conditions in a general unistage stratified sampling with unequal probabilities.

The popular *bootstrap* methodology innovated by Efron (1979) has the flexibility to be used not only for variance estimation but also for the approximation of a distribution and a confidence interval building, for data with imputed values and for small area problems. Basically, it incorporates SRSWR resampling schemes and other variations (while SRSWOR is more suitable for jackknife methodology). Let $X_1, \ldots, X_n$ be observed data and θ an unknown parameter of interest. Let $\widehat{\theta}_n = T_n(X_1, \ldots, X_n)$ be an estimator of θ. The standard inference on θ is based on a properly standardized version of its sampling distribution or suitable analytical approximations if exact distribution is not properly manageable; studentization and asymptotic normality being most common. In bootstrapping, first we approximate the (usually unknown) probability distribution of the $X_1, \ldots, X_n$ by its empirical counterpart $F_n(.)$ and generate a sample $X_1^*, \ldots, X_n^*$ from $F_n(.)$ by selecting a SRSWR of size n from the original sample; this is a bootstrap sample. In this model, $\theta^* = \widehat{\theta}_n$ and $\widehat{\theta}_n^* = T_n(X_1^*, \ldots, X_n^*)$. Then the exact bootstrap distribution of $\widehat{\theta}_n^*$ and its characteristics can be developed. In practice, we generate, say, B bootstrap samples $X_1^{*b}, \ldots, X_n^{*b}$, $b = 1, \ldots, B$, compute $\widehat{\theta}_n^{*b}$ in each sample and estimate the distribution and other characteristics from values $\widehat{\theta}_n^{*b}$, $b = 1, \ldots, n$. The bootstrap is consistent if the bootstrap distribution, that is, the conditional distribution of $\widehat{\theta}_n^*$ given the original sample $X_1, \ldots, X_n$ is asymptotically the same as the asymptotic distribution of $\widehat{\theta}_n$. We refer to Shao and Tu (1995, Chapter 3), where general theory and consistency results for bootstrap are explained in more detail. This theoretical claim comes with a price tag: (i) results are asymptotic and need to be validated in small samples by extensive simulation studies and (ii) if the asymptotic distribution of θ_n is different from a Gaussian one, then there is no guarantee of this asymptotic equivalence. (For example, if the underlying distribution is a stable one of index $\alpha < 2$, the bootstrap distribution will be asymptotically normal with large but finite variance, while the actual asymptotic distribution will still be stable, but not normal and infinite variance.)

If the sample $X_1, \ldots, X_n$ is selected from a finite population of size N WR and the bootstrap samples are also selected WR, then bootstrap provides asymptotically valid results. On the other hand, if the original sample is selected WOR, we need to be more careful. If the bootstrap sample is selected WR, then the bootstrap does not provide an asymptotically valid approximation, if the SF n/N is not negligible. *Without replacement bootstrap*, Gross (1980), is the following modification of the bootstrap procedure. For $N = kn$ where k is an integer, select a sample $X_1, \ldots, X_n$ of size n from the original population WOR and replicate each element of the sample exactly k times to create the bootstrap population of size N. From this population, select a bootstrap sample $X_1^*, \ldots, X_n^*$ of size n WOR. Then the SF is the same both in the original and the bootstrap sample. The conditional expectation of the bootstrap sample mean $\overline{X}_n^*$ given $X_1, \ldots, X_n$ is $\overline{X}_n$ and the conditional variance of $\overline{X}_n^*$ is

$$s_n^{*2} = \frac{n-1}{n^2} \cdot \frac{N-n}{N-1} s_n^2, \tag{35}$$

where $\overline{X}_n$ and s_n^2 are the sample mean and variance defined in (3). If $N \neq kn$, a randomization can be used as proposed by Bickel and Freedman (1984). The same procedure can be applied to stratified sampling, independently for each stratum. Then, if $\widehat{\theta} = \sum_{h=1}^{H} w_h \widehat{X}_{n_h}$ is the stratified sample mean as in Section 2 and τ^2 its variance given by (6), if $\widehat{\theta}^* = \sum_{h=1}^{H} w_h \widehat{X}^*_{n_h}$ is the bootstrap version of $\widehat{\theta}$ and

$$\tau^{*2} = \sum_{h=1}^{H} w_h^2 \frac{n_h - 1}{n_h^2} \cdot \frac{N_h - n_h}{N_h - 1} s^2_{n_h} \tag{36}$$

the bootstrap variance of $\widehat{\theta}^*$, the (conditional) distribution of $(\widehat{\theta}^* - \widehat{\theta})/\tau^*$ consistently estimates the distribution of $(\widehat{\theta} - \theta)/\tau$ (Bickel and Freedman, 1984), see also Chao and Lo (1985). On the other hand, the bootstrap variance (36) is not an unbiased estimator of the variance (6) of $\widehat{\theta}$, and if the sample sizes n_h remain bounded and $H \to \infty$ (case of many strata of small sizes), the bias of τ^{*2} persists even if $n = n_1 + \cdots + n_H \to \infty$.

Rao and Wu (1988) considered *rescaled bootstrap* WR that yields an unbiased variance estimator of the stratified sample mean when sampling within the strata is considered WR. They established the consistency of this procedure by the second-order Edgeworth expansion for large number of strata. They extended their method to stratified sampling WOR and two-stage cluster sampling and considered also unequal probability sampling without replacement (UPSWOR).

Sitter (1992a) and Chen and Sitter (1993) considered another variant of the bootstrap in stratified sampling called *mirror-match bootstrap*. The bootstrap sample in each stratum h is generated in such a way that a subsample of size n'_h is selected WOR from the sample of size n_h with the same SF f_h as in the original sampling scheme. This is done independently k_h times, replacing the subsample each time to obtain a vector of size n_h. The value of $k_h = 1/f_h$ is supposed to be an integer; otherwise a randomization should be used as proposed by the authors. Due to independent subsamples, the method yields unbiased variance estimator of the stratified sample mean. Asymptotic validity and second-order correctness based on Edgeworth expansion are also proved in case that the number of strata $H \to \infty$ and n_h remain bounded (Sitter, 1992a) and also in case that H is bounded and $n_h \to \infty$ for each $h = 1, \ldots, H$, Chen and Sitter (1993).

Asymptotic results for bootstrap in sample surveys are further discussed in Booth et al. (1994), Shao and Tu (1995), see also Shao (1996), Chaudhary and Sen (1998), and Helmers and Wegkamp (1998). Some extension of bootstrap procedures to probability proportional size (pps) sampling can be found in Sverchkov and Pfeffermann (2004). The main developments and the practical impact of the bootstrap to survey sampling are discussed in Shao (2003) and Lahiri (2003a). See also Chapter 28 in this handbook on resampling methods in surveys that includes a discussion of asymptotics of resampling methods in the context of small area estimation.

4. Estimation of population size: Asymptotics

The estimation of the total size of a population including, in particular, mobile populations (such as the number of fish in a lake) is of great importance in a variety of biological environmental and ecological studies. Of the methods available for obtaining

information about the size of such populations, the ones based on *capture, marking, release and recapture* (CMRR) of individuals, originated by Petersen (1896), have been extensively studied and adapted in practice. The Petersen method is a two-sample experiment and amounts to marking (or tagging) a sample of a given number of individuals from a closed population of unknown size (N) and then returning it into the population. The proportion of marked individuals appearing in the second sample estimates the proportion marked in the population, providing in turn, the estimate of the population size N. Schnabel (1938) considered a multisample extension of the Petersen method, where each sample captured commencing from the second is examined for marked members and then every member of the sample is given another mark before being returned to the population. For this method, the computations are simple, successive estimates enable the field worker to see the performance of his method as sampling progresses, and the method can be adapted for a wide range of capture conditions.

For the statistical formulation of the CMRR procedure, we use the following notations. Let N be the total population size (finite and unknown), $k \geq 2$ be the number of samples, n_i, $i \geq 1$ the size of the ith sample, m_i be the number of marked individuals in the ith sample, $u_i = n_i - m_i$, $i = 1, \ldots, k$, and M_i be the number of marked individuals in the population just before the ith sample is drawn (i.e., $M_i = \sum_{j=1}^{i-1} u_j$, $i = 1, \ldots, k$. Conventionally, we let $M_1 = u_1 = 0$ and $M_{k+1} = M_k + n_k - m_k = \sum_{j=1}^{k}(n_j - m_j)$. Now, the conditional distribution of m_i, given M_i, and n_i is given by

$$L_N^{(i)}(m_i|M_i, n_i) = \binom{M_i}{m_i}\binom{N - M_i}{n_i - m_i}\Big/\binom{N}{n_i}, \quad i = 2, \ldots, k, \tag{37}$$

so that the (partial) likelihood function is

$$L_N(n_1, \ldots, n_k) = \prod_{i=2}^{k} L_N^{(i)} = \prod_{i=2}^{k}\left\{\binom{M_i}{m_i}\binom{N - M_i}{n_i - m_i}\Big/\binom{N}{n_i}\right\}. \tag{38}$$

Note that

$$L_N/L_{N-1} = N^{-(k-1)}\left\{\prod_{i=1}^{k}(N - n_i)\right\}(N - M_{k+1})$$

so that

$$L_N/L_{N-1} \lessgtr 1 \text{ according to } (1 - N^{-1}M_{k+1}) \lessgtr \prod_{j=1}^{k}(1 - N^{-1}n_j). \tag{39}$$

Now, (39) provides the solution for the *maximum likelihood estimator* (MLE) of N. For the Petersen scheme, that is, $k = 2$, (39) reduces to

$$L_N/L_{N-1} \lessgtr 1 \text{ according to } N \lessgtr n_1 n_2/m_2 \tag{40}$$

so that $[n_1 n_2/m_2] = \widehat{N}_2$ is the MLE of N. For $k \geq 3$, in general, (39) needs an iterative solution for locating MLE of N. Note that based on $L_N^{(i)}$, the MLE of N is given by $\widehat{N}_i = [n_i M_i/m_i]$ for $i = 2, \ldots, k$. It is of natural interest to study the relationship between the MLE $\widehat{N}$ from (39) and the $\widehat{N}_j$, $j = 2, \ldots, k$, when $k \geq 3$. Before doing so we may note that, by virtue of (37),

$$P(m_i = 0|M_i, n_i) = \binom{N - M_i}{n_i} \Big/ \binom{N}{n_i} > 0 \quad \text{for every } i = 2, \ldots, k,$$

so that the MLE $\widehat{N}_i$ do not have finite moments of any positive order. To eliminate this drawback, we may proceed as in Chapman (1951) and consider the modified MLE

$$\check{N}_i = (n_i + 1)(M_i + 1)/(m_i + 1) - 1 \quad \text{for } i = 2, \ldots, k. \tag{41}$$

Asymptotically (as $N \to \infty$), both $\widehat{N}_i$ and $\check{N}_i$ behave identically and hence this modification is well recommended. Using the normal approximation to the hypergeometric distribution, one readily obtains from (41) that

$$N^{-1/2}(\widehat{N}_2 - N) \text{ is asymptotically } \mathcal{N}(0, \gamma^2(\alpha_1, \alpha_2)), \tag{42}$$

whenever for some $0 < \alpha_1, \alpha_2 \leq 1$, $n_1/N \to \alpha_1$, and $n_2/N \to \alpha_2$ as $N \to \infty$, where

$$\gamma^2(a, b) = (1-a)(1-b)/ab \geq [(2-a-b)/(a+b)^2], \quad 0 < a, b \leq 1, \tag{43}$$

and where the equality sign in (43) holds when $a = b$.

For the case of $k \geq 3$, a little more delicate treatment is needed for the study of the asymptotic properties of the MLEs as well as their interrelations. Using some martingale characterizations, such asymptotic studies have been made by Sen and Sen (1981) and Sen (1982a,b). First, it follows from Sen and Sen (1981) that a very close approximation N^* to the actual MLE $\widehat{N}$ in (39) is given by the solution of

$$N^* = \left[\sum_{s=2}^{k} \widehat{N}_s m_s/(N^* - M_s)(N^* - n_s)\right] \Big/ \left[\sum_{s=2}^{k} m_s/(N^* - M_s)(N^* - n_s)\right], \tag{44}$$

where the MLE $\widehat{N}_i$ are defined as in after (40). Two other approximations, listed in Seber (1973), are given by

$$\widetilde{N} = \left[\sum_{s=2}^{k} \widehat{N}_s m_s/(\widetilde{N} - M_s)\right] \Big/ \left[\sum_{s=2}^{k} m_s/(\widetilde{N} - M_s)\right] \tag{45}$$

and

$$\widetilde{\widetilde{N}} = \left[\sum_{s=2}^{k} \widehat{N}_s m_s\right] \Big/ \left[\sum_{s=2}^{k} m_2\right]. \tag{46}$$

$\widetilde{N}$ works out well when the $n_j, 2 \leq j \leq k$, are all equal or $N^{-1}n_j$ are all small, while $\widetilde{\widetilde{N}}$ is quite suitable when in addition, the $N^{-1}M_i, i = 2, \ldots, k$ are all small. For both (44) and (45), an iterative solution works out very well, as discussed is Sen and Sen (1981). If we let

$$n_i = N\alpha_i, \quad 0 < \alpha_i \leq 1, \quad \beta_i = \prod_{j=1}^{i}(1 - \alpha_j), \quad i = 1, \ldots, k, \tag{47}$$

then we have (see Sen and Sen, 1981)

$$N^{-1/2}(N^* - N) \text{ is asymptotically } \mathcal{N}(0, \sigma^{*2}) \tag{48}$$

where

$$\sigma^{*2} = \left[\sum_{s=2}^{k} \alpha_s(1-\beta_{s-1})/\beta_s\right]^{-1}.$$

Other estimators and asymptotic results parallel to (48) are given in Sen (1988).

In many situations when the n_j are very small compared with N, the m_j are also very small (may even be equal to 0 with a positive probability). This may push up the variability of the estimators considered earlier. For this reason, often an *inverse sampling* scheme is recommended. In this setup, at the sth stage, the sample units are drawn one by one until a preassigned number m_s of the marked units appear so that the sample size n_s is a r.v. while m_s is fixed in advance for $s = 2, \ldots, k$. For this inverse sampling scheme, parallel to (37), we have

$$\begin{aligned} L_N^{(i)}(n_i|M_i, m_i) &= \binom{N}{n_i-1}^{-1}\binom{M_i}{m_i-1}\binom{N-M_i}{n_i-m_i} \\ &\quad \times \{(M_i-m_i+1)/(N-n_i+1)\} \\ &= \left\{m_i\binom{M_i}{m_i}\binom{N-M_i}{n_i-m_i}\right\} \Big/ \left\{n_i\binom{N}{n_i}\right\}, \quad i = 2, \ldots, k, \end{aligned} \tag{49}$$

and (38) can be modified accordingly. Note that (39) and (40) are not affected so that the MLE remains the same. It follows from Bailey (1951) that $\check{N}_i = (M_i+1)n_i/m_i - 1$, $i = 2, \ldots k$, are unbiased estimators of N. Note that the exact variance of $\check{N}_2$ is equal to $(n_1 - m_2 + 1)(N+1)(N-n_1)/m_2(n_1+2)$ so that on letting $m_2 = \alpha_1^* \cdot \alpha_1 N$ we have, parallel to (42) and (43), that

$$N^{-1/2}(\check{N}_2 - N) \text{ is asymptotically } \mathcal{N}(0, (1-\alpha_1)(1-\alpha_1^*)/\alpha_1\alpha_1^*). \tag{50}$$

Note that α_1^* plays the same role as α_2 in (42) and (43). The main advantage of this inverse sampling scheme is that the estimates have finite moments of positive orders, although the amount of sampling (i.e., $n_2 + \cdots + n_k$) is not predetermined (but is a r.v.). Inverse sampling schemes are the precursors of *sequential sampling tagging* considered by Chapman (1952), Goodman (1953), Darroch (1958), and others. Darling and Robbins (1967) and Samuel (1968) have studied some related problems on stopping times arising in sequential sampling tagging for the estimation of the population size N, and the asymptotic theory plays a vital role in this context. Lack of stochastic independence of the r.v. at successive stages of drawing and nonstationarity of their marginal distributions call for a nonstandard approach for rigorous study of the asymptotic properties of the MLE of N in a multistage or sequential sampling procedure. Using a suitable martingale characterization, this asymptotic theory has been developed in Sen (1982a,b, 1987). These developments have been discussed in detail in Sen (1988). Applications to rare species size estimation have also been discussed there.

5. Sampling with varying probabilities: Asymptotics

Hansen and Hurwitz (1943) initiated the use of unequal selection probabilities leading to more efficient estimators of the population total. If N and n stand for the number of units in the population and sample, respectively, and if $Y_1, \ldots, Y_N$ and $y_1, \ldots, y_n$ denote the

values of these units in the population and sample, respectively, then one may consider the following sampling WR scheme. Let $P = (P_1, \ldots, P_N)$ be positive numbers that are normalized in such a way that $P_1 + \cdots + P_N = 1$. Typically, one may consider a measure S_i of the size of the ith unit in the population and set $P_i = S_i/(\sum_{i=1}^N S_i)$ for $i = 1, \ldots, N$. Now, corresponding to the sample entries $y_1, \ldots, y_n$, the associated P_s are denoted by $p_1, \ldots, p_n$. Here, sampling is made WR and the jth unit in the population is drawn with the probability P_j for $j = 1, \ldots, N$. Then, the Hansen-Hurwitz estimator of the population total $Y = Y_1 + \cdots + Y_N$ is

$$\widehat{Y}_{\mathrm{HH}} = n^{-1}(y_1/p_1 + \cdots + y_n/p_n). \tag{51}$$

This estimator is unbiased and its sampling variance is given by

$$\begin{aligned} \mathrm{Var}(\widehat{Y}_{\mathrm{HH}}) &= n^{-1}\left\{\sum_{i=1}^N Y_i^2/p_i - Y^2\right\} \\ &= (2n)^{-1}\sum_{1\le i\ne j\le N} P_iP_j\{Y_i/P_i - Y_j/P_j\}^2. \end{aligned} \tag{52}$$

We may further note that

$$\begin{aligned} S_{n\mathrm{HH}}^2 &= [n(n-1)]^{-1}\sum_{i=1}^n (y_i/p_i - \widehat{Y}_{\mathrm{HH}})^2 \\ &= [2n^2(n-1)]^{-1}\sum_{1\le i\ne j\le n}\{y_i/p_i - y_j/p_j\}^2 \end{aligned} \tag{53}$$

is an unbiased estimator of $\mathrm{Var}(\widehat{Y}_{\mathrm{HH}})$. Since sampling is made WR and the y_i/p_i are independent with mean Y and variance $\sum_{i=1}^N Y_i^2/p_i - Y^2 (= \sigma_{N\mathrm{HH}}^2$, say), standard large sample theory is adoptable to verify that as n increases,

$$nS_{n\mathrm{HH}}^2/\sigma_{N\mathrm{HH}}^2 \to 1 \text{ in probability}, \tag{54}$$

$$n^{1/2}(\widehat{Y}_{\mathrm{HH}} - Y) \text{ is asymptotically } \mathcal{N}(0, \sigma_{N\mathrm{HH}}^2), \tag{55}$$

so that by (54) and (55),

$$n^{1/2}(\widehat{Y}_{\mathrm{HH}} - Y)/S_{n\mathrm{HH}} \text{ is asymptotically } \mathcal{N}(0, 1). \tag{56}$$

The situation becomes quite different when sampling is made WOR. On one hand, one has generally more efficient estimators; on the other hand, the exact theory becomes so complicated that one is naturally inclined to rely mostly on the asymptotics. To encompass diverse sampling plans (WOR), we identify the population with the set $\mathcal{S} = \{1, \ldots, N\}$ of natural integers and denote the sample by s. A sampling design may then be defined by the probabilities $P(s)$, $s \subset \mathcal{S}$, associated with all possible samples. In particular, we let

$$\pi_i = \sum_{s\ni i} P(s), \quad i = 1, \ldots, N, \tag{57}$$

where the sum $\sum_{s\ni i}$ extends over all s containing i. These π_i are termed the *first-order inclusion probabilities.* Similarly, the *second-order inclusion probabilities* are

defined as

$$\pi_{ij} = \sum_{s \ni i,j} P(s), \quad i \neq j = 1, \ldots, N. \tag{58}$$

The classical *Horvitz–Thompson* (1952) *estimator* of the population total Y is then expressible as

$$\widehat{Y}_{\mathrm{HT}} = \sum_{i \in s} (Y_i/\pi_i). \tag{59}$$

The variance of this unbiased estimator of Y is

$$V(\widehat{Y}_{\mathrm{HT}}) = \mathrm{Var}(\widehat{Y}_{\mathrm{HT}}) = \sum_{i=1}^{N} (\pi_i^{-1} - 1) Y_i^2 + \sum_{1 \leq i \neq j \leq N} (\pi_{ij}/\pi_i \pi_j - 1) Y_i Y_j. \tag{60}$$

When the number of units (n) in the sample s is fixed, an alternative expression for the variance in (60), due to Sen (1953) and Yates and Grundy (1953), is

$$\sum_{1 \leq i < j \leq N} (\pi_i \pi_j - \pi_{ij})(Y_i/\pi_i - Y_j/\pi_j)^2. \tag{61}$$

Then, an unbiased estimator of $V(\widehat{Y}_{\mathrm{HT}})$ is

$$\widehat{V}(\widehat{Y}_{\mathrm{HT}}) = \frac{1}{2} \sum_{i \in s} \sum_{j \in s} (\pi_i \pi_j/\pi_{ij} - 1)(Y_i/\pi_i - Y_j/\pi_j)^2. \tag{62}$$

It is clear that if the Y_i are all (exactly or closely) proportional to the corresponding π_i, then (61) is (exactly or closely) equal to 0; this point advocates the choice of the π_i as proportional to the size of the units, and on that count, pps sampling is quite a reasonable option.

Now, in the context of sampling (WOR) with varying probabilities, various sampling designs have been considered by various authors; for a survey of such procedures, see, for example, Brewer and Hanif (1983), Tillé (2006), or Chapter 2 in this handbook.

Among these methods, *rejective sampling* may be defined as in Hájek (1964) as sampling WR with drawing probabilities $\alpha_1, \ldots, \alpha_N$ at each draw, conditioned on the requirement that all drawn units are distinct. The α_i are positive numbers adding up to 1. As soon as one obtains a replication, one rejects the whole partially built up sample and starts completely new. In this scheme, the inclusion probabilities π_i can be computed, as in Hájek (1964), in terms of the α_i.

A related sampling plan, known as *Sampford sampling* (Sampford, 1967), is defined in a similar manner, where the first unit in the sample is drawn from the population with the probabilities $\alpha_i^{(1)} = n^{-1}\pi_i$, $i = 1, \ldots, N$, and in the subsequent $(n-1)$ draws, one samples the units with drawing probabilities $\alpha_i^{(*)} = \alpha \pi_i (1 - \pi_i)^{-1}$, $i = 1, \ldots, N$ WR, where α is so selected that $\sum_{i=1}^{N} \alpha_i^{(*)} = 1$. Here also, a sample is accepted only if all the selected units are distinct. For both these schemes, a rejection of the accumulating sample is made when at any intermediate stage, a repetition occurs. On the contrary to rejective sampling, where the inclusion probabilities are computed from the drawing probabilities only approximatively, in Sampford sampling, the inclusion probabilities take the values π_i exactly.

Chao (1982) proposed a general sampling procedure with unequal probabilities that gives exact values of inclusion probabilities and that can be easily implemented by statistical programs. Let $\mathcal{S}$ be a population of N units and $\mathcal{S}_k$ a subpopulation of the first k units, $k \geq n$. At step k, a sample s_k of size n is selected from the population $\mathcal{S}_k$ with prescribed first-order probabilities $\pi(k, i), i = 1, \ldots, k$. In the next step, unit $k+1$ is selected from the population $\mathcal{S}_{k+1}$ with the probability $\pi(k+1, k+1)$. If unit $k+1$ is not sampled, then $s_{k+1} = s_k$, that is, the old sample is retained. Otherwise, one unit in s_k is replaced at random by unit $k+1$. The procedure starts with s_n and maintains the fixed sample size n at each step; the final sample is $s = s_N$. For this procedure, inclusion probabilities of all orders can be easily computed. Asymptotic properties of the above mentioned procedures will be discussed in the frame of asymptotic theory of *Poisson sampling* in the next section. In a *successive sampling* plan, one draws units one by one with drawing probabilities $P_1, \ldots, P_N$, and, if a replication occurs at any draw, that particular one is rejected, and the drawing is continued in this manner until one has the prefixed number n of distinct units in the sample. Following Rosén (1972a,b, 1974), let $I_1, \ldots, I_n$ be the indices in the (random) order in which they appear in the sample of size n, and let

$$\Delta(r, n) = P(r \in s) = \pi_r, \quad r = 1, \ldots, N \tag{63}$$

be the inclusion probability of unit r in the sample of size n. For this scheme, the Horvitz–Thompson estimator in (59) has the following form:

$$\widehat{Y}_{\mathrm{HT}} = \sum_{r \in s} Y_r / \Delta(r, n) = \sum_{i=1}^{n} Y_{I_i} / \Delta(I_i, n). \tag{64}$$

We may remark that if $P_1 = \cdots = P_N = N^{-1}$, the $\Delta(r, n)$ all reduce to n/N so that (64) is given by $N(n^{-1} \sum_{r \in S} Y_r)$, which is the standard estimator under the usual equal probability sampling scheme (WOR). In the more general case where the P_i are not all equal, the $\Delta(r, n)$ can be obtained (see Rosén, 1972a,b) as follows: for every $n \geq 1$ and $r_1, \ldots, r_n$, $1 \leq r_1 \neq \cdots \neq r_n \leq N$, let

$$P(r_1, \ldots, r_n) = P_{r_1} \times \left\{ \prod_{k=2}^{n} P_{r_k} \left[1 - \sum_{j=1}^{k-1} P_{r_j} \right]^{-1} \right\}. \tag{65}$$

Then,

$$\Delta(r, n) = \sum_{j=1}^{n} \left\{ \sum_{(j)} P(r_1, \ldots, r_n) \right\}, \tag{66}$$

where the summation $\sum_{(j)}$ extends over all permutations of $(r_1 \ldots, r_n)$ over $(1, \ldots, N)$, subject to the constraint that $r_j = r$, $j = 1, \ldots, n$, for $r = 1, \ldots, N$.

The varying probability structure and the complicated forms of the expressions in (65) and (66) introduce certain complications in the study of the asymptotic distribution theory of the Horvitz–Thompson estimator (or other estimators available in the literature). Rosén (1970, 1972a,b) considered an alternative approach (through the *coupon collector's problem)* and provided some deeper results in this context. To illustrate this approach, we first consider a coupon collector problem. For $N \geq 1$, let

$$\Omega_N = \{(a_{N1}, P_{N1}), \ldots, (a_{NN}, P_{NN})\} \tag{67}$$

be a sequence of coupon collector's situations, where the a_{Nj} and P_{Nj} are real numbers, the P_{Nj} are positive and $\sum_{j=1}^{N} P_{Nj} = 1$. Consider also a (double) sequence $\{J_{Nk}, k \geq 1\}$ of (row-wise) i. i. d. r.v., where, for each $N \geq 1, k \geq 1$,

$$P\{J_{Nk} = s\} = P_{Ns} \quad \text{for } s = 1, \ldots, N. \tag{68}$$

Further, for $k \geq 1$, put

$$X_{Nk} = \begin{cases} a_{NJ_{Nk}} & \text{if } J_{Nk} \notin \{J_{N1}, \ldots, J_{Nk-1}, \} \\ 0 & \text{otherwise.} \end{cases} \tag{69}$$

The r.v. $B_{Nn} = \sum_{k=1}^{n} X_{Nk}$ is called *bonus sum after n coupons*. Further, denote

$$\nu_{Nm} = \inf\{k : \text{number of distinct } J_{N1}, \ldots, J_{Nk} = m, \} \quad m \geq 1. \tag{70}$$

Note that for each N, the ν_{Nm} are positive integer-valued r.v. Then, as Rosén (1970, 1972a) has shown, with $X^*_{Nnk} = X_{Nk}/\Delta(J_{Nk}, n)$,

$$\widehat{Y}_{\mathrm{HT}} \overset{\mathcal{D}}{=} \sum_{k=1}^{\nu_{Nn}} X^*_{Nnk} = B^*_{N\nu_{Nn}}, \text{ say,} \tag{71}$$

where $\overset{\mathcal{D}}{=}$ stands for equality in distributions and for a given n, $B^*_{N\nu_{Nn}}$ in (71) is the bonus sum after ν_{Nn} coupons in the collector's situation $\Omega^*_{Nn} = \{(a^*_{N1}, P_{N1}), \ldots, (a^*_{NN}, P_{NN})\}$, where $a^*_{Nj} = a_{NJ}/\Delta(j, N)$, $j = 1, \ldots, N$. Thus, the asymptotic normality of (randomly stopped) bonus sums (for the reduced coupon collector's situation) provides the same result for the Horvitz–Thompson estimator. A similar treatment holds for many other related estimators in successive sampling with varying probabilities (WOR).

Towards this goal, we may note as in Rosén (1972a) that under some regularity conditions on the a_{Ni} and the P_{Ni}, as N increases,

$$\Delta(s, n) = 1 - \exp\{-P_{Ns}t(n)\} + o(N^{-1/2}), \quad s = 1, \ldots, N, \tag{72}$$

where the function $t(\cdot) = \{t(x), x \geq 0\}$ is defined implicitly by

$$N - x = \sum_{k=1}^{N} \exp\{-t(x)P_{Nk}\} \quad x \geq 0, \tag{73}$$

and, therefore, depends on $P_{N1} \ldots, P_{NN}$. Given this asymptotic relation, we may write for every $s = 1, \ldots, N$,

$$a^*_{Ns} = a_{Ns}/\Delta(s, n) = (1 - \exp\{-P_{Ns}t(n)\})^{-1}a_{Ns} + o(N^{-1/2}). \tag{74}$$

It also follows from Rosén (1970) that under the same regularity conditions,

$$\nu_{Nn}/t(n) \to 1 \text{ in probability,} \tag{75}$$

whenever n/N is bounded away from 0 and 1. Consequently, if we define the bonus sum for the reduced coupon collector's situation by $B^*_{Nnk} = \sum_{i=1}^{k} X^*_{Nni}$, $k \geq 1$ then we need to verify that (i) the normalized version of $B^*_{Nnt(n)}$ is asymptotically normal

and (ii) $n^{-1/2}\max\{|B^*_{Nnk} - B^*_{Nnt(n)}| : |k/t(n) - 1| \le \delta\} \to 0$ in probability. The latter condition is known in the literature as the Anscombe (1952) "uniform continuity in probability" condition. A stronger result, which ensures both (i) and (ii), relates to the weak convergence of the partial sequence $\{(B^*_{Nnk} - EB^*_{Nnk})/\{\mathrm{var}(B^*_{Nnt(n)})\}^{1/2}; k \le t(n), \}$; it has been established by Sen (1979) through a martingale approach. For simplicity of presentation, we consider the case of the original coupon collector's situation (the same result continues to hold for the reduced situation, too). Let us denote

$$\Phi_{Nn} = \sum_{s=1}^{N} a_{Ns}[1 - \exp\{-nP_{Ns}\}], \quad n \ge 0, \tag{76}$$

$$d^2_{Nn} = \sum_{s=1}^{N} a^2_{Ns}\exp\{-nP_{Ns}\}[1 - \exp\{-nP_{Ns}\}] - n\left(\sum_{s=1}^{N} a_{Ns}P_{Ns}\exp\{-nP_{Ns}\}\right)^2, \quad n \ge 0. \tag{77}$$

Then, under the usual (Rosén) regularity conditions, it follows that $d^2_{Nn} = \mathcal{O}(n)$ whenever n/N is finite and bounded away from 0 and

$$(B_{Nn} - \Phi_{Nn})/d_{Nn} \text{ is asymptotically } \mathcal{N}(0, 1). \tag{78}$$

See also Sen (1979, 1995) for martingale approach. For the reduced coupon collector's situation Ω^*_N and for n replaced by $t(n)$, we have, parallel to (77),

$$\begin{aligned} d^{*2}_{Nt(n)} &= \sum_{s=1}^{N} a^{*2}_{ns}\exp\{-t(n)P_{Ns}\}[1 - \exp\{-t(n)P_{Ns}\}] \\ &\quad - t(n)\left[\sum_{s=1}^{N} a^*_{Ns}P_{Ns}\exp\{-t(n)P_{Ns}\}\right]^2 \\ &= \sum_{s=1}^{N} Y^2_s\exp\{-t(n)P_{Ns}\}/[1 - \exp\{-t(n)P_{Ns}\}] \\ &\quad - t(n)\left[\sum_{s=1}^{N} Y_sP_{Ns}\exp\{-t(n)P_{Ns}\}/(1 - \exp\{-t(n)P_{Ns}\})\right]^2, \end{aligned} \tag{79}$$

where we may note that the P_{Ns} are all specified numbers so that by (73), $t(n)$ is a known quantity. As a result, we obtain that as N increases and n/N is bounded away from 0 (and is finite),

$$(\widehat{Y}_{\mathrm{HT}} - Y)/d^*_{Nt(n)} \text{ is asymptotically } \mathcal{N}(0, 1). \tag{80}$$

Further, if we take the sample observations as $y_j\ (= Y_{I_{Nj}})$, $j = 1, \ldots, n$, and denote by $P_{NI_{Nj}} = p_{Nj}$, $j = 1, \ldots, n$, we may set

$$U^{(1)}_{Nn} = n^{-1}\sum_{j=1}^{n} y^2_j\exp\{-t(n)p_{Nj}\}/[1 - \exp\{-t(n)p_{Nj}\}]^2, \tag{81}$$

$$U_{Nn}^{(2)} = n^{-1} \sum_{j=1}^{n} y_j p_{Nj} \exp\{-t(n) p_{Nj}\} / [1 - \exp\{-t(n) p_{Nj}\}]^2, \tag{82}$$

$$V_{Nn} = U_{Nn}^{(1)} - t(n)[U_{Nn}^{(2)}]^2. \tag{83}$$

Then it follows that as n increases, $V_{Nn}/d_{Nt(n)}^{*2}$ converges in probability to 1 so that in (80), $d_{N(t(n)}^{*}$ may be replaced by $V_{Nn}^{1/2}$.

We may remark that if the a_{Ns} are all nonnegative, the bonus sum B_{Nn} is nondecreasing in n so that we may define

$$U_N(t) = \min\{k : B_{Nk} \geq t\} \quad \text{for every } t \geq 0. \tag{84}$$

Then, $U_N(t)$ is termed the *waiting time to obtain the bonus sum t in the coupon collector's situation* Ω_N. Note that, by definition,

$$P\{U_N(t) > x\} = P\{B_{N\lfloor x \rfloor} < t\} \quad \text{for all } x, t > 0, \tag{85}$$

where $\lfloor x \rfloor$ denotes the integer part of x. Therefore, the asymptotic distribution of the normalized version of the waiting time can readily be obtained from (78). See also Rosén (1970) and Sen (1988) for more detail.

Asymptotic results like (78) were extended to weak invariance principles for bonus sums and waiting times in coupon collector's problems by Sen (1979, 1980). Asymptotic problems in successive sampling were also studied in Chaudhary and Sen (2002). These results provide needed theoretical justification of assumption of asymptotic normality and complement resampling results on variance estimation in complex sample survey, see Sen (1995) and Chaudhary and Sen (1998, 2002).

Besides the sampling strategies considered so far, there are some others, considered elsewhere. Among these, mention should be made of one special approach proposed by Rao et al. (1962). They considered a simple procedure of UPSWOR leading to an estimator having a smaller variance than in the case of sampling WR. Moreover, their procedure provides an unbiased sample estimator of the variance that is always positive. Both single-stage and two-stage designs were considered by them. In the single-stage design, let p_t be the probability of drawing the tth unit in the first draw from the whole population for $t = 1, \ldots, N$. They suggested that the population of N units be first divided at random into n groups of sizes $N_1, \ldots, N_n$, respectively, where $N = N_1 + \cdots + N_n$. Within each group, a sample of size one is drawn with probabilities proportional to p_t (for t belonging to the set of indices in the ith group), and this is done independently for each of the n groups. Thus, if the tth unit falls in group i, the actual probability that it will be selected is p_t/π_i where π_i is the sum over all values of p_t, for which t belongs to the set of indices in the ith group. If $y_1, \ldots, y_n$ denote the sampled units from the n groups, then the estimator of the population total is $\widehat{Y}_n = \pi_1 y_1/q_1 + \cdots + \pi_n y_n/q_n$, where $q_1, \ldots, q_n$ refers to the particular values of the p_t for these chosen units. $\widehat{Y}_n$ is an unbiased estimator of the population total, and the sampling variance of $\widehat{Y}_n$ is given by

$$V(\widehat{Y}) = [N(N-1)]^{-1} \left(\sum_{i=1}^{n} N_i^2 - N \right) \left(\sum_{t=1}^{N} Y_t^2/p_t - Y^2 \right),$$

where Y_t denotes the value associated with the tth unit in the population (and $Y = Y_1 + \cdots + Y_N$). It is clear from the above that this variance is a minimum when all the N_i are equal (i.e., $N/n = R$ is a positive integer and $N_1 = \cdots = N_n = R$). Thus, in actual practice, these N_i should be taken as close to each other as possible. An estimator of $V(\widehat{Y})$ is given by

$$\widehat{V}(\widehat{Y}) = \left(N^2 - \sum_{i=1}^{n} N_i^2\right)^{-1} \left(\sum_{i=1}^{n} N_i^2 - N\right)\left(\sum_{i=1}^{n} \pi_i((y_i/q_i) - \widehat{Y})^2\right).$$

In the case of the two-stage design, the tth primary unit ($t = 1, \ldots, N$) is composed of M_t second-stage units (subunits) so that following the selection of n primary units as in the single-stage design, for the tth primary unit selected, one draws a sample of m_t subunits WORs and with equal probabilities from the M_t subunits. Again, sampling is done independently for the different groups. The estimator of the population total Y is given by

$$\sum_{i=1}^{n} \left\{ \pi_i(M_i/m_i) \left(\sum_{j=1}^{m_i} Y_{ij} \right) / q_i \right\}.$$

Parallel expressions for the sampling variance of this estimator and an estimator of this sampling variance have also been provided by Rao et al. (1962). Thus, in either design, the procedure has the advantage of exact variance formulae (and their estimators) for any population size N and sample size n. When n, the number of groups, is large and the groupings are made randomly (as has been prescribed by them), the asymptotic theory developed earlier remains applicable under quite general regularity conditions. However, this random division of the N units into n groups may introduce some uncontrolled feature, although from the efficiency point of view it leads to improved estimators. Hartley and Rao (1962) have also considered an alternative sampling scheme with UPSWOR. The N units in the population are listed in a random order and their x_i (sizes) are cumulated; a systematic selection of n elements from a "random start" is then made on the cumulation. They were able to provide an asymptotic variance formula for their estimator. Comparing the two procedures by Rao et al. (1962) and Hartley and Rao (1962), we see that the former enjoys the advantage of exact variance formula for any population size, while the latter assumes N to be large; but in terms of the sampling variance, the former may lead to an estimator with a slightly larger variance than the latter (in many situations). Ohlsson (1986) proved the asymptotic normality of the Rao–Hartley–Cochran estimator of the population total by using a martingale approach. He also introduced a class of unbiased variance estimators for the sampling procedure by Rao et al. (1962) and established general conditions for their consistency (Ohlsson, 1989a). In Ohlsson (1989b), the asymptotic normality of the population total estimator in a general setting is established, which includes a wide class of two-stage sampling procedures.

Krewski and Rao (1981) considered general stratified multistage designs relating to a sequence $\{\Pi_H\}$ of finite populations (with H strata in Π_H), by which the primary sampling units are selected WR and independent subsamples are taken within those primary sampling units selected more than once. The asymptotic normality of both linear and nonlinear statistics is studied under the assumption that $H \to \infty$, and in the same setup,

the consistency of the variance estimators obtained by using the linearization, jackknifing, and BRR methods is established. Because of the independence of the subsamples, standard asymptotic theory, discussed before, remains applicable in this context.

Francisco and Fuller (1991) considered a sequence $\{\Pi_{H_\nu}\}$ of stratified finite populations with H_ν strata, where the population in each stratum is supposed to be a random sample of a finite number of clusters selected from an infinite superpopulation with a common distribution function F. Stratified random sample of clusters selected WOR from the population Π_{H_ν} is considered, and conditions under which the sample distribution function is asymptotically normal estimator of the population distribution function are established as $\nu \to \infty$ (and H_ν is increasing with ν). Asymptotic properties of the sample quantiles were also studied. Shao (1994), under the same sampling design as Krewski and Rao (1981), considered two general types of statistics for a complex survey, namely, smooth L-statistics with weights generated by smooth functions (they include, e.g., trimmed sample means or weighted deciles and variances) and sample quantiles. He obtained asymptotic normality for the smooth L-statistics and derived that their asymptotic variances can be consistently estimated by jackknifing. He also proposed an estimator of the asymptotic variance of the sample pth quantile under weaker conditions than Francisco and Fuller (1991).

Systematic procedures (random or ordered) for sampling with varying probabilities were also considered by Madow (1949) and Hartley (1966), among others. Iachan (1983) developed an asymptotic theory of systematic sampling from a finite population of r.v. that arise from a second-order stationary process.

Besides the systematic procedures, there are other procedures due to Narain (1951), Midzuno (1952), Yates and Grundy (1953), and Sen (1953), among others. Most of these procedures work out well for small values of n (viz., for $n \leq 4$), but as n increases, these procedures become prohibitively cumbrous. We refer to Brewer and Hanif (1983) for some detailed discussions of these procedures when n is not large. However, as regards the asymptotic theory, a lot of work remains to be accomplished.

6. Large entropy and relative samplings: Asymptotic results

In this section, as before, $\mathcal{S}$ denotes a population of N units, $s \subset \mathcal{S}$ a sample and P a probability distribution defined on the set of all subsets of $\mathcal{S}$. The population units take values $Y_1, \ldots, Y_N$. The probability of inclusion of unit i into the sample is $\pi_i = \sum_{s \ni i} P(s)$ and the probability of inclusion of units i and j is denoted by $\pi_{ij} = \sum_{s \ni i,j} P(s)$, respectively. We will use the notation $\pi_i(P)$ and $\pi_{ij}(P)$ to underline the dependence on the sampling scheme P. Size of the sample is $K = \sum_{i=1}^N I_i$, where I_i denotes the indicator of inclusion of unit i into sample s, that is, a r.v. with value 1 if $s \in$ and 0 otherwise. *Poisson sampling* with parameters $0 < p_i < 1, i = 1, \ldots, N$, by which the population units are sampled independently with probabilities p_i, is defined for any $s \subset S$ by the probability

$$P_o(s) = \prod_{i \in s} p_i \prod_{i \in S - s} (1 - p_i). \tag{86}$$

The indicators of inclusion in Poisson sampling are independent r.v. satisfying $P(I_i = 1) = p_i = 1 - P(I_i = 0)$, and the inclusion probabilities satisfy the identity

$\pi_i(P_o) = p_i$ for $i = 1, \ldots, N$. The sample size K is a r.v. with $E(K) = \sum_{i=1}^N p_i$. Poisson sampling plays an important role in defining and analyzing other sampling procedures. Assume that Ω is a set of all samples of fixed size n. Then for Poisson sampling (86) with $P_o(\Omega) > 0$, the sampling design (plan)

$$R(s) = P_o(s|\Omega) = \frac{P_o(s)}{P_o(\Omega)}, \quad s \in \Omega \tag{87}$$
$$= 0 \quad \text{otherwise}$$

is called *rejective sampling* of size n or *conditional Poisson sampling* with inclusion probabilities $\pi_i(R) = \sum_{s \ni i} R(s)$. The inclusion probabilities satisfy the condition $\sum_{i=1}^N \pi_i(R) = n$. It is known (Hájek, 1959), see also Hájek (1981, Theorem 3.4), that the rejective sampling (87) maximizes the *entropy* $H(P) = -\sum_{s \in \Omega} P(s) \log P(s)$ among all sampling plans P on Ω (e.g., among all plans of fixed sample size n) with probabilities of inclusion equal to $\pi_i(R)$. For two probability designs P_1 and P_2 defined on Ω, the Kullback–Leibler *divergence* P_1 from P_2 is defined by

$$D(P_1, P_2) = \sum_{s \in \Omega} P_1(s) \log \frac{P_1(s)}{P_2(s)}. \tag{88}$$

Since $D(P_1, P_2) \geq 0$ and $D(P_1, P_2) = 0$ if and only if $P_1(s) = P_2(s)$ for all $s \in \Omega$, any sampling P on Ω will be close to the rejective sampling R if the divergence $D(P, R)$ from the rejective sampling is small.

If we define a sampling plan by conditional Poisson sampling (87), the problem arises how to evaluate the probabilities of inclusion, because the parameters $p_1, \ldots, p_N$ may not yield exact values of them. The same problem appears when the population $\mathcal{S}$ is divided into strata $\mathcal{S}_1, \ldots, \mathcal{S}_m$, and we consider conditional Poisson sampling given fixed strata sample sizes $n_h, h = 1, \ldots, m$. For rejective sampling (87) of size n, Hájek (1964) developed the following asymptotic approximation of $\pi_i(R)$ by means of $p_1, \ldots, p_N$. If we suppose that

$$d(p) = \sum_{i=1}^N p_i(1 - p_i) \to \infty,$$

then

$$\pi_i(R) = p_i \left[1 - \frac{(\bar{\bar{p}} - p_i)(1 - p_i)}{d(p)} + o(d(p)^{-1}) \right], \quad i = 1, \ldots, N, \tag{89}$$

where

$$\bar{\bar{p}} = \sum_{i=1}^N p_i^2(1 - p_i)/d(p)$$

and $o(d(p)^{-1})$ is a remainder term such that $d(p)o(d(p)^{-1}) \to 0$ uniformly for all $1 \leq i \leq N$ as $d(p) \to \infty$. Notice that $d(p) \to \infty$ implies $n \to \infty$ and $N - n \to \infty$. For inclusion probabilities $\pi_{ij}(R)$, Hájek (1964) established the approximation,

$$\pi_{ij}(R) = \pi_i(R)\pi_j(R) \left[1 - \frac{(1 - \pi_i(R))(1 - \pi_j(R))}{d(\pi)} + o(d(\pi)^{-1}) \right],$$

or equivalently,

$$\pi_i(R)\pi_j(R)-\pi_{ij}(R) = \frac{\pi_i(R)(1-\pi_i(R))\pi_j(R)(1-\pi_j(R))}{d(\pi)}[1+o(1)], \quad (90)$$

where $d(\pi)$ is defined as $d(p)$ with p_i replaced by $\pi_i(R)$. Arratia et al. (2005) established a general asymptotic expansion of the inclusion probabilities $\pi_i(R)$ and $\pi_{ij}(R)$ around p_i in terms of decreasing powers of N. For the stratified conditional Poisson sampling, an asymptotic relation between the inclusion probabilities π_i, π_{ij} and the parameters of Poisson sampling was conjectured in (Hájek, 1981, Conjecture 14.1) and justified by Prášková (1995). Recently, algorithms and recursive procedures were developed to compute the probabilities $\pi_i(R)$ and $\pi_{ij}(R)$ from p_i that enable to construct Horvitz-Thomspon estimator (59) as well as variance estimators and compute their characteristics, see, for example, Chen et al. (1994), Chen (2000), and Aires (1999, 2000). It was remarked in the previous section that rejective sampling can be alternatively defined as a conditional sampling WR given the condition that all the selected units are different. Then

$$\begin{aligned} R(s) &= c\prod_{i\in s}\alpha_i, \quad K(s)=n \\ &= 0 \quad \text{otherwise}, \end{aligned} \quad (91)$$

where $0 \le \alpha_1, \cdots, \alpha_N \le 1$, $\sum_{i=1}^N \alpha_i = 1$, are drawing probabilities and c is a constant such that $\sum_{s\subset S} R(s) = 1$. Though there is one-to-one relation between the inclusion probabilities in Poisson sampling and drawing probabilities in rejective sampling, that is,

$$\alpha_i = \frac{p_i/(1-p_i)}{\sum_{j=1}^N p_j/(1-p_j)}, \quad (92)$$

connection between α_i and $\pi_i(R)$ remains on asymptotic level only (Hájek, 1981).

In the sequel, we will consider definition (87) of rejective sampling (i.e., as conditional Poisson sampling) and review some asymptotic results. Before we do it, we shall introduce the following notation. For any sample design P defined on Ω with inclusion probabilities $\pi_i(P)$ denote

$$d(P) = \sum_{i=1}^N \pi_i(P)(1-\pi_i(P)), \quad (93)$$

$$G(P) = \sum_{i=1}^N Y_i(1-\pi_i(P))/d(P), \quad (94)$$

$$\sigma^2(P) = \sum_{i=1}^N [Y_i - G(P)\pi_i(P)]^2\left(\frac{1}{\pi_i(P)}-1\right), \quad (95)$$

$$A_\epsilon = \{i : |Y_i - G(P)\pi_i(P)| > \epsilon\pi_i(P)\sigma(P)\}, \quad (96)$$

$$L(\epsilon) = \sum_{i\in A_\epsilon}[Y_i - G(P)\pi_i(P)]^2\left(\frac{1}{\pi_i(P)}-1\right)/\sigma^2(P), \quad (97)$$

$$\mathrm{e} = \inf\{\epsilon : L(\epsilon) \le \epsilon\}. \quad (98)$$

Notice that $\sigma^2(P)$ can be written as

$$\sigma^2(P) = \sum_{i=1}^{N} \left[\frac{Y_i}{\pi_i(P)} - G(P) \right]^2 \pi_i(P)(1 - \pi_i(P)) \tag{99}$$

$$= \sum_{i=1}^{N} Y_i^2 \left(\frac{1}{\pi_i(P)} - 1 \right) - d(P)G^2(P). \tag{100}$$

Now, let us consider the Horvitz–Thompson estimator (59) and its variance (60) under rejective sampling R. Utilizing relation (90) and inserting into (60), we get

$$V_R(\widehat{Y}_{\mathrm{HT}}) = \sum_{i=1}^{N} (Y_i - G(R)\pi_i(R))^2 \left(\frac{1}{\pi_i(R)} - 1 \right) [1 + o(1)] \tag{101}$$

$$= \sigma^2(R)[1 + o(1)],$$

where $o(1) \to 0$ as $d(R) = \sum_{i=1}^{N} \pi_i(R)(1 - \pi_i(R)) \to \infty$. Formula (101) enables us to approximate the variance (60) in terms that do not depend on inclusion probabilities $\pi_{ij}(R)$. The result belongs to Hájek (1964). Hájek (1964) also established the asymptotic normality of a linear estimator of the population total in the form $\widehat{Y}_H = \frac{X}{n} \sum_{i \in s} \frac{Y_i}{x_i}$ with $X = \sum_{i=1}^{N} x_i$, where x_i are some constants. Such estimator can be considered an approximation of the Horvitz–Thompson estimator (59) with inclusion probabilities proportionate to nx_i/X.

Víšek (1979), using the theory of characteristic functions and the asymptotic relations (89) between the inclusion probabilities of Poisson and rejective sampling, developed the asymptotic normality of the Horvitz–Thompson estimator in rejective sampling under some conditions on the probabilities p_i of the corresponding Poisson sampling and under some centering conditions. Prášková (1984), using the same setup, obtained the rate of convergence of the Horvitz–Thompson estimator to $\mathcal{N}(0, 1)$ with rate $O(N^{-1/2})$ as $N \to \infty$, $n \to \infty$ and with n/N bounded away from 0 and 1. Minimizing divergence from the rejective sampling and utilizing Hájek results, Berger (1998a) established the asymptotic normality of the Horvitz–Thompson estimator (59) for any sampling design P of fixed size. More exactly, let $\widehat{Y}_{\mathrm{HT}}$ be the Horvitz–Thompson estimator (59) of the population total Y under sampling scheme P, let $D(P, R)$ be the divergence of P from the rejective sampling R defined by (88) and

$$T(P) = \frac{\widehat{Y}_{\mathrm{HT}} - Y}{\sigma(P)}. \tag{102}$$

Then, in the notation (93)–(98), as $\mathrm{e} \to 0$, $d(P) \to \infty$, and $D(P, R) \to 0$, $T(P)$ is asymptotically $\mathcal{N}(0, 1)$. Under the assumption that

$$\sum_{i=1}^{N} \left(\frac{Y_i}{\pi_i(P)} \right)^4 \le b_1 N, \quad \sigma^2(P) \ge b_2 N$$

for some positive constants b_1, b_2 (this implies that n/N remains bounded away from 0 and 1), Berger (1998a) also established the rate of convergence to the normal distribution given by the inequality

$$\mathrm{P}(T(P) \le x) - \Phi(x)| \le \frac{k}{N} + 2\sqrt{D(P, R)}, \tag{103}$$

where Φ denotes the distribution function of the standard normal distribution and k is a positive constant. In his paper, he checked that the above conditions are satisfied for Sampford and successive sampling and completed and generalized results for the Horvitz–Thompson estimator obtained by Prášková (1982, 1984), Víšek (1979), and Rosén (1972a,b). Berger (2005a) also proved that the divergence from the rejective sampling tends to zero for Chao's (1982) sampling procedure, which implies the asymptotic normality of the Horvitz–Thompson estimator for this sampling as well.

We have seen in (101) that the Hájek-type approximation (95) (resp. (100)) can be considered as the asymptotic variance of the Horvitz–Thompson estimator for any sampling design P of fixed sample size, close to rejective sampling in the sense of decreasing Kullback–Leibler divergence. Berger (1998b) considered a slight modification of $\sigma^2(P)$,

$$\sigma_0^2(P) = \frac{N}{N-1}\left[\sum_{i=1}^{N} Y_i^2\left(\frac{1}{\pi_i(P)} - 1\right) - d(P)G^2(P)\right] \tag{104}$$

and an approximation of (62), obtained by replacing each total sum in $\sigma_0^2(P)$ by the corresponding Horvitz–Thompson estimator,

$$\begin{aligned}\widehat{\sigma}^2(P) &= \frac{n}{n-1}\left[\sum_{i\in s}\left(\frac{Y_i}{\pi_i(P)}\right)^2(1-\pi_i(P)) - \widehat{d}(P)\widehat{G}^2(P)\right] \\ &= \frac{n}{n-1}\sum_{i\in s}\left(\frac{Y_i}{\pi_i(P)} - \widehat{G}(P)\right)^2(1-\pi_i(P)),\end{aligned} \tag{105}$$

where

$$\widehat{G}(P) = \frac{1}{\widehat{d}(P)}\sum_{i\in s}\frac{Y_i}{\pi_i(P)}(1-\pi_i(P)), \tag{106}$$

$$\widehat{d}(P) = \sum_{i\in s}(1-\pi_i(P)). \tag{107}$$

It can be easily checked that $\sigma_0^2(P)$ and $\widehat{\sigma}^2(P)$ given in (104) and (105), respectively, coincide with the variance (61) of the Horvitz–Thompson estimator of the population total and its unbiased estimator (62) when SRSWOR is used. Berger (1998b) proved that $\sigma^2(P)$ (resp. $\sigma_0^2(P)$) well approximate the variance $V_P(\widehat{Y}_{\mathrm{HT}})$ for any sampling design P with decreasing divergence from the rejective sampling. Moreover, for the Yates-Grundy estimator $\widehat{V}_P(\widehat{Y}_{\mathrm{HT}})$, see (62), and approximation $\widehat{\sigma}^2(P)$,

$$\frac{\widehat{V}_P(\widehat{Y}_{\mathrm{HT}})}{\widehat{\sigma}^2(P)} \to 1 \text{ as } D(P, R) \to 0$$

with a rate which does not depend on the sample, $D(P, R)$ is divergence from the rejective sampling R defined by (88). The asymptotic results show that the Hájek variance approximation is valid for Sampford and successive (Berger, 1998b) as well as for the Chao sampling procedures (Berger, 2005a).

We close this part concerning asymptotics in rejective sampling with some remarks on convergence to Poisson distribution. Considering a sequence of sampling designs (87), the νth of which refers to a population $Y_{\nu_1}, \ldots, Y_{\nu_{N_\nu}}$ of size N_ν and sample of size n_ν, Prášková (1988) proved the convergence of the sample sum $\sum_{i\in s} Y_{\nu_i}$ to the Poisson distribution both in local and global sense, using a method of characteristic function. The result was generalized by Rao et al. (1991) to conditional Poisson sampling from populations of nonnegative r.v. Milbrodt (1987), considering the Hájek (1964) method of correction of Poisson sampling, established a convergence of the Horvitz–Thompson estimator (59) in rejective sampling to any infinitely divisible law under conditions analogous to those formulated by Hájek (1960) for SRSWOR. More specifically, under the assumption $d(P_\nu) \to \infty$, as $\nu \to \infty$, where $d(P_\nu)$ relates to Poisson sampling with probabilities $p_{\nu 1}, \ldots, p_{\nu N_\nu}$, and the assumptions

$$\lim_{\nu\to\infty} \sum_{j=1}^{N_\nu} Y_{\nu_j} = \lambda, \quad \lim_{\nu\to\infty} \sum_{j=1}^{N_\nu} \frac{Y_{\nu_j}^2}{p_{\nu_j}} = \lambda, \tag{108}$$

$$\lim_{\nu\to\infty} \max_{1\le j\le N_\nu} p_{j_\nu} = 0, \quad \lim_{\nu\to\infty} \max_{1\le j\le N_\nu} \frac{Y_{i_\nu}^2}{p_{i_\nu}} = 0, \tag{109}$$

$$\lim_{\nu\to\infty} \sum_{|Y_{i_\nu}/p_{i_\nu}-1|>\epsilon} \frac{Y_{i_\nu}^2}{p_{i_\nu}} = 0, \quad \epsilon > 0, \tag{110}$$

the limiting law of the corresponding Horvitz–Thompson estimators is the Poisson distribution with parameter λ. The above assumptions imply that $N_\nu \to \infty, n_\nu \to \infty$ but $n_\nu/N_\nu \to 0$.

Rosén (1997a) introduced a new class of sampling schemes with varying inclusion probabilities called *order sampling*. Order sampling of size n is defined as follows: to each unit $i \in S$ there is associated a distribution function F_i on $[(0, \infty)$ with density f_i. Let $Q_1, \ldots, Q_N$ be independent r.v. with distributions $F_1, \ldots, F_N$. The unit i is included in the sample if the realized value Q_i is among the n smallest realized values of $Q_1, \ldots, Q_N$. The sampling scheme with F_i being the distribution function of the uniform distribution on $[0, \theta_i^{-1}], \theta_i > 0, i = 1, \ldots, N$, is called *sequential Poisson sampling* (Ohlsson, 1995) or *uniform sampling*. Successive sampling of size n of units $i_1 < i_2 \cdots < i_n$ can be obtained by order sampling scheme with exponential distributions, that is, for $F_i(t) = 1-\exp(-t\theta_i), \theta_i > 0, i = 1, \ldots, N$. Successive sampling is therefore called *exponential sampling,* too. Rosén (1997a) established conditions for asymptotic normality of the sample sum $\sum_{i\in s} Y_i$ for general order sampling as follows.

Let us consider stochastic processes $J_i(s) = \chi_{[Q_i\le s]}, 0 \le s < \infty, i = 1, \ldots, N$, where $\chi_{[A]}$ denotes the indicator of set A. Then we have $EJ_i(s) = F_i(s)$, $\mathrm{Var}J_i(s) = F_i(s)(1 - F_i(s))$, $0 \le s < \infty$. Let $\xi > 0$ be an arbitrary fixed real number,

$$N(t) = \sum_{i=1}^{N} J_i(t\xi), \quad L(t) = \sum_{i=1}^{N} Y_i J_i(t\xi), \quad 0 \le t < \infty,$$

and $\tau_n = \inf\{t: J(t) = n\}$ be the hitting time to level n. Then the distribution of the sample sum $\sum_{i\in s} Y_i$ is the same as that of $L(\tau_n)$ and asymptotically normal with parameters μ

and σ^2 as $n \to \infty$, $N - n \to \infty$, where

$$\mu = \sum_{i=1}^{N} Y_i F_i(\xi), \tag{111}$$

$$\sigma^2 = \sum_{i=1}^{N} (Y_i - \phi)^2 F_i(\xi)(1 - F_i(\xi)), \tag{112}$$

$$\phi = \sum_{i=1}^{N} Y_i f_i(\xi) \Big/ \sum_{i=1}^{N} f_i(\xi), \tag{113}$$

and ξ solves the equation $\sum_{i=1}^{N} F_i(t) = n$.

Since inclusion probabilities in order sampling are difficult to be computed, Rosén (1997a) introduced the following estimator of the population total $Y = \sum_{i=1}^{N} Y_i$:

$$\widehat{Y}_{\mathrm{OS}} = \sum_{i \in s} \frac{Y_i}{F_i(\xi)}. \tag{114}$$

This estimator is consistent and asymptotically normal with the mean Y and the variance

$$\sum_{i=1}^{N} \left(\frac{Y_i}{F_i(\xi)} - \gamma \right)^2 F_i(\xi)(1 - F_i(\xi)), \tag{115}$$

where

$$\gamma = \sum_{i=1}^{N} \frac{Y_i}{F_i(\xi)} f_i(\xi) \Big/ \sum_{i=1}^{N} f_i(\xi). \tag{116}$$

If $F_i = F$ for all $i = 1, \ldots, N$, then $F(\xi) = \frac{n}{N}$ and $\gamma = \frac{1}{n} Y$ (and coincides with $G(P)$, see (94)) if SRSWOR is used. Similarly, a Hájek-type modification of the asymptotic variance is

$$\sigma^2(F) = \frac{N}{N-1} \sum_{i=1}^{N} \left(\frac{Y_i}{F_i(\xi)} - \gamma \right)^2 F_i(\xi)(1 - F_i(\xi)). \tag{117}$$

Clearly, $N/(N-1) \to 1$ as $N \to \infty$, and with this correction term, the formula (117) in case $F_i = F$ for all i yields the asymptotic variance of the Horvitz–Thompson estimator in SRSWOR. Replacing each total sum in (117) by the estimator of type (114), we get the following estimator of $\sigma^2(F)$,

$$\widehat{\sigma}^2(F) = \frac{n}{n-1} \sum_{i \in s} \left(\frac{Y_i}{F_i(\xi)} - \widehat{\gamma} \right)^2 (1 - F_i(\xi)), \tag{118}$$

where

$$\widehat{\gamma} = \sum_{i \in s} \frac{Y_i}{F_i(\xi)^2} f_i(\xi) \Big/ \sum_{i \in s} \frac{f_i(\xi)}{F_i(\xi)}. \tag{119}$$

Consistency of this variance estimator was justified in Rosén (1997a). From the asymptotic unbiasness and asymptotic normality, and the fact that the Horvitz–Thompson

estimator is the unique unbiased linear estimator of the population total, Rosén (1997a) formulated a conjecture that inclusion probabilities $\pi_i \approx F_i(\xi),\ i = 1, \ldots, N$.

In Rosén (1997b), the order sampling is studied with fixed order distribution shape H and intensities $\theta_1 \ldots, \theta_N$, that is, it is assumed that $F_i(t) = H(\theta_i t)$ with $\theta_i > 0,\ i = 1, \ldots, N$, where H is a distribution function with a density h such that $h(t) \geq 0,\ t \geq 0$. For given $\lambda_1, \ldots, \lambda_N$, $0 \leq \lambda_i \leq 1$, $\sum_{i=1}^N \lambda_i = n$ (called *target inclusion probabilities*), the order sampling with shape distribution H and intensities $\theta_i = H^{-1}(\lambda_i)$ is considered. Then, $F_i(t) = H(tH^{-1}(\lambda_i))$ and the relation $\sum_{i=1}^N F_i(\xi) = n$ is solved with $\xi = 1$, which yields $F_i(\xi) = \lambda_i$. With this $F_i(\xi)$, Rosén (1997b) proved that the optimal shape distribution that minimizes the asymptotic variance (117) is the distribution,

$$H(t) = \frac{t}{1+t}, \quad 0 \leq t < \infty \text{ with density } h(t) = \frac{1}{(1+t)^2}. \tag{120}$$

The distribution (120) is the standard Pareto distribution and the corresponding order sampling scheme is called *Pareto sampling.*

With given target inclusion probabilities $\lambda_1, \ldots, \lambda_N$, the expressions (114), (116), and (117) applied to Pareto sampling take the form

$$\widehat{Y}_{\mathrm{R}} = \sum_{i \in s} \frac{Y_i}{\lambda_i}, \tag{121}$$

$$\gamma = \sum_{i=1}^N Y_i(1-\lambda_i) \Big/ \sum_{i=1}^N \lambda_i(1-\lambda_i), \tag{122}$$

$$\begin{aligned} \sigma^2(F) &= \frac{N}{N-1} \sum_{i=1}^N \left(\frac{Y_i}{\lambda_i} - \gamma \right)^2 \lambda_i(1-\lambda_i) \\ &= \frac{N}{N-1} \left[\sum_{i=1}^N Y_i^2(\lambda_i^{-1} - 1) - \gamma^2 \sum_{i=1}^N \lambda_i(1-\lambda_i) \right]. \end{aligned} \tag{123}$$

Notice that the expressions (122) and (123) coincide with (94) and (104), respectively, with inclusion probabilities $\pi_i(P)$ replaced by the target inclusion probabilities λ_i. In the same way,

$$\widehat{\gamma} = \sum_{i \in s} \frac{Y_i}{\lambda_i}(1-\lambda_i) \Big/ \sum_{i \in s}(1-\lambda_i) \tag{124}$$

coincides with $\widehat{G}(P)$ in (106) and the variance estimator

$$\widehat{\sigma}^2(F) = \frac{n}{n-1} \sum_{i \in s} \left(\frac{Y_i}{\lambda_i} - \widehat{\gamma} \right)^2 (1-\lambda_i) \tag{125}$$

coincides with (105). Then, conditions for asymptotic normality of the estimator (114) of the population total in Pareto sampling can be reformulated and simplified as follows from Rosén (1997b): for $k = 1, 2 \ldots$, consider Pareto sampling of size n_k with target inclusion probabilities $\lambda_{k1}, \ldots, \lambda_{kN_k}$ from a population of size N_k on which the variable Y_k takes the values $(Y_{k1}, \ldots, Y_{kN_k})$. Let $\widehat{Y}_k$ be the estimator of the population total Y_k in accordance with (121), let σ_k^2 and γ_k be in accordance with (123) and (122), respectively.

Then $(\widehat{Y}_k - Y_k)/\sigma_k$ is asymptotically standard normal, as $n_k \to \infty$ for $k \to \infty$, and

$$\limsup_{k\to\infty} \frac{1}{n_k} \sum_{i=1}^{N_k} \lambda_{ki}^2 < 1,$$

$$\max_i \left| \frac{Y_{ki}}{\lambda_{ki}} - \gamma_k \right| \to 0 \text{ as } k \to \infty.$$

Rosén (2000) also studied the relation between the inclusion probabilities and the target inclusion probabilities and showed that under general conditions, for Pareto, uniform and exponential sampling, $\pi_i/\lambda_i \to 1$ as the size n of the sample increases. Some numerical algorithms to find the exact inclusion probabilities in Pareto and other sampling schemes and comparative simulations studies are given in Aires (1999, 2000), Aires and Rosén (2005), and Ng and Donadio (2006).

As we have seen in the last two sections, the general asymptotics for UPSWOR sampling schemes under the two approaches initiated by Hartley and Rao (1962) and Hájek (1964) have gone through various extensions and generalizations. Basically, the two approaches rest on somewhat different regularity assumptions and they also differ in their methodological treatise. Chaudhary and Sen (2002) appraised these two apparently different approaches with a view to eliminating some apparent anomalies, too. They observed that the basic assumption for validating asymptotic normality in the approach of Hartley and Rao (1962) along the probabilistic analysis of Hájek (1964) may be reconciled under an extra condition which amounts to allowing only small variation in the varying probabilities – a condition not totally needed for the Hájek (1964) approach. Chaudhary and Sen (2002) also considered the estimation of the asymptotic variance of linear estimators in UPSWOR with further justification for resampling methods for this variance estimation problem.

7. Successive subsampling with varying probabilities: Asymptotics

Subsampling or multistage sampling is often adopted in practice and has a great variety of applications in survey sampling. Typically, we may consider a finite population of N units with variate values $a_{N1}, \ldots, a_{NN}$. Consider a successive sampling scheme where items are sampled one after the other (WOR) in such a way that at each draw, the probability of drawing item s is proportional to a number P_{Ns} if item s has not already appeared in the earlier draws, for $s = 1, \ldots, N$, where $P_{N1} \ldots, P_{NN}$ are positive numbers, adding up to 1. We like to consider a multistage extension of this sampling scheme. Here, each of the N items in the population (called the *primary units*) is composed of a number of smaller units (*subunits*), and it may be more economic to select first a sample of n primary units and then to use subsamples of subunits in each of these selected primary units. Suppose that the sth primary unit has M_s subunits with variate values b_{sj}, $j = 1, \ldots, M_s$ so that $a_{Ns} = b_{s1} + \cdots + b_{sM}$, for $s = 1, \ldots, N$. For each s, we conceive a set $\{P_{sj}^0, 1 \le j \le M_s\}$ of positive numbers (such that $\sum_{j=1}^{M_s} P_{sj}^0 = 1$) and consider a successive sampling scheme (WOR), where m_s (out of M_s) subunits are chosen. Then, as in (64), an estimator of a_{Ns} can be framed for each of the n selected primary units. Finally, these estimates can be combined as in (64) to yield the estimator of the total $A_N = a_{N1} + \cdots + a_{NN}$. The procedure can be extended to the multistage

case in a similar way. This scheme may be termed SSVPWOR. To study the asymptotic theory, first we may note that a Horvitz–Thompson estimator of a_{Ns} is

$$\widehat{a}_{Ns} = \sum_{j=1}^{M_s} \omega^*_{sj} b_{sj} / \Delta^*_s(j, m_s), \tag{126}$$

where the b_{sj} are defined as before, ω^*_{sj} is equal to 1 or 0 according to whether the jth subunit in the sth primary unit belongs to the subsample of size m_s or not, $j = 1, \ldots, M_s$, and $\Delta^*_s(j, m_s)$ is the probability that the jth subunit belongs to the subsample of m_s subunits from the sth primary unit, $1 \le j \le M_s$, $s = 1, \ldots, N$. Combining (64) and (126), we may consider the natural estimator

$$\begin{aligned} \widehat{A}_{N(\mathrm{HT})} &= \sum_{s=1}^{N} \omega_{Ns} \widehat{a}_{Ns} / \Delta(s, n) \\ &= \sum_{s=1}^{N} \sum_{j=1}^{M_s} \omega_{Ns} \omega^*_{sj} b_{sj} / [\Delta(s, n) \Delta^*(j, m_s)], \end{aligned} \tag{127}$$

where ω_{Ns} is equal to 1 or 0 according to whether the sth primary unit is in the sample of n primary units from the population $s = 1, \ldots, N$, and the inclusion probabilities $\Delta(s, n)$ are defined as in (63). Note that for each selected primary unit s, for the estimator $\widehat{a}_{Ns}$ in (126), one may use the theory discussed in Section 5. This, however, leads to a multitude of stopping numbers and thereby introduces complications in a direct extension of the Rosén approach to SSVPWOR. A more simple approach based on some martingale constructions has been worked out in Sen (1980), and we present the basic asymptotic theory as follows.

Our primary interest is to present the asymptotic theory of the estimator $\widehat{A}_{N(\mathrm{HT})}$ in (127). In this context, as in earlier sections, we allow N to increase. As $N \to \infty$, we assume that n, the primary sample size, also increases, in such a way that n/N is bounded away from 0 and 1, while the m_s (i.e., the subsample sizes) for the selected primary units may or may not be large. For this situation, the asymptotic theory rests heavily on the structure of the primary unit sampling but allow the sampling scheme for the subunits to be rather arbitrary (not necessarily a SSVPWOR), while we assume that the primary units are sampled in accordance with a SSVPWOR scheme. A second situation may arise where the number of primary units (i.e., N) is fixed or divided into a fixed number of strata, and within each stratum a sample of secondary units is drawn according to SSVWPOR scheme. This situation, however, is congruent to the stratified sampling scheme under SSVPWOR, for which the theory in Section 5 extends readily. Hence, we shall not enter into detailed discussions of this second scheme. With the notations introduced before, we now set

$$a^0_{Ns} = E(\widehat{a}_{Ns}) \text{ and } \sigma^2_{Ns} = \mathrm{Var}(\widehat{a}_{Ns}) \quad \text{for } s = 1, \ldots, N; \tag{128}$$

$$A^0_N = \sum_{s=1}^{N} a^0_{Ns} = E(\widehat{A}_{N(\mathrm{HT})}). \tag{129}$$

In order that $A^0_N = A_N$, it is therefore preferred to have unbiased estimators at the subunit stage so that $a^0_{Ns} = a_{Ns}$ for every s. Otherwise, the bias may not be negligible.

Also, for every N, we consider a nondecreasing function $t_N = \{t_N(x) : 0 \le x \le N\}$ by letting

$$N - x = \sum_{s=1}^{N} \exp\{-P_{Ns} t_N(x)\}, \quad x \in (0, N). \tag{130}$$

Let, then,

$$\begin{aligned}\delta_{Nn}^2 = &\sum_{s=1}^{N} [a_{Ns}^0]^2 \exp\{-P_{Ns} t_N(n)\}[1 - \exp\{-P_{Ns} t_n(n)\}]^{-1} \\ &+ \sum_{s=1}^{N} \sigma_{Ns}^2 [1 - \exp\{-P_{Ns} t_N(n)\}]^{-1} \\ &- t_N(n) \left[\sum_{s=1}^{N} a_{Ns}^0 P_{Ns} \exp\{-P_{Ns} t_N(n)\}/(1 - \exp\{-P_{Ns} t_N(n)\}) \right]^2. \end{aligned} \tag{131}$$

Finally, we assume that the subunit estimators $\widehat{a}_{Ns}$ satisfy a Lindeberg-type condition, namely, that for every $\eta > 0$,

$$\max_{1 \le s \le N} E[(\widehat{a}_{Ns} - a_{Ns}^0)^2 I(|\widehat{a}_{Ns} - a_{Ns}^0| > \eta N^{1/2})] \to 0 \quad \text{as } N \to \infty. \tag{132}$$

The other regularity conditions are, of course, the compatibility of the probabilities $P_{N1}, \ldots, P_{NN}$ and the sizes $a_{N1}, \ldots, a_{NN}$ (in the sense that for each sequence, the ratio of the maximum to the minimum entry is asymptotically finite). Then, we have the following result:

$$(\widehat{A}_{N(\mathrm{HT})} - A_N^0)/\delta_{Nn} \text{ is asymptotically } \mathcal{N}(0, 1). \tag{133}$$

Actually, we may consider a stochastic process $\xi_N = \{\xi_N(t);\ c < t < 1\}$ (where $c > 0$), by letting $\xi_N(t) = N^{-1/2}(\widehat{A}_{N(\mathrm{HT})}^{(t)} - A_N^0)$, where $\widehat{A}_{N(\mathrm{HT})}^{(t)}$ is the estimator in (127) based on the sample size $n = [Nt]$ (for the primary sample), $t \in [c, 1]$. Then, the process ξ_N converges in law to a Gaussian function on $[c, 1]$. The proofs of these results are based on some asymptotic theory for an *extended coupon collector's problem,* where in (67) through (70), the real (nonstochastic) elements a_{Ns} are replaced by suitable random variables X_{Ns}, $s = 1, \ldots, N$. For details of these developments, we may refer to Sen (1980).

Note that in the above development, apart from the uniform integrability condition in (132), we have not imposed any restriction on the estimates $\widehat{a}_{Ns}$. Thus we are allowed to make the subsample sizes m_s arbitrary, subject to the condition that (132) holds. In this context, we may note that if the m_s are also large, then the σ_{Ns}^2 defined by (128) will be small, so that in (131) the second sum on the right-hand side will be of smaller order of magnitude (compared with the first sum), and hence, in (133), δ_{Nn} may be replaced by d_{Nn}^* defined by (79), where the Y_s are to be replaced by a_{Ns}^0. In this limiting case, we therefore observe that subsampling does not lead to any significant increase of the variance compared with SSVPWOR of the primary units and complete enumeration in the second stage. Note, however, that in many practical problems, subsampling is more suitable, because it does not presuppose the knowledge of the values of the primary

units $\{a_{Ns}\}$, and a complete census within the selected primary units may be much more expensive than use of the estimates $\{\widehat{a}_{Ns}\}$, based on a handful of subunits.

So far, we have considered sampling WOR. In SSVP sampling WR, the theory of sampling with varying probabilities and WR, discussed in the beginning of Section 5, readily extends. In (51), instead of the primary units y_j, we need to use their estimates $\widehat{y}_j$ derived from the respective subsamples. As in (131), this will result in an increased variability due to the individual variances of the second-stage estimators. However, when sampling WR, this strategy yields simplifications in the treatment of the relevant asymptotic theory, and (55) and (56) both extend to this subsampling scheme without any difficulty.

8. Conclusions

In this chapter, we have reviewed some recent results in finite population asymptotics. We may remark that in FPS, the usual treatment for the asymptotic theory (valid for independent r.v.) may not be directly applicable. But in most of these situations, by appeal either to some appropriate permutation structures (for equal probability sampling) or to some martingale theory (for sampling with varying probabilities as well), the asymptotic theory has been established under quite general regularity conditions. These provide theoretical justifications of the asymptotic normality of different estimators (under diverse sampling schemes) when the sample size(s) may or may not be nonstochastic. Many details of asymptotic results in FPS based on martingale theory and permutation principles can be found in the article by Sen (1988).

We conclude this chapter with some brief remarks on topics that we have not considered so far. Some of them are discussed in other chapters of this handbook like *two-phase sampling* that is described in detail in Chapter 3. We mention the paper by Chen and Rao (2007) only, where asymptotic results for a class of estimators under various two-sample designs are developed. We also did not discuss in detail higher order asymptotic results based on Edgeworth expansions and saddle point approximations in FPS. We may refer to the book by Thompson (1997) for an explanation of relevant asymptotic theory and some basic references. What concerns (generalized) occupancy problems, in view of the asymptotic normality results referred to in the text, both jackknifing and bootstrap methods may be worked out conveniently. However, these require more extensive methodological investigations, and we are to relegate such developments for future research.

Small area estimation, particularly arising in *spatial sampling* problems, refers to a large number of small subpopulations and small samples from them. The situation is different from stratified sampling where the number of strata is generally small to moderate while the sample sizes for the strata are generally not small. Since a direct survey estimator based only on the data from the small area yields large standard errors, a need exists to develop new procedures with the aim to increase the precision of small area estimates. These statistical methods are reviewed in Chapter 32 of this handbook. Asymptotic results obtained there concern the improvement of the estimation of the mean square errors and covering accuracy of prediction intervals when the number of areas increases to infinity while the sample sizes remain fixed. We also refer to the book by

Rao (2003a) on small area estimations that provides a detailed account of all the results developed until then, and to Chapter 28 of this handbook, where resampling methods including for small area estimation are discussed. Second-order accurate nonnegative estimators of mean squared prediction errors in small area estimation are developed in Lahiri et al. (2007). Empirical and hierarchical Bayes methods have been found to be very useful in small area estimation (for some asymptotic results see, e.g., Butar and Lahiri, 2003), while for large samples there are general equivalence results for Bayes and likelihood procedures. As such, we refer to Chapter 29.

Sampling procedures from populations, units of which are labeled by times or positions in space, are subject of intensive research in the last two decades. Asymptotic theory for time dependent data, however, is beyond the scope and extent of this chapter and we do not present it here.

Acknowledgment

The authors thank the Editor and the referees for very constructive comment. The research of the first author was supported by grants MSM 0021620839 and GAČR 201/06/0186.

Sample Surveys: Inference and Analysis, Vol. 29B
ISSN: 0169-7161
© 2009 Elsevier B.V. All rights reserved
DOI: 10.1016/S0169-7161(09)00241-7

Some Decision-Theoretic Aspects of Finite Population Sampling

*Yosef Rinott**

1. Introduction

Decision theory provides tools and insights for understanding, comparing, and selecting sampling and estimation procedures. In this chapter, we present a small sample of the extensive literature on decision-theoretic aspects of sampling from finite populations, without attempting to give a comprehensive survey of the best possible results and references.[1] Technical details are sometimes omitted for the sake of simplicity.

The chapter is quite theoretical, dealing with the foundations of *finite population sampling* and inference through simple designs and models rather than the complex ones in modern use. It is hoped that a practitioner may find these basic ideas of interest, albeit theoretical. However, it seems that a student or teacher of statistical decision theory can definitely benefit from the wealth of ideas that exist in the area of finite population sampling. It provides setups and examples that add an interesting perspective to the standard illustrations given in most statistical decision theory courses, where a *sample* is often restricted to mean i.i.d. observations.

The task of estimating the mean, say, of a given finite population of size N by measuring $n < N$ units does not seem to involve any probability structure, unlike other statistical setups where it is assumed at the outset that the data consist of random or noisy observations. By random sampling, statisticians introduce noise or randomness that did not exist in the original problem. It is well known that the introduction of random sampling can avoid biases and allow important notions such as *unbiased estimation* and *confidence intervals*. While many statisticians (and most standard books on sampling) take random sampling as so self-evident that questions like "why do statisticians use dice or other random devices and add randomness or noise to the task" seem unwarranted[2], it is, in fact, an intriguing question that merits more than intuitive answers. Indeed, there is a large body of literature showing formally and precisely that certain relevant optimality criteria can only be achieved by random sampling designs.

* Partially supported by Israel Science Foundation grant 473/04.

[1] For a scholarly survey of results until 1987 and numerous references, see Chaudhuri and Vos (1988).

[2] but see Valliant et al. (2000) for a refreshing change.

Emphasis in this chapter is placed on optimal inference. In the context of finite populations, optimality is most often expressed in terms of minimax results, which in general require random strategies. Other decision-theoretic notions such as loss and risk, admissibility, sufficiency, completeness, unbiasedness, uniformly minimum variance (UMV), Bayes procedures, and more, will also be discussed in connection with finite population sampling.

2. Notations and definitions

The following notation will be used throughout the chapter. A list of main notations appears in Section 8.

1. The ***population*** $\mathcal{Y} = (y_1, \ldots, y_N)$ is a vector of values of some measurements with index set $\mathcal{N} = \{1, \ldots, N\}$, where the population size N is assumed to be known whenever it is needed. Here $i \in \mathcal{N}$ denotes the ***label*** of the i-th population ***unit*** whose value is y_i. In this chapter, we assume that $\mathcal{Y} \in \mathbb{R}^N$, so that each y_i is a univariate measurement (although in many applications more than one variable is measured for each unit). Some of the ideas could be extended to more general measurements, but this will not be done here. $\mathcal{Y}$ is an unknown ***parameter***, and so is any function $\theta(\mathcal{Y})$ such as $\overline{\mathcal{Y}} = \frac{1}{N}\sum_{i=1}^N y_i$, $V(\mathcal{Y}) = \frac{1}{N}\sum_{i=1}^N (y_i - \overline{\mathcal{Y}})^2$, $Max(\mathcal{Y}) = \mathrm{Max}_{1\le i\le N} y_i$, or $Med(\mathcal{Y}) = \mathrm{Median}_{1\le i\le N} y_i$.

   The set of possible $\mathcal{Y}$'s is denoted by Υ, the ***parameter space***; unless otherwise stated (towards the end of Section 6), we shall **always** assume that Υ is a ***symmetric parameter space***, that is, a symmetric subset of $\mathbb{R}^N$ in the sense that if $\mathcal{Y} = (y_1, \ldots, y_N) \in \Upsilon$, then so does every permutation of $\mathcal{Y}$. In particular, any set of the form $\Upsilon = \Lambda \times \cdots \times \Lambda$, a product of some set N times, satisfies this assumption. The set $\Omega(\mathcal{Y})$ of all permutations of a given vector $\mathcal{Y} = (y_1, \ldots, y_N)$ is, of course, also symmetric. As usual, the parameter space Υ is known to the statistician.

   If the parameter $\theta(\mathcal{Y})$ remains constant under permutations of $\mathcal{Y}$, we say that it is a ***symmetric parameter***. The above examples are all of this kind.

2. A ***sampling design*** $\mathcal{P}$ is a probability function on the space of all subsets S of $\mathcal{N}$. Unless otherwise stated, we assume ***noninformative sampling***, also known as ***ignorable sampling***; that is, the probability $\mathcal{P}(S)$ does not depend on the parameter $\mathcal{Y}$. Formally, $\mathcal{P}(S \mid \mathcal{Y}) = \mathcal{P}(S)$. In the Bayesian or superpopulation context of Section 6, $\mathcal{Y}$ is also random, $\mathcal{P}(S \mid \mathcal{Y})$ becomes a conditional probability, and ignorability is equivalent to independence of S and $\mathcal{Y}$.

   In certain examples, we allow the design $\mathcal{P}$ to depend on known covariates or auxiliary variables; see below. The ***inclusion probability*** of a unit is defined by $\alpha_i = \mathcal{P}(\{i \in S\}) = \sum_{S: S \ni i} \mathcal{P}(S)$, the probability that unit i is in the sample S. Here S is the ***set*** of drawn labels (without order and repetitions). By a simple sufficiency argument given in Remark 1 below, we can ignore designs that take an order of the elements in the sampled set into account or allow repetitions.

   The set S is called the ***sample***, and its size, $|S|$, is the ***sample size***. If $\mathcal{P}(S) > 0$ implies $|S| = n$, then the design $\mathcal{P}$ is said to have a ***fixed sample size***.

Simple random sampling without replacement of size n, abbreviated **SRS**, is denoted by $\mathcal{P}_s$ and satisfies $\mathcal{P}_s(S) = 1/\binom{N}{n}$ if $|S| = n$, and zero otherwise.

When ***auxiliary*** information is available in the form of positive values $(x_1, \ldots, x_N)$, where x_i is some *known* value of a variable pertaining to unit $i \in \mathcal{N}$, it can be used in the design and in estimation. For example, when $x_i > 0$, the design having a fixed sample size n, defined by $\mathcal{P}_{\text{ppas}}(S) = \sum_{i\in S} x_i / \left[N\bar{\mathcal{X}}\binom{N-1}{n-1}\right]$ if $|S| = n$, where $\bar{\mathcal{X}} = \frac{1}{N}\sum_{i=1}^{N} x_i$, is of this kind (Lahiri, 1951). The notation *ppas* stands for ***probability proportional to aggregate size***, in this case, to the aggregate size of the auxiliary variables in S, $\sum_{i\in S} x_i$. It can be implemented by first choosing one unit from the population, say i, with probability $x_i/N\bar{\mathcal{X}}$, and then adding a subset of $n-1$ additional units chosen from the remaining $N-1$ units uniformly, that is, with equal probabilities for all subsets of size $n-1$. See Rao and Vijayan (1977) and Hedayat and Sinha (1991) for details and references on this design and a discussion of drawing mechanisms for design implementation, and Cassel et al. (1977) for further references.

3. The ***data*** consist of the set of pairs $\{(i, y_i) : i \in S\}$, that is, the y-values and their labels for the units in the sample S. We set

$$D = D[S, \mathcal{Y}] = \{(i, y_i) : i \in S\}. \tag{1}$$

For $S = \{i_1, \ldots, i_n\}$, let $\mathcal{Y}_S$ be the **multiset** $\{y_{i_1}, \ldots, y_{i_n}\}$, with equal y-values listed separately provided that they have different labels. In other words, $\mathcal{Y}_S$ can be viewed as the sequence $(y_{i_1}, \ldots, y_{i_n})$, where the order is ignored. For example, if $S = \{1, 2, 3\}$ and $y_1 = y_2 = 13$ and $y_3 = 7$, then $\mathcal{Y}_S = \{13, 13, 7\}$ in any order.

REMARK 1. *By sufficiency arguments (Basu, 1958) we shall consider the data D as above, that is, without taking into account the order (if known) in which the sample was drawn; when the sampling procedure allows repetitions of units, as in sampling with replacement, repetitions will also be ignored and each repeated unit will be counted once. Since the relevant data D consist only of the set of drawn labels S and their y-values, we shall only consider* ***designs $\mathcal{P}$ on the space of (unordered) subsets (with no repetitions)*** *of $\mathcal{N}$. The sufficiency of D is intuitively obvious: no information is added by measuring a unit more than once, or specifying the order in which the measurements were taken. A formal statement and proof follow. We denoted designs which ignore the order of labels and repetitions by $\mathcal{P}$ and the corresponding data by $\mathcal{D}$. In the proposition below, we consider designs that are probability measures on ordered multisets of $\mathcal{N}$, so repetitions are allowed, and the data contain information on order and repetitions. In this case, the sampling design and data are denoted by bold-face letters $\mathbf{P}$ and $\mathbf{D}$, respectively, and the sample is an ordered multiset (allowing repetitions) denoted by $\mathbf{S}$, distributed according to $\mathbf{P}$.*

PROPOSITION 2. *Let $\mathbf{P}$ be a sampling design on* ***ordered multisets*** *which we denote by $\mathbf{S}$, and consider the data $\mathbf{D} = \{(i, y_i) : i \in \mathbf{S}\}$, a multiset that includes information on the order and repetitions in the sample. Let $S = r(\mathbf{S}) = \{i : i \in \mathbf{S}\}$;*

that is, S is the set formed from $\mathbf{S}$ *when repetitions and order are ignored, and let* $D = r(\mathbf{D}) = \{(i, y_i) : i \in S\}$. *Then* D *is a* ***sufficient statistic*** *for the parameter* $\mathcal{Y}$.

PROOF. *For a design* $\mathbf{P}$ *as above, the conditional probability of* $\mathbf{D} = \{(i, y_i) : i \in \mathbf{S}\}$ *given* D, *where* $\mathbf{S}$ *is an ordered multiset and the parameter is* $\mathcal{Y}$, *satisfies*

$$P(\mathbf{D}|D) = \begin{cases} \mathbf{P}(\mathbf{S}) / \sum_{\mathbf{S}': r(\mathbf{S}')=S} \mathbf{P}(\mathbf{S}') & \text{if} \quad r(\mathbf{D}) = D \\ 0 & \text{otherwise}. \end{cases} \tag{2}$$

Since the right-hand side of Eq. (2) depends only on D *and not on the parameter, it follows that* D *is sufficient.* □

4. An ***estimator*** $t = t(D) = t(\{(i, y_i) : i \in S\})$ is a function of the data. We use various notations for $t(D)$, namely $t(D[S, \mathcal{Y}])$, or $t(S, \mathcal{Y})$. It should be emphasized that $t(S, \mathcal{Y})$ depends only on the data $\{(i, y_i) : i \in S\}$, that is, the labels and the labeled y-values in the sample. Note that when the sample size $|S|$ is not fixed, then implicit in the notation is the assumption that t is a function defined on arguments of different dimensions.

   If the estimator $t(S, \mathcal{Y}) = t(D[S, \mathcal{Y}])$ can be expressed as a function of $\mathcal{Y}_S$ alone, we write $t(S, \mathcal{Y}) = t(\mathcal{Y}_S)$ and say that t is ***symmetric*** (or invariant). Such an estimator depends on the y-values in the sample and not on their labels.

   Examples of symmetric statistics are the sample mean $\bar{y}_S = \frac{1}{|S|} \sum_{i \in S} y_i$ and variance $\frac{1}{|S|-1} \sum_{i \in S} (y_i - \bar{y}_S)^2$. However, the ***Horvitz–Thompson estimator*** $t_{\mathrm{HT}} = \sum_{i \in S} y_i/\alpha_i$ does require knowledge of the labels associated with each y-value, and, thus, it is **not** symmetric.

   When ***auxiliary*** information $(x_1, \ldots, x_N)$ is available for every unit in the population, it can be used in the sampling design and in estimation. For example, consider the ***ratio estimator*** of $\bar{\mathcal{Y}}$ defined by $t_R = (\bar{y}_S/\bar{x}_S)\bar{\mathcal{X}}$, where $\bar{x}_S = \frac{1}{|S|} \sum_{i \in S} x_i$; we denote it by t_R since $\bar{y}_S/\bar{x}_S$ is an estimator of the ratio $R = \bar{\mathcal{Y}}/\bar{\mathcal{X}}$. The estimator t_R is **not** symmetric since the computation of $\bar{x}_S$ requires knowledge of the labels in S (but not their pairing with the y-values). Note that if $\bar{\mathcal{X}}$ is known, and the population consists of the pairs, that is, $\mathcal{Z} = \{z_1 = (y_1, x_1), \ldots, z_N = (y_N, x_N)\}$, then t_R is a symmetric estimator for the population $\mathcal{Z}$.

   In this chapter, we assume $t \in \mathbb{R}$ and any value in $\mathbb{R}$ is allowed, regardless of the parameter space. For example, a proportion in a population of size N (see Section 3.5.2) is necessarily a rational number of the form k/N, but we allow an estimator t of this proportion to assume any real value; if certain values are undesired, the loss function should reflect it.
5. A pair $(\mathcal{P}, t)$ consisting of a sampling design and an estimator is called a ***strategy***. A ***class of strategies*** consists of all pairs $(\mathcal{P}, t)$ such that $\mathcal{P}$ belongs to some class of sampling designs and t belongs to some class of estimators.
6. A ***loss function*** $L(\tau, \mathcal{Y})$ represents a penalty paid in an estimation problem when the estimator assumes the value τ, and the value of the parameter is $\mathcal{Y}$. If $\theta = \theta(\mathcal{Y})$ is a parameter and $t = t(S, \mathcal{Y})$ is an estimator of θ, we may use the notation $L(t, \theta)$

for the loss. A common example is $L(t, \theta) = (t - \theta)^2$, the ***quadratic*** loss function (= squared error loss). A loss function is said to be ***symmetric*** if $L(\tau, \mathcal{Y})$ remains constant when $\mathcal{Y}$ is replaced by any permutation of its coordinates for any fixed τ. Clearly, if $\theta(\mathcal{Y})$ is a ***symmetric parameter***, that is, if it remains constant under permutations of $\mathcal{Y}$, then so does $L(\tau, \theta(\mathcal{Y}))$, and the loss is symmetric.

7. The ***risk*** of a strategy $(\mathcal{P}, t)$ for the population $\mathcal{Y}$ is the expected loss defined by

$$R(\mathcal{P}, t; \mathcal{Y}) := E_{\mathcal{P}} L(t, \mathcal{Y}) = \sum_S \mathcal{P}(S) L(t(D[S, \mathcal{Y}]), \mathcal{Y}) \tag{3}$$

where the sum extends over all subsets of $\mathcal{N}$.

An important special case is $R(\mathcal{P}, t; \mathcal{Y}) := \mathrm{MSE}(\mathcal{P}, t; \mathcal{Y}) := E_{\mathcal{P}}(t - \theta)^2$; one reason for the interest in this measure is that by Chebychev's inequality it provides a lower bound on confidence interval coverage: $\mathcal{P}(|t - \theta(\mathcal{Y})| \le c) \ge 1 - MSE(\mathcal{P}, t; \mathcal{Y})/c^2$ for each $\mathcal{Y} \in \Upsilon$. For unbiased estimators, the MSE coincides with the variance, which plays a role in the construction of confidence intervals based on the normal approximation. It is well known that the MSE of an estimator can be decomposed into the sum of its variance and the square of its bias.

8. The strategy $(\mathcal{P}, t)$ is said to be ***unbiased*** for $\theta = \theta(\mathcal{Y})$ if

$$E_{\mathcal{P}} t := \sum_S \mathcal{P}(S) t(D[S, \mathcal{Y}]) = \theta(\mathcal{Y}) \tag{4}$$

for all $\mathcal{Y} = (y_1, \ldots, y_N) \in \Upsilon$. In this case, we say that t is $\mathcal{P}$-**unbiased**.

Note that if $\mathcal{P}$ satisfies $\alpha_i = n/N$ for all $i = 1, \ldots, N$, then the sample mean $\bar{y}_S = \frac{1}{n} \sum_{i \in S} y_i$ is unbiased for the population average $\bar{\mathcal{Y}}$ and, more generally for any design $\mathcal{P}$, so is the estimator t_{HT}/N where $t_{\mathrm{HT}} = \sum_{i \in S} y_i/\alpha_i$ is the Horvitz–Thompson estimator, since by setting I_i to be the indicator of the event that $i \in S$ we have $E_{\mathcal{P}} I_i = \alpha_i$ and therefore,

$$E_{\mathcal{P}}[t_{\mathrm{HT}}/N] = E_{\mathcal{P}} \frac{1}{N} \sum_{i=1}^{N} I_i y_i/\alpha_i = \frac{1}{N} \sum_{i=1}^{N} y_i E_{\mathcal{P}} I_i/\alpha_i = \bar{\mathcal{Y}}. \tag{5}$$

More generally, the estimator $t = \frac{1}{N} \sum_{i \in S} y_i c_i(S)/\alpha_c(i)$, where $\alpha_c(i) = \sum_{S: S \ni i} c_i(S) \mathcal{P}(S)$, is easily seen to be $\mathcal{P}$-unbiased for $\bar{\mathcal{Y}}$. When $c_i(S) \equiv 1$ it reduces to t_{HT}.

Under SRS, the ratio estimator $t_R = (\bar{y}_S/\bar{x}_S)\bar{\mathcal{X}}$ is, in general, not unbiased. On the other hand, the strategy $(\mathcal{P}_{\mathrm{ppas}}, t_R)$, with $\mathcal{P}_{\mathrm{ppas}}$ defined above as probability proportional to $\sum_{i \in S} x_i$ sampling, is unbiased for $\bar{\mathcal{Y}}$, since

$$E_{\mathrm{ppas}}\, t_R = \sum_S \Big(\sum_{i \in S} x_i \Big) \Big/ \left[N \bar{\mathcal{X}} \binom{N-1}{n-1} \right] (\bar{y}_S/\bar{x}_S) \bar{\mathcal{X}}$$

$$= \binom{N-1}{n-1}^{-1} \frac{1}{N} \sum_S \sum_{i \in S} y_i = \bar{\mathcal{Y}}.$$

To compare the above notions for finite populations with standard statistical decision theory, we give the following concise definitions, to be followed by a short discussion. For further details see, for example, Ferguson (1967) and Lehmann and Casella (1998).

DEFINITION 3. • An **observation** is a random variable $X \sim P_\theta$ (i.e., X has the distribution P_θ), where $\theta \in \Theta$, the **parameter space**.

- A **decision rule** $\delta(X)$ is a function taking values in a decision space $\mathcal{A}$, or a distribution on $\mathcal{A}$ (which may depend on X) in which case δ is **randomized**. The decision space is sometimes identical to the parameter space.
- $L(a, \theta)$ is the **loss** due to a decision $a \in \mathcal{A}$, and if δ is randomized, we set $L(\delta, \theta) = E_\delta L(a, \theta)$ where $a \sim \delta = \delta(X)$. The **risk** is defined by $R(\delta, \theta) = EL(\delta(X), \theta)$, where the expectation is with respect to $X \sim P_\theta$. Given a prior distribution ρ for θ, the **Bayes risk** is $r(\rho, \delta) = E_\rho R(\delta, \theta) = \int R(\delta, \theta) d\rho(\theta)$.
- A decision rule δ_0 is **Bayes with respect to** ρ if $r(\delta_0, \rho) = \inf_\delta r(\delta, \rho)$. It is a **minimax** rule if $\sup_\theta R(\delta_0, \theta) = \inf_\delta \sup_\theta R(\delta, \theta)$. The rule δ_0 has a **uniformly minimal risk** among **unbiased** estimators of $g(\theta)$ if $E\delta_0(X) = g(\theta)$, that is, δ_0 is unbiased, and $R(\delta_0, \theta) \le R(\delta, \theta)$ for all $\theta \in \Theta$ and any unbiased rule δ. In the case of MSE risk, the latter δ_0 has the UMV among Unbiased estimators (UMVU) property.

In standard decision theory as given in Definition 3, the distribution of the data is prescribed as part of the problem, and optimization is done only with respect to the decision rule or the estimator. In contrast, when we study strategies in finite population sampling, we attempt to optimize over both the estimator and the sampling design. The latter determines the data collection method and the distribution of the data, and in this sense optimality in finite population sampling is more comprehensive than classical decision theory.

REMARK 4. *Henceforth, we consider only* ***nonrandomized*** *estimators unless otherwise stated (as when we consider nonconvex loss in Section 3.4, and the Rao–Hartley–Cochran strategy in Section 7.2). When the loss function $L(a, \theta)$ (or $L(\tau, \mathcal{Y})$) is convex in the variable a (or τ), as in the quadratic loss case, then randomized estimators can be replaced by nonrandom ones having a smaller risk. In fact, by Jensen's inequality the risk of a randomized estimator can only decrease when the estimator is replaced by its expectation (assuming it is finite), which is a nonrandomized estimator.*

One may now ask whether randomization in the sampling design can also be eliminated in a similar way under some convexity conditions, that is, can a design $\mathcal{P}$ be replaced by a deterministic sample with a smaller risk. However, this cannot be done since the relevant space is not convex: there is no "average" or "expected" set for a given design.

3. Minimax strategies

3.1. Definitions and discussion

DEFINITION 5. A strategy $(\mathcal{P}_0, t_0)$ is said to be **minimax** relative to a given class of strategies if it belongs to this class, and

$$\sup_{\mathcal{Y} \in \Upsilon} R(\mathcal{P}_0, t_0; \mathcal{Y}) \le \sup_{\mathcal{Y} \in \Upsilon} R(\mathcal{P}, t; \mathcal{Y}) \tag{6}$$

for every strategy $(\mathcal{P}, t)$ in the given class of strategies.

The estimator t_0 is said to be **minimax under** $\mathcal{P}$ in a class of estimators, if

$$\sup_{\mathcal{Y}\in\Upsilon} R(\mathcal{P}, t_0; \mathcal{Y}) \le \sup_{\mathcal{Y}\in\Upsilon} R(\mathcal{P}, t; \mathcal{Y}) \tag{7}$$

for any estimator t in the class.

In Eq. (6) and below, the sup may be replaced by max when the latter exists. With quadratic loss function, $\sup_{\mathcal{Y}\in\Upsilon} R(\mathcal{P}, t; \mathcal{Y})$ becomes $\sup_{\mathcal{Y}\in\Upsilon}\mathrm{MSE}(\mathcal{P}, t; \mathcal{Y})$, which is $\sup_{\mathcal{Y}\in\Upsilon}\mathrm{Var}_{\mathcal{P}}(t)$ for unbiased estimators.

A minimax strategy guarantees the lowest maximal risk, that is, the smallest risk in the *worst case* or the worst possible $\mathcal{Y}$. If we denote the left-hand side of Eq. (6) by v_0, then using the strategy $(\mathcal{P}_0, t_0)$, we are guaranteed a risk of at most v_0 whatever the population values $\mathcal{Y}$ are, and a lower value *for all* $\mathcal{Y}$ cannot be guaranteed.

It turns out that minimax strategies involve random sampling. Strategies that avoid randomization are in general not minimax and hence may yield very poor estimates for certain populations $\mathcal{Y}$. Randomization guarantees that the sample "represents" the population (with probability that increases with the sample size). Any fixed sample could be very biased relative to certain populations. For example, the mean of a sample consisting of the first n labels from an ordered $\mathcal{Y}$ would be a poor estimate of the population mean, and such poor samples are avoided with high probability by randomization. This is why regulatory agencies insist on randomization, and perhaps also in order to prevent biased experimenters who have some partial knowledge of $\mathcal{Y}$ from choosing a biased sample that would prove their point rather than yield good estimates.

Minimax strategies are particularly relevant in zero-sum games, where maximizing one's own gain is equivalent to minimizing one's opponent's gain. The view of a statistical problem as a game between a statistician who chooses a strategy $(\mathcal{P}_0, t_0)$ and *nature* which "chooses" the parameter value, appears in well-known texts such as Blackwell and Girshick (1954) and Ferguson (1967). Random sampling is equivalent to a mixed strategy of the statistician, that is, a strategy which chooses the action (in this case, the sample S) at random according to a certain probability law (which in our case is $\mathcal{P}$). In general, minimax strategies are mixed strategies. Thus, the minimax criterion leads naturally to random sampling. One may argue that nature should not be considered a strategic player who uses the worst possible (for the statistician) or least favorable $\mathcal{Y}$ as a player in a zero-sum game, and question the minimax approach and the relevance of zero-sum games. However, the protection against a worst-case population appears quite reasonable when prior knowledge of the populations is very limited.

Brewer (1963) and Royall (1970b) present optimality results where the sup's in Eq. (6) are replaced by expectations with respect to a prior (superpopulation model) on $\mathcal{Y}$ satisfying certain conditions that are expressed in terms of covariates. The resulting optimal design, which may be very sensitive to the choice of a prior, is nonrandom: averaging over a prior replaces the need for averaging by a random design. This approach is analogous to average-case or probabilistic analysis of algorithms in computer science, whereas the minimax approach pertains to worst-case evaluations.

While protecting against the worst case in the parameter space, minimax rules may sometimes be relatively unsatisfactory in other parts of that space. An example is given in Section 3.5.

3.2. *Some minimax results through symmetry (invariance)*

Invariance or symmetry has a long history in statistics. Symmetrization of strategies (as in Eq. (9) below) appears in Blackwell and Girshick (1954), and in Kiefer (1957) with reference to work of Hunt and Stein from the 1940s.

The first step towards finding minimax strategies through symmetry is to show that it suffices to search among strategies consisting of symmetric estimators and a (conditional) SRS design. This is formulated in Proposition 6. Part of the notation below follows Stenger (1979).

Let Π denote the group of permutations of $\mathcal{N} = \{1, \ldots, N\}$, and for $\pi \in \Pi$, $S \subseteq \mathcal{N}$, and $\mathcal{Y} = (y_1, \ldots, y_N)$, define

$$\pi S = \{\pi(i) : i \in S\}, \quad \pi\mathcal{Y} = (y_{\pi^{-1}(1)}, \ldots, y_{\pi^{-1}(N)}). \tag{8}$$

For a design $\mathcal{P}$ let

$$\bar{\mathcal{P}}(S) = \sum_{\pi\in\Pi} \mathcal{P}(\pi S)/N!, \quad \bar{t}_{\mathcal{P}}(S, \mathcal{Y}) = \frac{1}{N!\bar{\mathcal{P}}(S)} \sum_{\pi\in\Pi} t(\pi S, \pi\mathcal{Y})\mathcal{P}(\pi S). \tag{9}$$

Note that $\bar{\mathcal{P}}(S)$ is a probability on subsets S of $\mathcal{N}$. If the design $\mathcal{P}$ concentrates on sets of size n, then $\bar{\mathcal{P}}$ is a uniform probability with $\bar{\mathcal{P}}(S) = 1/\binom{N}{n}$ on such sets, that is, $\bar{\mathcal{P}}(S) = \mathcal{P}_s(S)$, which is SRS. If $\mathcal{P}(|S| = m) = \gamma_m$, $m = 1, \ldots, N$, where $|S|$ denotes the size of S, then $\bar{\mathcal{P}}(S) = \gamma_{|S|}/\binom{N}{|S|}$. In this case, $\bar{\mathcal{P}}(S)$ is uniform over all sets of a given size, and we call it a ***conditional*** *SRS*. Moreover, $\bar{\mathcal{P}} = \mathcal{P}$ if and only if $\mathcal{P}$ is a conditional SRS.

Let us now consider a random pair (π, S), consisting of a random permutation and a random set having the joint distribution $(\pi, S) \sim \frac{\mathcal{P}(\pi S)}{N!}$. It is easy to see that $\sum_\pi \sum_S \frac{\mathcal{P}(\pi S)}{N!} = 1$, and by Eq. (9) we have $\sum_\pi \frac{\mathcal{P}(\pi S)}{N!} = \bar{\mathcal{P}}(S)$, the marginal distribution of S. The conditional distribution of π given S is the ratio of the joint distribution $\frac{\mathcal{P}(\pi S)}{N!}$ and marginal distribution $\bar{\mathcal{P}}(S)$ of S. We summarize this notation as follows:

$$(\pi, S) \sim \frac{\mathcal{P}(\pi S)}{N!}, \quad S \sim \bar{\mathcal{P}}(S), \quad \pi|S \sim \frac{\mathcal{P}(\pi S)}{N!\bar{\mathcal{P}}(S)}. \tag{10}$$

Then, $\bar{t}_{\mathcal{P}}(S, \mathcal{Y}) = E[t(\pi S, \pi\mathcal{Y}) \mid S] = E_{\pi|S}[t(\pi S, \pi\mathcal{Y})]$.

Note that $D[\pi S, \pi\mathcal{Y}] = \{(\pi(i), y_{\pi^{-1}\pi(i)}) : i \in S\} = \{(\pi(i), y_i) : i \in S\}$, and so $\bar{t}_{\mathcal{P}}(S, \mathcal{Y})$ does not depend on y-values outside of $\mathcal{Y}_S$. The same is true for any $t(\pi S, \pi\mathcal{Y})$ with known π. Assume without loss of generality that $S = \{1, \ldots, n\}$ and use the notation $\pi(i) = j_i$ for $i \in S$. From Eq. (9) we have

$$\bar{t}_{\mathcal{P}}(S, \mathcal{Y}) = c \sum_{\{j_1,\ldots,j_n\}} t(\{(j_1, y_1), \ldots, (j_n, y_n)\})\mathcal{P}(\{j_1, \ldots, j_n\}),$$

where c is a constant. The sum is over all subsets of size n and does not depend on S, and it follows that $\bar{t}_{\mathcal{P}}(S, \mathcal{Y})$ depends on $\mathcal{Y}_S$ (and $\mathcal{P}$), but not on S; that is, $\bar{t}_{\mathcal{P}}$ is a symmetric estimator.

It is now easy to see that for a symmetric estimator $t(\mathcal{Y}_S)$ we have

$$t(\pi S, \pi\mathcal{Y}) = t(S, \mathcal{Y}) \quad \text{and, therefore,} \quad \bar{t}_{\mathcal{P}}(S, \mathcal{Y}) = t_{\mathcal{P}}(S, \mathcal{Y}). \tag{11}$$

From the definitions in Eq. (9) it is easy to see that if $(\mathcal{P}, t)$ is an **unbiased strategy** for a *symmetric* parameter $\theta = \theta(\mathcal{Y})$, that is, a parameter satisfying $\theta(\pi\mathcal{Y}) = \theta(\mathcal{Y})$ for all $\pi \in \Pi$, then so is the strategy $(\bar{\mathcal{P}}, \bar{t})$.

The following proposition (see Gabler (1990) and references therein for closely related results) implies that for a minimax strategy relative to designs of fixed sample size, it suffices to search among strategies $(\mathcal{P}_s, t)$, where $t = t(\mathcal{Y}_S)$ is symmetric and $\mathcal{P}_s$ is SRS with the same sample size. More generally, it shows that for any strategy $(\mathcal{P}, t)$ there is a strategy consisting of a conditional SRS design having the same distribution of sample size as $\mathcal{P}$ and a symmetric estimator $t(\mathcal{Y}_S)$ with a smaller maximal risk. The proposition requires symmetry and convexity of L. A closely related result that does not require convexity of the loss is Proposition 13. The latter proposition provides an interpretation of the third expression in Eq. (13) below in terms of a randomized estimator. We defer the discussion to Section 3.4 for the sake of simplicity at this point.

Recall that a loss function L is symmetric if it remains constant under permutations of $\mathcal{Y}$; that is, $L(\tau, \mathcal{Y}) = L(\tau, \pi\mathcal{Y})$. With the above definitions we have,

PROPOSITION 6. *Let $L(\tau, \mathcal{Y})$ be a symmetric loss function that is convex in τ for each $\mathcal{Y} \in \Upsilon$, a symmetric parameter space. Then for $\bar{\mathcal{P}}, \bar{t}$ defined in Eq. (9),*

$$\sup_{\mathcal{Y}\in\Upsilon} R(\bar{\mathcal{P}}, \bar{t}; \mathcal{Y}) \le \sup_{\mathcal{Y}\in\Upsilon} R(\mathcal{P}, t; \mathcal{Y}). \tag{12}$$

PROOF.

$$\begin{aligned} R(\bar{\mathcal{P}}, \bar{t}; \mathcal{Y}) &= \sum_S L(\bar{t}(S, \mathcal{Y}), \mathcal{Y})\bar{\mathcal{P}}(S) \overset{(0)}{\le} \sum_S \sum_\pi \frac{\mathcal{P}(\pi S)}{N!\bar{\mathcal{P}}(S)} L(t(\pi S, \pi\mathcal{Y}), \mathcal{Y})\bar{\mathcal{P}}(S) \\ &= \sum_S \sum_\pi \frac{\mathcal{P}(\pi S)}{N!} L(t(\pi S, \pi\mathcal{Y}), \mathcal{Y}) \overset{(1)}{=} \sum_\pi \sum_S \frac{\mathcal{P}(S)}{N!} L(t(S, \pi\mathcal{Y}), \mathcal{Y}) \\ &\overset{(2)}{=} \frac{1}{N!} \sum_\pi \sum_S \mathcal{P}(S) L(t(S, \pi\mathcal{Y}), \pi\mathcal{Y}) = \frac{1}{N!} \sum_\pi R(\mathcal{P}, t; \pi\mathcal{Y}); \end{aligned} \tag{13}$$

□

Jensen's inequality applied to the convexity of the loss function L implies the inequality marked by (0); a further explanation is given below. The equality (1) was obtained by substituting S for πS (both range over all subsets of $\mathcal{N}$ under the summation on S) and (2) follows because by symmetry $L(\tau, \mathcal{Y}) = L(\tau, \pi\mathcal{Y})$.

A simple way to understand the above inequality (0) is to note that under the definitions of Eq. (10),

$$\sum_S \sum_\pi \frac{\mathcal{P}(\pi S)}{N!\bar{\mathcal{P}}(S)} L(t(\pi S, \pi\mathcal{Y}), \mathcal{Y})\bar{\mathcal{P}}(S) = E_S\{E_{\pi|S}[L(t(\pi S, \pi\mathcal{Y}), \mathcal{Y})]\},$$

and the inequality becomes $E_S L(E_{\pi|S}[t(\pi S, \pi\mathcal{Y})],\ \mathcal{Y}) \le E_S\{E_{\pi|S}[L(t(\pi S, \pi\mathcal{Y}), \mathcal{Y})]\}$.

From Eq. (13) we have $R(\bar{\mathcal{P}}, \bar{t}; \mathcal{Y}) \le \max_\pi R(\mathcal{P}, t; \pi\mathcal{Y})$ since the maximum is larger than the average, and by the symmetry (permutation invariance) of the parameter space

Υ, it follows that

$$\sup_{\mathcal{Y}\in\Upsilon} R(\bar{\mathcal{P}}, \bar{t}; \mathcal{Y}) \le \sup_{\mathcal{Y}\in\Upsilon} R(\mathcal{P}, t; \mathcal{Y}). \qquad \blacksquare$$

Note that for a (conditional) SRS design $\mathcal{P}$ we have $\mathcal{P} = \bar{\mathcal{P}}$, and we conclude from Proposition 6 that for such designs it suffices to consider symmetric estimators when the goal is to minimize maximal risk with convex loss. This is formulated in the corollary below, which appears in Royall (1970a).

COROLLARY 7. Let $\mathcal{P}$ be a conditional SRS design. Under the conditions of Proposition 6

$$\sup_{\mathcal{Y}\in\Upsilon} R(\mathcal{P}, \bar{t}; \mathcal{Y}) \le \sup_{\mathcal{Y}\in\Upsilon} R(\mathcal{P}, t; \mathcal{Y}).$$

Our next goal is to establish a minimax result for unbiased strategies. First, we need the following lemma on completeness, due to Royall (1968). As usual, completeness will be used to obtain uniqueness of unbiased estimators. A function $h(y_1, \ldots, y_n)$ is said to be ***symmetric*** if it is invariant under permutations of its arguments, that is, it depends only on the set of (unordered) values $\{y_1, \ldots, y_n\}$. We can then write $h(\mathcal{Y}_S)$, since $\mathcal{Y}_S$ is a set of (unordered) y-values.

LEMMA 8. *Let Υ be any product parameter space*

$$\Upsilon = \Lambda^N = \Lambda \times \cdots \times \Lambda, \tag{14}$$

and let $\mathcal{P}$ be a design such that $\mathcal{P}(S) > 0$ implies $|S| = n$. Then $\mathcal{Y}_S$ is complete; that is, for any symmetric function h, $E_{\mathcal{P}} h(\mathcal{Y}_S) = 0$ for all $\mathcal{Y} = (y_1, \ldots, y_N) \in \Upsilon$ implies that $h(y_1, \ldots, y_n) = 0$ for any $y_i \in \Lambda,\ i = 1, \ldots, n$.

PROOF. First consider $\mathcal{Y} = (a, \ldots, a) \in \Upsilon$ (here, e.g., we use the structure of Υ given by Eq. (14)) and compute the expectation under this value of the parameter. Then $0 = E_{\mathcal{P}} h(\mathcal{Y}_S) = \sum_S \mathcal{P}(S) h(a, \ldots, a)$ implies $h(a, \ldots, a) = 0$. Assuming without loss of generality that $\mathcal{P}(S) > 0$ for $S = \{1, \ldots, n\}$ choose now $\mathcal{Y} = (b, a, \ldots, a) \in \Upsilon$. Then $0 = E_{\mathcal{P}} h(\mathcal{Y}_S) = ph(a, \ldots, a) + qh(b, a, \ldots, a)$ with $q > 0$, and we conclude that $h(b, a, \ldots, a) = 0$. The result follows by continuing in the same manner (induction). □

The next theorem shows that relative to the class of unbiased strategies, there exist minimax strategies that involve SRS. For a closely related result see Theorem 3.10 in Cassel et al. (1977) and references therein. Remark 11 compares their result to Theorem 9 below. Such a result is not true without restricting the class to unbiased strategies (see Remark 11 below). Unbiasedness is ubiquitous in applications. This is quite natural since avoiding bias is often given as a justification for random sampling. However, unbiasedness alone does not guarantee good estimation; see, for example, Basu's (1971) famous circus-elephants weighing example for a ridiculously poor unbiased estimator.

THEOREM 9. *Let Υ be any product parameter space: $\Upsilon = \Lambda^N = \Lambda \times \cdots \times \Lambda$, and let $\mathcal{P}_s$ denote SRS of size n. If there exists any unbiased strategy $(\mathcal{P}, t)$ for the parameter $\theta = \theta(\mathcal{Y})$ with $\mathcal{P}$ having a fixed sample size n, then there exists a unique symmetric estimator $t_0 = t_0(\mathcal{Y}_S)$ depending only on $\mathcal{Y}_S$, such that the strategy $(\mathcal{P}_s, t_0)$ is unbiased.*

If $\theta = \theta(\mathcal{Y})$ is a symmetric parameter, and the loss function $L(\tau, \theta)$ is convex in τ for each $\mathcal{Y} \in \Upsilon$, then,

$$\sup_{\mathcal{Y}\in\Upsilon} R(\mathcal{P}_s, t_0; \mathcal{Y}) \leq \sup_{\mathcal{Y}\in\Upsilon} R(\mathcal{P}, t; \mathcal{Y}) \tag{15}$$

*for any unbiased strategy $(\mathcal{P}, t)$ for θ, having a fixed samples size n. In other words, the strategy $(\mathcal{P}_s, t_0)$ is **minimax** relative to the above class of strategies $(\mathcal{P}, t)$.*

PROOF. As mentioned after Eq. (11), if $(\mathcal{P}, t)$ is unbiased then so is $(\bar{\mathcal{P}}, \bar{t})$. Since $\mathcal{P}$ concentrates on sets of size n, we have $\bar{\mathcal{P}} = \mathcal{P}_s$. Lemma 8 implies that a symmetric unbiased estimator is unique (take h in the lemma to be the difference between two unbiased estimators to obtain that they are the same). It follows that the strategy $(\bar{\mathcal{P}}, \bar{t})$ is the same for all unbiased $(\mathcal{P}, t)$, and the result follows from (12) with $t_0 = \bar{t}$.

The sup's in Eq. (15) may be infinite, in which case the result is uninteresting, and it is empty if no unbiased strategies exist. □

COROLLARY 10. Let $\Upsilon = \Lambda^N$. If $\theta = \theta(\mathcal{Y}) = \frac{1}{N}\sum_{i=1}^N y_i = \bar{\mathcal{Y}}$, the population mean, and the loss function $L(\tau, \theta)$ is convex in τ for each $\mathcal{Y} \in \Upsilon$, then for the sample mean $\bar{y}_S = \frac{1}{n}\sum_{i\in S} y_i$ we have

$$\sup_{\mathcal{Y}\in\Upsilon} R(\mathcal{P}_s, \bar{y}_S; \mathcal{Y}) \leq \sup_{\mathcal{Y}\in\Upsilon} R(\mathcal{P}, t; \mathcal{Y}) \tag{16}$$

for any unbiased strategy $(\mathcal{P}, t)$ for θ, having a fixed sample size n. In other words, the strategy $(SRS, \bar{y}_S)$ is **minimax** relative to the above class of strategies $(\mathcal{P}, t)$.

For the population variance $\theta = V(\mathcal{Y}) = \frac{1}{N}\sum_{i=1}^N (y_i - \bar{\mathcal{Y}})^2$, set $v = \frac{N-1}{N(n-1)}\sum_{i\in S}(y_i - \bar{y}_S)^2$, an unbiased estimator. Then $(SRS,\ v)$ is minimax relative to unbiased strategies having sample size n.

PROOF. The population mean $\theta = \bar{\mathcal{Y}}$ is a symmetric parameter, and the sample mean $\bar{y}_S$ is a $\mathcal{P}_s$-unbiased estimator, that is, it is unbiased for SRS of size n. The result follows from Theorem 9. Similarly, v above is symmetric and a $\mathcal{P}_s$-unbiased estimator of the population variance, which is a symmetric parameter. □

REMARK 11. *Theorem 3.10 of Cassel et al. (1977) states a result similar to Theorem 9 for the special case of $\theta = \bar{\mathcal{Y}}$, the population mean, and for the quadratic loss function (MSE). It states that in this case an unbiased strategy $(\mathcal{P}_1,\ \bar{y}_S)$ with sample size n is minimax relative to the class of unbiased strategies with sample size n, for any such $\mathcal{P}_1$ satisfying $\alpha_i = n/N$ for $i = 1, 2, \ldots, N$, and it seems that all they require of the parameter space is for it to be symmetric.*

In the counterexamples below, we also consider quadratic loss and estimation of the population mean. The first example shows that the assumption $\alpha_i = n/N$ does not suffice.

Set $N=4$, $n=2$, and $\Upsilon=\{0, 2a\}^4$. Then for quadratic loss a straightforward calculation shows that $\max_{\mathcal{Y}\in\Upsilon} R(\mathcal{P}_s, \bar{y}_S; \mathcal{Y}) = a^2/2$ *and the maximum is attained at* $\mathcal{Y} = (0, 0, 0, 2a)$. *On the other hand, the design defined by* $\mathcal{P}_1(\{1, 2\}) = \mathcal{P}_1(\{3, 4\}) = 1/2$ *satisfies* $\alpha_i = n/N = 1/2$, *but for* $\mathcal{Y} = (0, 0, 2a, 2a)$ *one easily gets* $R(\mathcal{P}_1, \bar{y}_S; \mathcal{Y}) = a^2$ *and clearly* $(\mathcal{P}_1, \bar{y}_S)$ *is not minimax.*

The next example shows that even in the case of SRS it is not enough to assume that Υ *is symmetric. Set* $\Upsilon = \Omega(1, 2, 3) \bigcup \Omega(11, 12, 13)$, *that is, the set consisting of the two indicated vectors and all their permutations. Here* $N = 3$. *Then for SRS with* $n = 1$ *or* $n = 2$, *there clearly exists an (unbiased) estimator* t *which is always exactly correct, and hence satisfies* $R(\mathcal{P}_s, t; \mathcal{Y}) = 0$, *whereas* $R(\mathcal{P}_s, \bar{y}_S; \mathcal{Y}) > 0$, *so that* $(\mathcal{P}_s, \bar{y}_S)$ *is not minimax. This happens because one observation provides complete information about the population up to permutations (recall that the parameter space is assumed known).*

Finally, we show by a simple example that the unbiasedness condition is not redundant. (For MSE, this will become clear also in Section 3.5.) In fact, a ***biased*** *estimate* t *may satisfy* $\max_{\mathcal{Y}\in\Upsilon} R(\mathcal{P}_s, t; \mathcal{Y}) < \max_{\mathcal{Y}\in\Upsilon} R(\mathcal{P}_s, \bar{y}_S; \mathcal{Y})$. *Take* $N = 2, \Upsilon = \{0, 1\}^2$, *and* $n = 1$. *Then,* $\max_{\mathcal{Y}\in\Upsilon} R(\mathcal{P}_s, \bar{y}_S; \mathcal{Y}) = (1/2)^2$. *The (biased) estimator* t *defined by* $t(0) = 1/4$, $t(1) = 3/4$, *satisfies* $\max_{\mathcal{Y}\in\Upsilon} R(\mathcal{P}_s, t; \mathcal{Y}) = (1/4)^2$.

The restriction to unbiased estimators may be replaced by linearity and invariance conditions and similar results to Theorem 9 still hold. Note that for the case of estimating $\bar{\mathcal{Y}}$, *for example, the minimax strategy* $(SRS, \bar{y}_S)$ *does not depend on* Υ. *Without unbiasedness or similar restrictions, the minimax strategy depends on* Υ, *and finding it may be difficult. For nonsymmetric parameter spaces the problem becomes even harder. See Proposition 17 below for a minimax rule on symmetric product parameter spaces for quadratic loss (MSE), without an unbiasedness condition.*

3.3. Symmetric estimators and nonconvex loss

The next proposition is a special case of a general result on invariance, Theorem 8.6.4 of Blackwell and Girshick (1954), who applied it in the context of sampling. It says that for **symmetric estimators** the maximal risk is minimized by (conditional) SRS designs. In particular, designs having a fixed sample size can be replaced by SRS. Since $R(\mathcal{P}, t; \mathcal{Y})$ is linear in $\mathcal{P}$ (but not in t), **convexity of the loss** L **is not required**. Also, we require no conditions on Υ other than symmetry, which is always assumed. Recall that for a given design $\mathcal{P}$ the corresponding $\bar{\mathcal{P}}$ is defined in Eq. (9).

PROPOSITION 12. *For any symmetric estimator* t, *design* $\mathcal{P}$, *and symmetric loss function* L,

$$\sup_{\mathcal{Y}\in\Upsilon} R(\bar{\mathcal{P}}, t; \mathcal{Y}) \le \sup_{\mathcal{Y}\in\Upsilon} R(\mathcal{P}, t; \mathcal{Y}).$$

If the design $\mathcal{P}$ *has a fixed sample size, say* n, *then*

$$\sup_{\mathcal{Y}\in\Upsilon} R(\mathcal{P}_s, t; \mathcal{Y}) \le \sup_{\mathcal{Y}\in\Upsilon} R(\mathcal{P}, t; \mathcal{Y}),$$

where, $\mathcal{P}_s$ *denotes SRS of size* n.

PROOF.

$$\begin{aligned}
\sup_{\mathcal{Y}\in\Upsilon} R(\bar{\mathcal{P}}, t; \mathcal{Y}) &= \sup_{\mathcal{Y}\in\Upsilon} \sum_S \bar{\mathcal{P}}(S) L(t(S, \mathcal{Y}), \mathcal{Y}) \\
&= \sup_{\mathcal{Y}\in\Upsilon} \sum_S \sum_\pi \frac{\mathcal{P}(\pi S)}{N!} L(t(S, \mathcal{Y}), \mathcal{Y}) \\
&= \sup_{\mathcal{Y}\in\Upsilon} \sum_\pi \sum_S \frac{\mathcal{P}(S)}{N!} L(t(\pi^{-1} S, \mathcal{Y}), \mathcal{Y}) \\
&\overset{(1)}{=} \sup_{\mathcal{Y}\in\Upsilon} \sum_\pi \sum_S \frac{\mathcal{P}(S)}{N!} L(t(S, \pi\mathcal{Y}), \mathcal{Y}) \\
&\overset{(2)}{=} \sup_{\mathcal{Y}\in\Upsilon} \sum_\pi \sum_S \frac{\mathcal{P}(S)}{N!} L(t(S, \pi\mathcal{Y}), \pi\mathcal{Y}) \\
&\leq \frac{1}{N!} \sum_\pi \sup_{\mathcal{Y}\in\Upsilon} \sum_S \mathcal{P}(S) L(t(S, \pi\mathcal{Y}), \pi\mathcal{Y}) \\
&= \frac{1}{N!} \sum_\pi \sup_{\mathcal{Y}\in\Upsilon} R(\mathcal{P}, t; \mathcal{Y}) = \sup_{\mathcal{Y}\in\Upsilon} R(\mathcal{P}, t; \mathcal{Y}); \qquad (17)
\end{aligned}$$

the equalities marked by (1) is obtained by the symmetry of t using the first part of Eq. (11), and (2) follows from the symmetry of L; the rest is straightforward. The second part of the proposition is a special case of the first, based on the fact that if $\mathcal{P}$ has a fixed sample size n, then $\bar{\mathcal{P}} = \mathcal{P}_s$. □

It is easy to provide a counterexample to the above result for t that is not symmetric. Take $N = 3$ and $n = 2$ and let $t(S, \mathcal{Y})$ be an estimator that takes a huge and irrelevant value when $3 \in S$. Clearly, one can choose values such that the design satisfying $\mathcal{P}(\{1, 2\}) = 1$ will violate the second inequality of Proposition 12.

3.4. Asymmetric and randomized estimators and nonconvex loss

So far, we have considered loss functions $L(\tau, \mathcal{Y})$ that are **convex** in τ, with one exception, namely Proposition 12. In sample survey applications, unbiased or nearly unbiased estimators are usually considered, and their variances or MSE are computed. This corresponds to quadratic loss, which is convex. In this section, we shall see (and it is well known) that for nonconvex loss functions and certain optimality criteria of statistical decision theory, randomized estimators become relevant, and we discuss them briefly in the next paragraph. Nonconvex loss functions arise, for example, in specific areas such as statistical classification, where a convex loss would tend to overemphasize misclassification of outliers, and more generally, when one wants to allow bounded loss over unbounded spaces. In this chapter, we treat general loss functions, including nonconvex ones, because we think that they may be useful and relevant, and because their discussion clarifies the analysis and shows what conditions are really needed, an issue that may be hidden in explicit calculations with quadratic loss.

For randomized estimators, standard decision theory suggests taking expectation of the loss over both the random estimator, and the design. This leads to the interpretation

of the loss for randomized estimators as given in Definition 3: $L(\delta, \theta) = E_\delta L(a, \theta)$ where $a \sim \delta = \delta(X)$. This interpretation, which in certain situations leads to optimality of randomized estimators as shown later, is perhaps relevant when a large number of similar estimation problems are considered together, with (roughly) the same value of the estimated parameter. The law of large numbers is then often used to justify the expectation. Perhaps one may consider repeated estimation of employment rates in a monthly Labor Force Survey to be such a situation. But in general, statistical agencies do not use randomized estimators, and their discussion in our context is theoretical.[3] The latter fact indicates that this approach to loss and randomized estimators is debatable. For example, when the randomization does not depend on the data, which is the case in most of the examples given later, this interpretation seems to violate the conditionality principle which many statisticians accept,[4] since it takes into account possible estimators which were not chosen to be used.

Like random sampling (which is, of course, used everywhere) randomized estimators may be seen as mixed strategies in game theoretical terminology; see, for example, Rubinstein (1991) and references therein for a discussion of the difficulty of interpreting mixed strategies in game theory, which pertains to statistics as well.

Below is a simple example showing that without convexity, randomized estimators must sometimes be taken into account. Consider for example the loss function of a *perfectionist* defined by $L(\tau, \theta) = 0$ if $\tau = \theta$ and $= 1$ otherwise,[5] and let θ be the population mean. Let $n = 1$, $\Upsilon = \{0, 1\}^2$, that is, $N = 2$. Then under SRS, for example, the randomized estimator t^* with $t^*(0) = 0$ or $= 1/2$ with probability $1/2$ each, and $t^*(1) = 1/2$ or $= 1$, again with probability $1/2$, is the minimax rule with risk = $1/2$. For convex loss function, a simple application of Jensen's inequality implies that we would achieve the same or smaller risk by averaging t^* to obtain the estimator t with $t(0) = 1/4$, $t(1) = 3/4$ (see Remark 11), which for quadratic loss is minimax. However, for the perfectionist's loss function the risk of t equals 1; it is an estimator that is never exactly correct.

The next result says that for any symmetric loss function (convex or not) and any strategy $(\mathcal{P}, t)$ having a fixed sample size n, one can find an estimator t^* such that the maximal risk of $(\mathcal{P}_s, t^*)$ is smaller than that of $(\mathcal{P}, t)$, where $\mathcal{P}_s$ denotes SRS of size n. This suggests that for minimax purposes or when considering maximal risk (and with the absence of auxiliary information), only SRS needs to be considered.

The estimator t^* turns out to be randomized, and its construction is given explicitly in Proposition 13 below. I cannot provide a reference for this proposition; it is probably not new, but if it is, it may be because little or no attention has been paid to nonconvex loss functions in finite population sampling. See Ferguson (1967 Theorem 4.3.1) for a related result, where randomized rules play a similar role. The proposition shows that the fact that SRS suffices for minimax considerations when estimating a symmetric parameter (which implies symmetric loss) is not related to convexity.

Given an estimator t, let $t_\pi(S, \mathcal{Y}) = t(\{(\pi(i), y_i) : i \in S\} = t(\pi S, \pi\mathcal{Y})$; see Eqs. (1) and (8) for notations. For example, if $t = t_{\mathrm{HT}}$ then $t_\pi(S, \mathcal{Y}) = \sum_{i \in S} y_i/\alpha_{\pi(i)}$. For a

[3] However, the Rao–Hartley–Cochran strategy mentioned in Section 7.2 is an example of a randomized estimator.

[4] See Helland (1995) for the history, a critical discussion, and references on the conditionality principle.

[5] A smoothed version of this function could be studied in similar ways.

strategy $(\mathcal{P}, t)$ with a fixed sample size, let t^* be the **randomized** estimator defined for a given S by

$$t^*(S, \mathcal{Y}) = t_\pi(S, \mathcal{Y}) = t(\pi S, \pi\mathcal{Y}) \quad \text{with probability} \quad \frac{\mathcal{P}(\pi S)}{N!\mathcal{P}_s(S)} \quad \text{for } \pi \in \Pi, \tag{18}$$

where Π is the permutation group over $\{1, \ldots, N\}$. Clearly $\sum_\pi \frac{\mathcal{P}(\pi S)}{N!\mathcal{P}_s(S)} = 1$.

PROPOSITION 13. *Let $L(\tau, \mathcal{Y})$ be a symmetric loss function (convex or not) and as always let Υ be a symmetric parameter space. Given a strategy $(\mathcal{P}, t)$ with fixed sample size n, let $t^*(S, \mathcal{Y})$ be the **randomized** estimator of Eq. (18). Then*

$$\sup_{\mathcal{Y}\in\Upsilon} R(\mathcal{P}_s, t^*; \mathcal{Y}) \le \sup_{\mathcal{Y}\in\Upsilon} R(\mathcal{P}, t; \mathcal{Y}). \tag{19}$$

PROOF. We just repeat part of the proof of Proposition 6. Using the notation of Eq. (8), but see also Eqs. (9) and (13), we have

$$\begin{aligned}
R(\mathcal{P}_s, t^*; \mathcal{Y}) &= \sum_S \sum_\pi \frac{\mathcal{P}(\pi S)}{N!\mathcal{P}_s(S)} L(t_\pi(S, \mathcal{Y}), \mathcal{Y})\mathcal{P}_s(S) \\
&= \sum_S \sum_\pi \frac{\mathcal{P}(\pi S)}{N!} L(t(\pi S, \pi\mathcal{Y}), \mathcal{Y}) \\
&\overset{(1)}{=} \sum_\pi \sum_S \frac{\mathcal{P}(S)}{N!} L(t(S, \pi\mathcal{Y}), \mathcal{Y}) \\
&\overset{(2)}{=} \frac{1}{N!} \sum_\pi \sum_S \mathcal{P}(S) L(t(S, \pi\mathcal{Y}), \pi\mathcal{Y}) \\
&= \frac{1}{N!} \sum_\pi R(\mathcal{P}, t; \pi\mathcal{Y}) \le \max_\pi R(\mathcal{P}, t; \pi\mathcal{Y}),
\end{aligned} \tag{20}$$

where the relations marked by (1) and (2) are explained under Eq. (13). The result now follows easily (compare to the proof of Proposition 6). □

A similar result to the above holds when $\mathcal{P}$ does not have a fixed sample size, in which case the left-hand side of Eq. (19) holds with $\mathcal{P}_s$ replaced by the conditional SRS design $\bar{\mathcal{P}}$.

The relation between Propositions 13 and 6 is as follows. If the loss is convex, we can replace t^* in Eq. (19) by its expectation, and obtain a lower bound by Jensen's inequality, and Proposition 6 follows; this is true also when the sample size is random.

We stated Propositions 12 and 13 separately only because in the context of finite population sampling, randomized estimators (which are needed to state Proposition 13) are esoteric. We could have stated just Proposition 13, since it implies Proposition 12 readily. To see this it suffices to note that by Eq. (11), if the estimator t is symmetric, then the estimator t^* defined in Eq. (18) is in fact nonrandomized, and $t^* = t$.

3.5. Minimax and Bayes estimators

Minimax estimators can be obtained from Bayesian calculations. An example of this approach concerning estimation of a proportion in a finite population is given with the purpose of demonstrating the technique. While minimizing the maximal risk by definition, the resulting minimax rule has a higher risk than the usual estimator, the sample proportion, in parts of the parameter space, and we discuss the comparison between the two estimators. Most of the following discussion and much more can be found in Lehmann and Casella (1998) and the references therein.

The problem of estimating a proportion in a finite population of size N by a sample of size n is first approximated by the standard decision-theoretic problem of estimating the parameter p from a binomial distribution, that is, a sample of iid Bernoulli(p) observations. The notation and terminology we need for the latter problem is that of Definition 3.

3.5.1. The binomial case

Consider $X \sim \text{Binomial}(n, p)$ and a Bayesian structure with a prior $p \sim \text{Beta}(a, b)$. For quadratic loss, it is well known that the Bayes estimator is the posterior expectation $d(X) = E(p|X)$. A standard calculation shows that the estimator,

$$d(X) = \frac{X}{n}\frac{\sqrt{n}}{1+\sqrt{n}} + \frac{1}{2(1+\sqrt{n})} \tag{21}$$

is Bayes with respect to the above prior when $a = b = \sqrt{n}/2$ and that it is an *equalizer*, that is, its risk is constant and does not depend on p. In fact, $E(d(X) - p)^2 = \frac{1}{4(1+\sqrt{n})^2}$. The following proposition is well known (see, e.g., Ferguson (1967) or Lehmann and Casella (1998)) and readily implies the minimax result of Corollary 15 below. For definitions see Definition 3.

PROPOSITION 14. *A Bayes estimator δ_0 having a constant risk (equalizer) is minimax. If δ_0 is uniquely Bayes with respect to a given prior, then it is the unique minimax estimator.*

PROOF. Let δ be another estimator, and assume δ_0 is Bayes with respect to ρ. The estimator δ_0 satisfies $r(\delta_0, \rho) = \int R(\delta_0, \theta)d\rho \le r(\delta, \rho) = \int R(\delta, \theta)d\rho$. As $R(\delta_0, \theta)$ is a constant not depending on θ, it follows that $R(\delta_0, \theta) \le \int R(\delta, \theta)d\rho \le \sup_\theta R(\delta, \theta)$ for all θ, and δ_0 is minimax. If another rule is minimax, then using the assumption of constant risk of δ_0, it is easy to see that it is also Bayes, and the uniqueness part follows. □

COROLLARY 15. The estimator $d(X)$ of Eq. (21) is the unique minimax estimator of p for quadratic loss.

For the estimator $d^*(X) = X/n$, which is UMVU, we have $E(d^*(X) - p)^2 = p(1-p)/n$, and we see that around $p = 1/2$ the estimator d is slightly better than d^* provided that n is not small, but d^* has smaller risk when p is not close to 1/2. Thus here, and in the developments below where a similar phenomenon occurs, one may argue about the quality of the estimator obtained by the minimax criterion.

3.5.2. Finite population sampling for proportion and mean

We now follow Lehmann and Casella (1998) and Hodges and Lehmann (1982). Related results appear in Bickel and Lehmann (1981). In Corollary 16 and Proposition 17 below, we obtain minimax results without restriction to unbiased estimators (compare to Corollary 10).

Consider SRS from a population of size N whose values are either 0 or 1, and we wish to estimate the parameter W/N where W is the number of ones. In fact, in this case $W/N = \bar{\mathcal{Y}}$. Consider the prior on W, which is a mixture of binomials with Beta(a, b) weights, that is,

$$P(W = w) = \int_0^1 \binom{N}{w} p^w (1-p)^{N-w} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} p^{a-1}(1-p)^{b-1} dp. \tag{22}$$

A reader who wants to avoid the calculations can look at numbered equations only. Let X be the number of ones drawn in an SRS of size n. Then $X|W \sim$ Hypergeometric, that is, $P(X = x|W) = \binom{W}{x}\binom{N-W}{n-x}/\binom{N}{n}$ and with standard calculations we have for some $c = c(x)$,

$$\begin{aligned} P(W = w|X = x) &= cP(X = x|W = w)P(W = w) \\ &= c\int_0^1 \binom{N-n}{w-x} p^{w+a-1} q^{N-w+b-1} dp \\ &= c\int_0^1 \binom{N-n}{k} p^k q^{N-n-k} \cdot p^{x+a-1} q^{n-x+b-1} dp\,; \end{aligned}$$

the last expression above is obtained by the substitution $k = w - x$ where $k = 0, \ldots, N - n$, and it is arranged so that under the integral we observe the Bin$(N - n, p)$ probability function in the variable k. It is now easy to see that $c = \Gamma(n+a+b)/[\Gamma(x+a)\Gamma(n-x+b)]$ is the normalizing constant so that $P(W = w|X = x)$ is a probability function. Using the Bin$(N - n, p)$ expectation we get,

$$E(W - x|X = x) = c\int_0^1 (N-n)p \cdot p^{x+a-1} q^{n-x+b-1} dp = \frac{(N-n)(x+a)}{n+a+b},$$

and, therefore, we obtain that the Bayes estimator is the linear estimator

$$d(x) = E(W/N|X = x) = \frac{(N+a+b)x + (N-n)a}{N(n+a+b)}. \tag{23}$$

To compute the MSE of d, we use the relations $E(X|W) = nW/N$ and $\mathrm{Var}(X|W) = Wn(N-W)(N-n)N^{-2}(N-1)^{-1}$ for the hypergeometric distribution of $X|W$, and the formula MSE(d) = Variance(d)+ [Bias(d)]^2. We then choose a, b which make the MSE constant (not dependent on W). The resulting equalizer estimator is given in Eq. (24) below and we obtain.

COROLLARY 16. Under SRS with sample size n and quadratic loss, the estimator

$$\begin{aligned} &d(X) = AX/n + B, \quad \text{where} \\ &\quad A = 1/[1 + \sqrt{(N-n)/(nN-n)}\ ], \quad B = (1-A)/2 \end{aligned} \tag{24}$$

is minimax among symmetric estimators (depending only on X, the number of ones in the sample) for the proportion W/N of ones in a finite population of zeros and ones.

We omit the calculations; clearly the obtained estimator is minimax, being an equalizer and Bayes. Naturally the estimator obtained in Eq. (24) converges to that of Eq. (21) for large N, and has similar properties: it is worse than the usual sample mean when W is not near $N/2$, and it is somewhat better near $N/2$.

As already mentioned, for parameter spaces that are not symmetric, minimax estimators depend on the parameter space, and their calculation may be difficult in the absence of further restrictions on the decision rules. Corollary 16 provides a minimax rule for estimating a proportion, in which case the parameter space is $\{0, 1\}^N$. We show next that for SRS on $\Upsilon = \Lambda^N$ and quadratic loss, Corollary 16 can be extended to provide a minimax estimator for any interval $\Lambda \subset \mathbb{R}$ (and more general sets). As noted by Lehmann and Casella (1998) (see references therein), this can be obtained from the previous discussion. See also Gabler (1990) for generalizations.

PROPOSITION 17. *Let $\Upsilon = \Lambda^N$, where $\Lambda = [a, b]$ for some $a \le b$. Let $\bar{\mathcal{Y}}$ and $\bar{y}_S$ denote the population and sample means. Under $\mathcal{P}_s$ = SRS with sample size n, and quadratic loss, the estimator $d_0 = (b-a)d((\bar{y}_S - a/(b-a)) + a$, where $d(z) = Az + B$ with A and B as defined in Eq. (24), is minimax for $\bar{\mathcal{Y}}$ relative to the class of all estimators. In the case $[a, b] = [0, 1]$ the minimax estimator is $d_0 = d(\mathcal{Y}_S) = A\bar{y}_S + B$. Moreover, the strategy $(\mathcal{P}_s, d_0)$ is minimax (see Definition 5) relative to the class of all strategies with a fixed sample size n.*

PROOF. We can assume first that $\Lambda = [0, 1]$, and then apply a linear transformation. By Corollary 7 we can restrict our attention to symmetric estimators t. By Corollary 16, the estimator of (24) is minimax among symmetric estimators when the parameter space is restricted to the set of extreme points of Υ, which we denote by $\Upsilon e = \{0, 1\}^N$. Let E below denote expectation with respect to the (prior) probability measure on Υe defined by $P(\mathcal{Y} = (y_1, \ldots, y_N)) = P(W = w)/\binom{N}{w}$ for any vector $(y_1, \ldots, y_N) \in \Upsilon e$, where $\sum_{i=1}^N y_i = w$, $w = 0, 1, \ldots N$, and the distribution of W is given in Eq. (22). Note that the estimator $d = A\bar{y}_S + B$ is Bayes with respect to this prior and an equalizer on Υe. We have,

$$\sup_{\mathcal{Y}\in\Upsilon} R(\mathcal{P}_s, t; \mathcal{Y}) \ge \sup_{\mathcal{Y}\in\Upsilon e} R(\mathcal{P}_s, t; \mathcal{Y}) \overset{(1)}{\ge} ER(\mathcal{P}_s, t; \mathcal{Y}) \overset{(2)}{\ge} ER(\mathcal{P}_s, d_0; \mathcal{Y})$$
$$\overset{(3)}{=} \sup_{\mathcal{Y}\in\Upsilon e} R(\mathcal{P}_s, d_0; \mathcal{Y}) \overset{(4)}{=} \sup_{\mathcal{Y}\in\Upsilon} R(\mathcal{P}_s, d_0; \mathcal{Y});$$

inequality (1) holds because an average is smaller than the maximum, (2) holds because d_0 is Bayes with respect to the given prior, and (3) holds because d_0 is an equalizer, that is, $R(\mathcal{P}_s, d; \mathcal{Y})$ is constant on Υe. Finally (4) follows from the fact that $R(\mathcal{P}_s, d; \mathcal{Y}) = \sum_S \mathcal{P}_s(S)(d(\mathcal{Y}_S) - \bar{\mathcal{Y}})^2$ is a convex function of $\mathcal{Y}$ and, therefore, its maximum is attained at the set of extreme points.

Proposition 6 readily implies that the strategy $(\mathcal{P}_s, d_0)$ is minimax as stated. □

It is easy to see that the above result holds for any bounded $\Lambda \subset \mathbb{R}$ satisfying $\{a, b\} \subseteq \Lambda \subseteq [a, b]$ for some $a, b \in \mathbb{R}$. By continuity arguments it also holds when $\Lambda = (a, b)$.

If Λ is not convex, the estimator may take a value that is not in the parameter space, which is allowed, as our decision space is always $\mathbb{R}$. Proposition 17 is trivial if Λ is unbounded since the maximal risk is always infinite.

For the parameter space $\Upsilon = \{\mathcal{Y} : \sum_{i=1}^{N}(y_i - \bar{\mathcal{Y}})^2 \leq M\}$, Bickel and Lehmann (1981) proved that under simple random sampling, the sample mean is minimax for quadratic loss. Since the variance of the sample mean is proportional to $\sum_{i=1}^{N}(y_i - \bar{\mathcal{Y}})^2$, this definition of the parameter space is equivalent to assuming that this variance is bounded by a given constant. The proof uses invariance, and a reduction of the problem to estimation of a translation parameter, and showing that the sample mean coincides with the Pitman estimator. If the population is divided into given **strata**, then the usual weighted average of the sample means in the strata is minimax when the parameter space is defined by the condition that its variance is bounded by a given constant. Results on optimal designs are also given.

Related results by Aggarwal (1959, 1966) for a superpopulation model are described in Section 7.3.

4. UMVU estimators

It may be natural to hope to find estimators that have a uniformly smallest risk in an interesting class of estimators. However, this short section describes a negative result, which indicates that such uniformly best estimators do not exist in interesting cases. This fact justifies "weaker" optimality criteria such as the minimax, which considers the maximal risk rather than the risk at each value of the parameter, or the Bayes risk, which averages the risk over the parameter space. For further references and discussions, see Cassel et al. (1977).

UMVU estimators are unbiased estimators whose MSE is smaller than that of any other unbiased estimator for each $\mathcal{Y} \in \Upsilon$. We consider more general risk functions than MSE but still use the term UMVU.

Definition 18. A $\mathcal{P}$-unbiased estimator t^* (of a parameter θ) that is in some class of estimators, is said to be UMVU in this class, under the design $\mathcal{P}$, if

$$R(\mathcal{P}, t^*; \mathcal{Y}) \leq R(\mathcal{P}, t; \mathcal{Y}) \quad \text{for all} \quad \mathcal{Y} \in \Upsilon \tag{25}$$

for any $\mathcal{P}$-unbiased estimator t of θ in this class.

We briefly discuss estimation of the population mean, and show that interesting cases of UMVU estimators do not exist; the condition that Eq. (25) hold for all $\mathcal{Y} \in \Upsilon$ is too strong.

Consider the so-called ***generalized difference estimator*** (Basu, 1971) defined for any design $\mathcal{P}$ with inclusion probabilities $\alpha_i > 0$ for all $i \in \mathcal{N}$ by

$$t_{\mathrm{GD}} = \sum_{i \in S} \frac{y_i - e_i}{\alpha_i} + \tilde{e}, \quad \text{where} \quad \tilde{e} = \sum_{i=1}^{N} e_i, \tag{26}$$

with known but arbitrary constants $\mathbf{e} = (e_1, \ldots, e_N)$. When $\mathbf{e} = \mathbf{0}$, we obtain the Horvitz–Thompson estimator. Note that for any $\mathbf{e}$ we have $\mathrm{E}_{\mathcal{P}}(t_{\mathrm{GD}}/N) = \bar{\mathcal{Y}}$.

PROPOSITION 19. *Let $\mathcal{P}$ be a design such that $\alpha_i > 0$ for all $i \in \mathcal{N}$ and $\alpha_i < 1$ for some $i \in \mathcal{N}$, and let $\theta = \bar{\mathcal{Y}}$. Let $\Upsilon = \Lambda^N$ with $\Lambda \subseteq \mathbb{R}$ such that $|\Lambda| \geq 2$. Consider a loss function $L(\tau, \theta)$ such that $L(\tau, \theta) \geq 0$ for all τ and θ, and $L(\tau, \theta) = 0$ if and only if $\tau = \theta$. Then no UMVU estimator in the class of unbiased estimators of the population mean θ exists.*

PROOF. If $\mathbf{e}$ happens to coincide with some $\mathcal{Y} \in \Upsilon$ then $t_{\mathrm{GD}}/N = \bar{\mathcal{Y}}$ for any sample S, and $R(\mathcal{P}, t_{\mathrm{GD}}/N; \mathcal{Y}) = 0$. It follows that a UMVU estimator t must satisfy $R(\mathcal{P}, t; \mathcal{Y}) = 0$ for all $\mathcal{Y} \in \Upsilon$. The result now follows readily. □

Since t_{GD} is in the class of unbiassed linear (or affine) estimators (a linear combination of the observations plus a constant), the proof shows that there is no UMVU estimator in this class. This was shown by Godambe (1955). For further references see Godambe and Joshi (1965).

Finally, we point out that *among symmetric unbiased estimators there do exist UMVU estimators* in a trivial manner. For example, the completeness result of Lemma 8 shows that $\bar{y}_S$ is the unique unbiased estimator of $\bar{\mathcal{Y}}$ that is symmetric, that is, an estimator of the form $t = t(\mathcal{Y}_S)$. Thus, $\bar{y}_S$ is trivially UMVU among symmetric estimators. More generally, if we restrict attention to the class of symmetric unbiased estimators of any parameter, then at most one such estimator exists, and it is trivially UMVU in this class.

5. Admissibility

DEFINITION 20. A strategy $(\mathcal{P}_0, t_0)$ is **admissible** in a class of strategies if there is no strategy $(\mathcal{P}, t)$ in this class satisfying

$$R(\mathcal{P}, t; \mathcal{Y}) \leq R(\mathcal{P}_0, t_0; \mathcal{Y}) \text{ for all } \mathcal{Y} \in \Upsilon \text{ with strict inequality for at least one } \mathcal{Y}.$$

An estimator t_0 is **admissible** under a design $\mathcal{P}_0$ in a class of estimators if there is no estimator t in this class satisfying

$$R(\mathcal{P}_0, t; \mathcal{Y}) \leq R(\mathcal{P}_0, t_0; \mathcal{Y}) \text{ for all } \mathcal{Y} \in \Upsilon \text{ with strict inequality for at least one } \mathcal{Y}.$$

If the first inequality in the above definition holds, we say that the strategy $(\mathcal{P}, t)$ ***dominates*** $(\mathcal{P}_0, t_0)$, and if the second inequality holds, we say that t ***dominates*** t_0 under $\mathcal{P}_0$.

Admissibility is in some sense a minimal property. If a strategy (estimator) is inadmissible, then there is a better strategy (estimator) that will perform better (or at least as well) under any of the criteria mentioned in this chapter. But an admissible strategy may still be very poor. For example, it is easy to construct a finite population estimation problem such that an estimator which is a constant guess that ignores the sample altogether is admissible in a wide class, but has an arbitrarily large risk on large parts of the parameter space. Admissibility is called *Pareto optimality* in the terminology of game theory.

The next two theorems, from Scott (1975), show that admissibility is a property of the support of the design $\mathcal{P}$ defined by $\mathcal{S}_{\mathcal{P}} = \{S \subseteq \mathcal{N} : \mathcal{P}(S) > 0\}$. In fact, if t_0 is admissible under a design $\mathcal{P}$, then it is admissible under any design having the same or

smaller support. Here convexity of the loss function and randomized rules play a role, as will be seen in the precise statements and proofs.

THEOREM 21. *If an estimator t_0 is admissible under a design $\mathcal{P}_0$ in the class of all estimators including randomized ones, then the same holds for any design $\mathcal{P}$ satisfying $\mathcal{S}_{\mathcal{P}} \subseteq \mathcal{S}_{\mathcal{P}_0}$.*

PROOF. Suppose to the contrary that there exists an estimator t_1 that dominates t_0 under $\mathcal{P}$. Using it, we will construct an estimator t^* that dominates t_0 under $\mathcal{P}_0$, contradicting our assumption.

Let $m = \max\{\mathcal{P}(S)/\mathcal{P}_0(S) : S \in \mathcal{S}_{\mathcal{P}}\}$ and $Q(S) = \mathcal{P}(S)/m\mathcal{P}_0(S)$. Then, by the conditions on the supports $1 \le m < \infty$ and $0 < Q(S) \le 1$. Here and in the next proof, we use the abbreviation $t(S)$ for $t(S, \mathcal{Y})$, that is, we suppress the parameter $\mathcal{Y}$. Consider the following randomized estimator t^*: if $S \in \mathcal{S}_{\mathcal{P}_0} \setminus \mathcal{S}_{\mathcal{P}}$, then $t^*(S) = t_0(S)$. If $S \in \mathcal{S}_{\mathcal{P}}$, then $t^*(S) = t_1(S)$ w.p. $Q(S)$ and $t^*(S) = t_0(S)$ w.p. $1 - Q(S)$. We claim that t^* dominates t_0 under $\mathcal{P}_0$. Indeed,

$$
\begin{aligned}
&R(\mathcal{P}_0, t^*; \mathcal{Y}) \\
&= \sum_{S \in \mathcal{S}_{\mathcal{P}}} Q(S)\mathcal{P}_0(S)L(t_1(S), \mathcal{Y}) + \sum_{S \in \mathcal{S}_{\mathcal{P}}} [1 - Q(S)]\mathcal{P}_0(S)L(t_0(S), \mathcal{Y}) + \sum_{S \in \mathcal{S}_{\mathcal{P}_0} \setminus \mathcal{S}_{\mathcal{P}}} \mathcal{P}_0(S)L(t_0(S), \mathcal{Y}) \\
&= m^{-1} \sum_{S \in \mathcal{S}_{\mathcal{P}}} \mathcal{P}(S)L(t_1(S), \mathcal{Y}) + \sum_{S \in \mathcal{S}_{\mathcal{P}}} [1 - Q(S)]\mathcal{P}_0(S)L(t_0(S), \mathcal{Y}) + \sum_{S \in \mathcal{S}_{\mathcal{P}_0} \setminus \mathcal{S}_{\mathcal{P}}} \mathcal{P}_0(S)L(t_0(S), \mathcal{Y}) \\
&\le m^{-1} \sum_{S \in \mathcal{S}_{\mathcal{P}}} \mathcal{P}(S)L(t_0(S), \mathcal{Y}) + \sum_{S \in \mathcal{S}_{\mathcal{P}}} [1 - Q(S)]\mathcal{P}_0(S)L(t_0(S), \mathcal{Y}) + \sum_{S \in \mathcal{S}_{\mathcal{P}_0} \setminus \mathcal{S}_{\mathcal{P}}} \mathcal{P}_0(S)L(t_0(S), \mathcal{Y}) \\
&= \sum_{S \in \mathcal{S}_{\mathcal{P}}} Q(S)\mathcal{P}_0(S)L(t_0(S), \mathcal{Y}) + \sum_{S \in \mathcal{S}_{\mathcal{P}}} [1 - Q(S)]\mathcal{P}_0(S)L(t_0(S), \mathcal{Y}) + \sum_{S \in \mathcal{S}_{\mathcal{P}_0} \setminus \mathcal{S}_{\mathcal{P}}} \mathcal{P}_0(S)L(t_0(S), \mathcal{Y}) \\
&= R(\mathcal{P}_0, t_0; \mathcal{Y})
\end{aligned}
$$

for all $\mathcal{Y}$, with a strict inequality for at least one $\mathcal{Y}$, where the inequality follows from the assumption that t_1 dominates t_0 under $\mathcal{P}$. Note that the estimator t^* is randomized. If $L(\tau, \mathcal{Y})$ is convex in τ, we can replace t^* by its expectation, and thus assume that t^* is nonrandomized. See Remark 4. In this case, Theorem 21 holds also for the class of nonrandomized estimators. □

For the next result, we slightly generalize Scott's (1975) formulation. Given a collection H of real valued functions defined on subsets of $\mathcal{N}$ and a corresponding collection of constants $C = \{c_h\}_{h \in H}$, define the class of designs

$$\mathcal{D}_{H,C} = \{\mathcal{P} : E_{\mathcal{P}} h(S) = c_h \text{ for all } h \in H\}.$$

For H consisting of the single function $h(S) = |S|$, and $c_h = n$, we obtain the class of designs with expected sample size $= n$. Taking H to be the set of indicator functions of all sets of size $\ne n$, and $c_h = 0$, we obtain the class of design having fixed sample size n. These two classes were considered by Scott.

Given an estimator $t_0(S, \mathcal{Y})$ and parameter $\theta = \theta(\mathcal{Y})$, the class of designs $\mathcal{P}$ under which t_0 is unbiased for θ is also of this kind. To see this define $h_{\mathcal{Y}}(S) = t_0(S, \mathcal{Y})$, and $c_h = c_{\mathcal{Y}} = \theta(\mathcal{Y})$, and set $H = \{h_{\mathcal{Y}} : \mathcal{Y} \in \Upsilon\}$.

The class of designs of having sample size n in some interval, the class of conditional SRS designs, see definition following Eq. (9), and other classes are of the above type, as are their intersections.

THEOREM 22. *If a strategy $(\mathcal{P}_0, t_0)$ is admissible in a class of strategies having designs in a given class $\mathcal{D} = \mathcal{D}_{H,C}$ and estimators in the class of all estimators including randomized ones, then the same holds for any strategy $(\mathcal{P}, t_0)$ such that $\mathcal{P} \in \mathcal{D}$, and $\mathcal{S}_{\mathcal{P}} \subseteq \mathcal{S}_{\mathcal{P}_0}$.*

PROOF. For $\mathcal{P}$ as above, suppose to the contrary that there exist a design $\mathcal{P}_1$ in $\mathcal{D}$, and an estimator t_1, such that the strategy $(\mathcal{P}_1, t_1)$ dominates $(\mathcal{P}, t_0)$. Set $\mathcal{P}^*(S) = \mathcal{P}_0(S) + m^{-1}(\mathcal{P}_1(S) - \mathcal{P}(S))$, and note that $\mathcal{P}^* \in \mathcal{D}$. Define $T(S) = \mathcal{P}_1(S)/m\mathcal{P}^*(S)$. Since $\mathcal{P}_0(S) - m^{-1}\mathcal{P}(S) \geq 0$, we have $\mathcal{P}^*(S) \geq m^{-1}\mathcal{P}_1(S)$, and therefore $0 \leq T(S) \leq 1$. Define the randomized estimator $t^*(S) = t_1(S)$ with probability $T(S)$ and $t^*(S) = t_0(S)$ with probability $1 - T(S)$. A calculation similar to the one in the proof of Theorem 21 shows that $R(\mathcal{P}^*, t^*; \mathcal{Y})$ dominates $R(\mathcal{P}_0, t_0; \mathcal{Y})$, a contradiction. □

Godambe and Joshi (1965) have shown that for any design, the Horvitz–Thompson estimator is admissible in the class of all unbiased estimators of a finite population total. The proof we give is essentially due to Ramakrishnan (1973), extended here from MSE to a more general convex loss function. The requirement $0 \in \Lambda$ is discussed after the proof.

THEOREM 23. *Let $\mathcal{P}$ be any design with $\alpha_i > 0$ for $i = 1, 2, \ldots, N$, and consider the parameter space Λ^N for some set Λ satisfying $0 \in \Lambda$. The Horvitz–Thompson estimator $t_{\mathrm{HT}}(S, \mathcal{Y}) = \sum_{i \in S} y_i/\alpha_i$ is admissible in the class of unbiased estimators for the parameter $\theta_N = \sum_{i=1}^N y_i$ provided that the loss function $L(t, \theta)$ is strictly convex in t and assumes its minimum, when $t = \theta$.*

PROOF. We assume $\mathcal{P}(S) > 0$ implies $|S| > 0$ to avoid trivialities. The proof is by induction on N. For $N = 1$ clearly $t_{\mathrm{HT}} = \theta_1$, and the result is obvious. The induction hypothesis is that for a population of size N, $R(\mathcal{P}, t; \mathcal{Y}) \leq R(\mathcal{P}, t_{\mathrm{HT}}; \mathcal{Y})$ for all $\mathcal{Y} \in \Lambda^N$ implies that $t = t_{\mathrm{HT}}$ with $\mathcal{P}$-probability 1, and it is easy to see that the desired admissibility follows. Let $(\mathcal{P}^*, t^*)$ be an unbiased strategy for θ_{N+1} on a population of size $N + 1$ denoted by U_{N+1}, and consider the population U_N of size N obtained by removing the last coordinate from U_{N+1}. On the latter population, we construct a strategy $(\mathcal{P}, t)$ by setting,

$$\begin{aligned}\mathcal{P}(S) &= \mathcal{P}^*(S) + \mathcal{P}^*(S, N+1), \quad t(S, \mathcal{Y}) \\ &= \frac{1}{\mathcal{P}(S)}[\mathcal{P}^*(S)t^*(S, \mathcal{Y}^*) + \mathcal{P}^*(S, N+1)t^*((S, N+1), \mathcal{Y}^*)]\end{aligned}$$

where, $(S, N+1) = S \cup \{N+1\}$, $\mathcal{Y} = (y_1, \ldots, y_N)$, and $\mathcal{Y}^* = (y_1, \ldots, y_N, 0)$. It is easy to see that $(\mathcal{P}, t)$ is unbiased for θ_N. For now let t_{HT} and t^*_{HT} denote the Horvitz–Thompson estimators for the designs $\mathcal{P}$ and $\mathcal{P}^*$, respectively. We claim that,

$$R(\mathcal{P}, t_{\mathrm{HT}}; \mathcal{Y}) = R(\mathcal{P}^*, t^*_{\mathrm{HT}}; \mathcal{Y}^*). \tag{27}$$

To see this, construct t_{HT} and t^*_{HT} on the same probability space (coupling) as follows. When a set $S \subset \{1, \ldots, N\}$ or $(S, N+1)$ is chosen with probability $\mathcal{P}^*$ as the sample

for t^*_{HT}, let S be the chosen set for t_{HT}. It then follows that $t_{\mathrm{HT}} = t^*_{\mathrm{HT}}$, and Eq. (27) follows. Next, we claim that,

$$R(\mathcal{P}, t; \mathcal{Y}) \leq R(\mathcal{P}^*, t^*; \mathcal{Y}^*). \tag{28}$$

It is easy to see that this follows by the convexity of L and Jensen's inequality, given that t is a convex combination of values of t^*. Moreover, the fact that L is strictly convex implies strict inequality in Eq. (28), whenever $t^*(S, \mathcal{Y}^*) \neq t^*((S, N+1), \mathcal{Y}^*)$, $\mathcal{P}^*(S) > 0$, and $\mathcal{P}^*(S, N+1) > 0$.

Assume that $R(\mathcal{P}^*, t^*; \mathcal{Y}^*) \leq R(\mathcal{P}^*, t^*_{\mathrm{HT}}; \mathcal{Y}^*)$ for all $\mathcal{Y}^* \in \Lambda^{N+1}$. Together with Eqs. (27) and (28) we then have,

$$R(\mathcal{P}, t; \mathcal{Y}) \leq R(\mathcal{P}, t_{\mathrm{HT}}; \mathcal{Y}) \text{ for all } \mathcal{Y} \in \Lambda^N \tag{29}$$

and so by the induction hypothesis,

$$t(S, \mathcal{Y}) = t_{\mathrm{HT}}(S, \mathcal{Y}) \tag{30}$$

for any S with $\mathcal{P}(S) > 0$. For such sets $S \subseteq \{1, \ldots, N\}$ we clearly have,

$$t_{\mathrm{HT}}(S, \mathcal{Y}^*) = t^*_{\mathrm{HT}}(S, \mathcal{Y}^*) \text{ for all } \mathcal{Y}^* \in \Lambda^{N+1}. \tag{31}$$

Moreover, strict inequality would hold in Eq. (29) for any $\mathcal{Y}$ such that $t^*(S, \mathcal{Y}^*) \neq t^*((S, N+1), \mathcal{Y}^*)$, $\mathcal{P}^*(S) > 0$, and $\mathcal{P}^*(S, N+1) > 0$ for $\mathcal{Y}^* = (y_1, \ldots, y_N, 0)$. But strict inequality is impossible since it would contradict the induction hypothesis, and therefore if $\mathcal{P}^*(S) > 0$ then either $\mathcal{P}^*(S, N+1) = 0$ or $t^*(S, \mathcal{Y}^*) = t^*((S, N+1), \mathcal{Y}^*)$. In either case, we then have $t(S, \mathcal{Y}) = t^*(S, \mathcal{Y}^*)$. This, together with Eqs. (30) and (31), implies that

$$t^*(S^*, \mathcal{Y}^*) = t^*_{\mathrm{HT}}(S^*, \mathcal{Y}^*) \text{ for all } \mathcal{Y}^* \in \Lambda^{N+1} \tag{32}$$

for any S^* not containing $N+1$ such that $\mathcal{P}^*(S) > 0$. We can repeat the argument with the label $N+1$ replaced by any j, and obtain Eq. (32) for any set S^* of size $\leq N+1$. Finally, Eq. (32) for the set $S^* = \{1, \ldots, N+1\}$ follows from this equality for all other sets S^* and from the fact that t^* and t^*_{HT} have the same expectation. This completes the induction step. □

The above result required $0 \in \Lambda$. Unlike in the case of Proposition 17, we cannot assume it "without loss of generality" by applying a linear transformation when $0 \notin \Lambda$. Indeed, if $\Lambda = \{a\}$ with $a \neq 0$, then the estimator $t = a$ is better than t_{TH} with respect to any design such that $\mathrm{Var} \sum_{i \in S} 1/\alpha_i > 0$. It is easy to construct less trivial examples. However, for $A = \{0, 1\}$ and using t_{HT}/N for estimating a proportion, the above admissibility result holds.

It is easy to construct examples with fixed or random sample size, where the Horvitz–Thompson estimator t_{HT}/N for a proportion is not in the interval $[0, 1]$ with positive probability (a trivial example is $n = 1$, $0 < \alpha_i < 1$ and $y_i \equiv 1$). In this case, it is clearly not admissible in the class of all estimators. This does not contradict Theorem 23, which requires unbiasedness (and allows designs with random sample size). For fixed-size sample designs, the Horvitz–Thompson estimator is admissible among all estimators when the parameter space is $\mathbb{R}^N$; see Joshi (1965, 1966). It follows that for SRS of any size n, the sample mean is an admissible estimator of the population mean. By Theorem 21, it follows that the sample mean is admissible for any fixed-size design.

The following example shows that if the sample size is random, t_{HT} may not be admissible in the class of all estimators: set $N = 2$ and $\mathcal{P}(\{1\}) = \mathcal{P}(\{1,2\}) = 1/2$. When the sample $\{1,2\}$ is selected, we have $t_{\mathrm{HT}} = y_1 + 2y_2$, and the (biased) estimator obtained by instead using $y_1 + y_2$ shows that t_{HT} is not admissible.

6. Superpopulation models

6.1. Background

In ***superpopulation models*** one assumes that the given population $\mathcal{Y} = (y_1, \ldots, y_N)$ is a realization of a random vector $Y = (Y_1, \ldots, Y_N)$ having a distribution $\mathcal{G}$. We shall refer to $\mathcal{G}$ as the ***prior***. Several possibilities arise: **1.** $\mathcal{G}$ is completely known. **2.** $\mathcal{G}$ belongs to a class having some known parameters and properties, for example, distributions with certain specified moments and possibly with some exchangeability properties. **3.** $\mathcal{G}$ depends on an unknown parameter ϕ, that is, $\mathcal{G} = \mathcal{G}_\phi$.

Design-based inference on the population $\mathcal{Y}$, as the name suggests, uses the sampling design (randomization distribution) only. Pure *model-based inference* on the population $\mathcal{Y}$, the prior $\mathcal{G}$, or the parameter ϕ, refers to inference where the sampling design plays no role, and the risk, for example, is defined as expectation with respect to $\mathcal{G}$ of the squared difference between the estimate and the estimand, conditioned on the sample.

A third approach combines the above two. Starting from the design-based risk $R(\mathcal{P}, t; Y)$, this approach studies the Bayes risk (see definition 3), that is, the expected risk with respect to $\mathcal{G}$, $E_{\mathcal{G}} R(\mathcal{P}, t; Y)$. The optimization goal is to find a strategy $(\mathcal{P}, t)$ that minimizes the latter expectation. For unbiased estimators and quadratic loss, this expectation becomes $\mathrm{E}_{\mathcal{G}} Var_{\mathcal{P}} t$, known in the sampling literature as the *anticipated variance*. It is often used to compare two design unbiased estimators when comparison of the $\mathcal{P}$-variances does not lead to clear conclusions.

It may happen that the superpopulation assumptions involve enough symmetry and randomness to make the sampling design inessential. For example, if Y is exchangeable under the superpopulation model and we use a symmetric estimator, then random sampling may be redundant since the data are assumed to be given in a random order.

We have already used the Bayesian approach, and, in fact, Eq. (22) can be seen as a prior of the above type; however, we used it only as a technical device to arrive at a minimax estimator, noting that the minimax criterion does not depend on the Bayesian structure.

We shall not discuss the philosophy and relevance of superpopulation models and model-based optimality criteria here. Some discussions and references can be found, for example, in Smith (1976), Särndal et al. (1992), Hedayat and Sinha (1991), and Cassel et al. (1977); the latter two books also contain a discussion that is closely related to the one that follows, with references and further results.

6.2. $\mathcal{P}$-unbiased estimators

In the discussion below, we consider $\mathcal{P}$-***unbiased estimators*** of the ***population mean*** $\bar{\mathcal{Y}}$. We shall consider ***quadratic loss*** and MSE, and the Bayes risk, which is the MSE integrated with respect to the prior $\mathcal{G}$. Theorem 30 shows that for any exchangeable prior (superpopulation model) the Bayes risk is minimized among $\mathcal{P}$-unbiased strategies by the strategy consisting of SRS (or any design with $\alpha_i = n/N$) and the sample mean. This

is generalized in Theorem 31 to the case of exchangeability of a linear transformation of the population values, and the optimal estimators are then the generalized difference estimators (see Eq. (26)) of which Horvitz–Thompson estimators form a special case. Note that these results involve $\mathcal{P}$-unbiasedness, which is a design-based criterion, and the Bayes risk, which is a model-based expectation over a design-based risk.

The results and techniques used next: sufficiency, completeness, and the Rao–Blackwell approach are close to those that led to Theorem 9. However, many detail are different. In particular, here the notions of sufficiency and completeness are with respect to the prior $\mathcal{G}$ rather than the design as in Section 3.2.

When we think of the population as fixed we denote it by $\mathcal{Y}$; when we want to emphasize that under the superpopulation model it is random, we denote it by Y. We used $\bar{\mathcal{Y}}$ and $\bar{y}_S$ for the population and sample means; we denote them by $\bar{Y}$ and $\bar{Y}_S$, when we want to emphasize that now the population is random, and when we take expectation with respect to $\mathcal{G}$. Given $Y = (Y_1, \ldots, Y_N)$ and a sample S, Y_S denotes the ***multiset*** containing all Y_i-values arising from distinct labels $i \in S$, in analogy to $\mathcal{Y}_S$ in the fixed population case. Similarly, we may express the data $D[S, \mathcal{Y}])$ as $D[S, Y])$, and when we want to describe an estimator, we may write $t(S, Y)$ instead of $t(S, \mathcal{Y})$, etc.

Note that we now have two sources of randomness, the sample $S \sim \mathcal{P}$ and the population $Y \sim \mathcal{G}$. Therefore, notations like $E_{\mathcal{P}}$, $E_{\mathcal{G}}$, and $E_{\mathcal{G},\mathcal{P}}$ for expectations will be used, where $E_{\mathcal{G},\mathcal{P}} = E_{\mathcal{G}} E_{\mathcal{P}}$. Unless otherwise stated, we consider designs that are ***noninformative*** or **ignorable**, that is $\mathcal{P}(S|Y) = \mathcal{P}(S)$, independent of Y. In words, the design does not depend on the population values Y. This assumption allows interchange of expectations with respect to $\mathcal{G}$ and $\mathcal{P}$. We discuss it further in Section 6.3.

Recall that the strategy $(\mathcal{P}, t)$ is ***unbiased*** for $\bar{\mathcal{Y}}$ (the population mean) if t is $\mathcal{P}$-***unbiased***, that is, if for all $\mathcal{Y} \in \Upsilon$, $E_{\mathcal{P}} t := \sum_S \mathcal{P}(S) t(D[S, \mathcal{Y}]) = \bar{\mathcal{Y}}$. Note that the latter expectation can also be interpreted as the conditional expectation $E\{t(D[S, Y]) \mid Y = \mathcal{Y}\}$. Recall also that for S satisfying $|S| = n$, $\bar{Y}_S = \frac{1}{n} \sum_{i \in S} Y_i$, that is, $\bar{Y}_S = \frac{1}{|S|} \sum_{i \in S} Y_i$.

Let $\mathbb{G}$ denote the class of ***exchangeable distributions***, that is, distributions that remain unchanged under permutations of the components of the vector Y.

LEMMA 24. *Let* $(\mathcal{P}, t = t(D[S, \mathcal{Y}]))$ *be an unbiased strategy for* $\bar{\mathcal{Y}}$. *Let* $Y = (Y_1, \ldots, Y_N) \sim \mathcal{G} \in \mathbb{G}$, *and* $E_{\mathcal{G}} Y_i = \mu_{\mathcal{G}}$. *Then,* $E_{\mathcal{G},\mathcal{P}} t(D[S, Y]) := E_{\mathcal{G}} \sum_S \mathcal{P}(S) t(S, Y) = \mu_{\mathcal{G}}$. *Also,* $E_{\mathcal{G},\mathcal{P}} \bar{Y}_S = \mu_{\mathcal{G}}$.

PROOF. The first part of the lemma is obvious. For the second part, note that for a general sampling design $\mathcal{P}$, $\bar{Y}_S$ is not necessarily $\mathcal{P}$-unbiased, so the first part does not imply the second. We have $\bar{Y}_S = \sum_{i=1}^N Y_i I_i / \sum_{j=1}^N I_j$, where $I_i = 1$ if $i \in S$ and 0 otherwise. Now $E_{\mathcal{G}}(\bar{Y}|S) = \mu_{\mathcal{G}} \sum_{i=1}^N I_i / \sum_{j=1}^N I_j = \mu_{\mathcal{G}}$, and the result follows. □

The next two easy lemmas show completeness and sufficiency. The classical Rao–Blackwell argument uses completeness and sufficiency as follows: given a statistic $t(X)$ which depends on some data $X \sim P_\theta$ (see Definition 3), and a sufficient statistic for θ, say $W(X)$, the estimator $t_0 = \mathrm{E}(t|W)$ is a statistic since it does not depend on θ by sufficiency. Also, t_0 has the same expectation as t but a smaller variance (by Jensen's inequality, or by a well-known variance decomposition formula). If W is complete, then t_0 is the unique estimator with the same expectation as t. This proves that it is a UMVU estimator of $\mathrm{E}t$ (see Definition 3). A version of this argument appears below, leading to Theorem 30.

LEMMA 25. *Let $Y = (Y_1, \ldots, Y_N) \sim \mathcal{G} \in \mathbb{G}$ and let $S \sim \mathcal{P}$. Consider the data $D = D[S, Y]$. Then Y_S is sufficient in the sense that $P(D|Y_S)$ does not depend on $\mathcal{G}$. (It does depend on $\mathcal{P}$, which is held fixed here.)*

PROOF. Just note that if $|S| = n$, then $P(D|Y_S) = \mathcal{P}(S)/n!$, where the $n!$ is due to the $n!$ equally likely (by exchangeability) ways of pairing the elements of S with those of Y_S. □

For the next lemma, we need two new conditions, which will henceforth be assumed. The first is that the ***parameter space is a product*** of the form $\Upsilon = \Lambda^N$, and the second is that the design $\mathcal{P}$ has a ***fixed sample size***, say n.

LEMMA 26. *Let $\mathbb{G}$ denote the class of exchangeable distributions over a product space Λ^N, and let $Y = (Y_1, \ldots, Y_N) \sim \mathcal{G} \in \mathbb{G}$, and $S \sim \mathcal{P}$, a given design with fixed sample size n. Let Y_S denotes the multiset containing all Y_i-values arising from distinct labels $i \in S$. Then Y_S is complete; that is, for any symmetric (permutation invariant) function h of n variables, if $E_{\mathcal{G},\mathcal{P}} h(Y_S) = 0$ for all $\mathcal{G} \in \mathbb{G}$ then $P_{\mathcal{G},\mathcal{P}}(h(Y_S) \neq 0) = 0$ for all $\mathcal{G} \in \mathbb{G}$.*

PROOF. The proof is similar to that of Lemma 8. For any $a \in \Lambda$, let $\mathcal{G}$ be the probability measure concentrated on $(a, \ldots, a) \in \Lambda^N$. Then clearly for this $\mathcal{G}$, $\mathrm{E}_{\mathcal{G},\mathcal{P}} h(Y_S) = 0$ implies $h(a, \ldots, a) = 0$. Now, let $\mathcal{G}$ be the exchangeable probability measure which concentrates on $(b, a, \ldots, a) \in \Lambda^N$ and all its permutations. This is used to prove $h(b, a, \ldots, a) = 0$ as in the proof of Lemma 8, and so on. □

As usual, completeness implies uniqueness of unbiased estimators, since if there existed two distinct unbiased estimators which are functions of Y_S, then their difference h would be a nonzero function whose expectation is zero, contradicting Lemma 26. The following example shows that a fixed sample size is, indeed, needed in Lemma 26. If the sample size is random with expectation n, then it is easy to see that the estimator $t = \frac{1}{n} \sum_{i \in S} Y_i$, where we divide by the expected sample size n rather than $|S|$, satisfies $E_{\mathcal{G},\mathcal{P}} t = \mu_{\mathcal{G}}$. The same holds for the estimator $\bar{Y}_S = \frac{1}{|S|} \sum_{i \in S} Y_i$ by Lemma 24, and unless the sample size is fixed, we have two distinct unbiased estimators of $\mu_{\mathcal{G}}$.

We shall now consider ***quadratic loss***. Then, $R(\mathcal{P}, t; Y) = \sum_S \mathcal{P}(S)(t(S, Y) - \bar{Y})^2 = Var_{\mathcal{P}}\, t$, and $\mathrm{E}_{\mathcal{G},\mathcal{P}}(t - \mu_{\mathcal{G}})^2 = E_{\mathcal{G}} \sum_S \mathcal{P}(S)(t(S, Y) - \mu_{\mathcal{G}})^2$, where the sum extends over all subsets S (of size n) of $\mathcal{N}$. The following lemma shows that for unbiased estimation of $\mu_{\mathcal{G}}$, the sample mean $\bar{Y}_S$ is optimal in the sense of minimizing $\mathrm{E}_{\mathcal{G},\mathcal{P}}(t - \mu_{\mathcal{G}})^2$. In fact, Lemma 27 holds for any convex loss function, but since quadratic loss is required in all the subsequent lemmas and theorems, we use quadratic loss also here.

LEMMA 27. *Let $\bar{Y}_S = \frac{1}{n} \sum_{i \in S} Y_i$, and let $t = t(D[S, Y])$ be an estimator satisfying $E_{\mathcal{G},\mathcal{P}} t(D[S, Y]) = \mu_{\mathcal{G}}$. Then,*

$$E_{\mathcal{G},\mathcal{P}}(\bar{Y}_S - \mu_{\mathcal{G}})^2 \leq E_{\mathcal{G},\mathcal{P}}(t - \mu_{\mathcal{G}})^2. \tag{33}$$

In particular this holds for any $\mathcal{P}$-unbiased estimator t of the population mean $\bar{\mathcal{Y}}$.

PROOF. The lemma follows by a standard Rao–Blackwell argument applied to the quadratic loss function, and using the facts that Y_S is sufficient and complete for the parameter $\mathcal{G}$ (see Lemma 26), and that $E_{\mathcal{G},\mathcal{P}}\bar{Y}_S = \mu_{\mathcal{G}}$. The latter equality and the last part about $\mathcal{P}$-unbiased estimators follow from Lemma 24. □

The next lemma is a standard variance decomposition. Recall the notation $\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i$, and the fact that for a $\mathcal{P}$-unbiased estimator of $\bar{\mathcal{Y}}$, we have $E_{\mathcal{P}}(t|Y) = \bar{Y}$.

LEMMA 28. *Let t be a $\mathcal{P}$-unbiased estimator of $\bar{\mathcal{Y}}$. Then, $\mathrm{Var}_{\mathcal{G},\mathcal{P}}\, t := E_{\mathcal{G},\mathcal{P}}(t-\mu_{\mathcal{G}})^2 = E_{\mathcal{G}}\mathrm{Var}_{\mathcal{P}}t + E_{\mathcal{G}}(\bar{Y}-\mu_{\mathcal{G}})^2 = E_{\mathcal{G}}\sum_S \mathcal{P}(S)(t(S,Y_S)-\bar{Y})^2 + E_{\mathcal{G}}(\bar{Y}-\mu_{\mathcal{G}})^2$.*

Lemmas 27 and 28 imply

LEMMA 29. *Let $\mathcal{P}$ be any design with fixed sample size n, $t = t(D[S,\mathcal{Y}])$ a $\mathcal{P}$-unbiased estimator of $\bar{\mathcal{Y}}$, and let $\mathcal{G}$ be any exchangeable (prior) distribution on the population $Y = (Y_1,\ldots,Y_N)$. Then,*

$$E_{\mathcal{G}}\mathrm{Var}_{\mathcal{P}}\bar{Y}_S = E_{\mathcal{G}}R(\mathcal{P},\bar{Y}_S;Y) \le E_{\mathcal{G}}\mathrm{Var}_{\mathcal{P}}t = E_{\mathcal{G}}R(\mathcal{P},t;Y).$$

The above result compares a $\mathcal{P}$-unbiased estimator t to the estimator $\bar{Y}_S$, which in general is not $\mathcal{P}$-unbiased. In fact, the strategy $(\mathcal{P},\bar{Y}_S)$ is unbiased if and only if $\alpha_i = n/N$. Note that $E_{\mathcal{G}}\mathrm{Var}_{\mathcal{P}}\bar{Y}_S = E_{\mathcal{G}}\sum_S \mathcal{P}(S)(\bar{Y}_S-\bar{Y})^2 = E_{\mathcal{G}}\sum_\pi \mathcal{P}(\pi S)(\bar{Y}_{\pi S}-\bar{Y})^2$ is constant as a function of $\mathcal{P}$ for all designs having sample size n, since by exchangeability $\bar{Y}_{\pi S}$ are identically distributed for all permutations π. Thus, we obtain the following theorem that compares the Bayes risk of unbiased strategies.

THEOREM 30. *Any strategy $(\mathcal{P}_0,\bar{Y}_S)$ with fixed sample size n, and $\alpha_i = n/N$, is optimal in the class of $\mathcal{P}$-unbiased (for the population mean) strategies $(\mathcal{P}, t = t(D[S,Y]))$ having sample size n, in the sense that for any $\mathcal{G}\in\mathbb{G}$,*

$$E_{\mathcal{G}}R(\mathcal{P}_0,\bar{Y}_S;Y) \le E_{\mathcal{G}}R(\mathcal{P},t;Y). \tag{34}$$

The above result can be generalized as follows. Suppose that we have reason to believe that our units are not exchangeable. For example, they may have different known average sizes a_i, that is, $\mu_i := E_{\mathcal{G}}Y_i = a_i$ and, more generally, $\mu_i = E_{\mathcal{G}}Y_i = a_i\mu + b_i$ and perhaps also $E_{\mathcal{G}}(Y_i-\mu_i)^2 = a_i^2\sigma^2$. The known constants, a_i, b_i, can be viewed as auxiliary information. This leads to Theorem 31 below, in which we assume that the variables $(Y_1-b_1)/a_1,\ldots,(Y_N-b_N)/a_N$ have an exchangeable prior (superpopulation model) with known constants $a_i > 0$ and b_i. We set $\sum_{i=1}^{N} a_i = N$ without loss of generality.

THEOREM 31. *Let $((Y_1-b_1)/a_1,\ldots,(Y_N-b_N)/a_N) \sim \mathcal{G}\in\mathbb{G}$, and*

$$t_{\mathrm{GD}_0} = \sum_{i\in S}\frac{Y_i-b_i}{\alpha_i} + \tilde{b}\,,\ \textit{where}\ \ \tilde{b} = \sum_{i=1}^{N} b_i.$$

Let $\mathcal{P}_0$ be any design having a fixed sample size n, and $\alpha_i = a_i n/N$. The strategy $(\mathcal{P}_0, \frac{1}{N}t_{\mathrm{GD}_0})$ is optimal in the class of $\mathcal{P}$-unbiased (for the population mean) strategies $(\mathcal{P}, t = t(D[S, Y_S]))$ having sample size n, in the sense that

$$E_{\mathcal{G}}R(\mathcal{P}_0, \tfrac{1}{N}t_{\mathrm{GD}_0}; Y) \le E_{\mathcal{G}}R(\mathcal{P}, t; Y). \tag{35}$$

PROOF. Define $Z_i = \frac{(Y_i - b_i)}{a_i} + \bar{b}$, where $\bar{b} = \frac{1}{N}\sum_{i=1}^{N} b_i$, and $Z = (Z_1, \ldots, Z_N)$. Then $\bar{Z}_S := \frac{1}{n}\sum_{i\in S} Z_i = \frac{1}{N}t_{\mathrm{GD}_0}$. The proof is the same as that of Theorem 30, applied to the above Z. □

Theorem 31 is due to Cassel et al. (1977). The special case of $b_i = 0$ shows that Horvitz–Thompson strategies, that is, any strategy $(\mathcal{P}_0, \frac{1}{N}t_{\mathrm{HT}})$ with $\alpha_i = a_i n/N$ and the corresponding estimate $\frac{1}{N}t_{\mathrm{HT}} = \frac{1}{N}\sum_{i\in S} y_i/\alpha_i$, have a minimal Bayes risk among $\mathcal{P}$-unbiased (for the population mean) strategies for priors such that the vector $(Y_1/a_1, \ldots, Y_N/a_N)$ is exchangeable. In this case, the expectations EY_i are proportional to some known constants, and any design of fixed sample size n with inclusion probabilities that are proportional to those constants and a corresponding Horvitz–Thompson estimator form an optimal strategy with respect to Bayes risk.

6.3. *Linear prediction*

We consider estimation of the population mean on the basis of a sample from the random population Y, where $Y \sim \mathcal{G}$, to be specified later. Under such a superpopulation model, the population mean $\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i$ is a random variable, which we are trying to predict. From the relation $\bar{Y} = \frac{n}{N}\bar{Y}_S + (1 - \frac{n}{N})\bar{Y}_{S^c}$, we see that when $D[S, Y_S]$ is observed, our task is to predict $\bar{Y}_{S^c} = \frac{1}{N-n}\sum_{i\in S^c} Y_i$, where $n = |S|$ and S^c denotes the complement of S. We consider momentarily the possibility that the design depends on the population values, in which case the ***design*** is said to be ***informative***, and write $\mathcal{P}(S|Y)$ for the probability of sampling S given that the population vector is Y. Let $g(Y)$, $y \in \mathbb{R}^N$ denote a density of the prior $\mathcal{G}$ (which may depend on a parameter). Then the ***predictive density*** of Y_{S^c}, the unobserved part of Y given the data, is

$$f(y_{S^c}|S, y_S) = \mathcal{P}(S|y_S, y_{S^c})g(y_S, y_{S^c}) \Big/ \int \mathcal{P}(S|y_S, y_{S^c})g(y_S, y_{S^c})dy_{S^c}.$$

The design is ***noninformative*** or **ignorable** if $\mathcal{P}(S|Y) = \mathcal{P}(S)$, independent of Y. ***Adaptive designs*** satisfy $\mathcal{P}(S|Y) = \mathcal{P}(S|Y_S)$; that is, the sample may depend on the observed y-values, but not on the unobserved ones. Such designs arise when the sample is selected sequentially: first some units are sampled, and the choice of the units that are added to the sample depends on the y's already observed; see Thompson and Seber (1996) and references therein. In either the ignorable or the adaptive case, we obtain

$$f(y_{S^c}|S, y_S) = g(y_S, y_{S^c}) \Big/ \int g(y_S, y_{S^c})dy_{S^c},$$

and we see that the design does not play a role in the predictive density. It is, therefore, not surprising that the predictive (model-based, Bayes) optimality result of Theorem 32 below is not stated in terms of a sampling design.

We need some definitions. A statistic $t = t(S, Y)$ is said to be a ***linear predictor*** (or estimator) if it is of the form $t(S, Y_S) = \sum_{i\in S} r_i Y_i + q$, where the constants r_i, q may depend on S.

Let s be a given subset of $\mathcal{N}$. Given a statistic $t = t(S, Y_S)$ we can fix the set s, and consider the random variable $t(s, Y_s)$ for the given fixed set s and $Y \sim \mathcal{G}$. The estimator t is said to be a $\mathcal{G}$-***unbiased predictor*** of the population mean $\bar{Y}$ if $E_{\mathcal{G}}(t(s, Y_s) - \bar{Y}) = 0$ for ***every*** fixed $s \subset \mathcal{N}$. See, for example, Cassel et al. (1977). $\mathcal{G}$-unbiased predictors of other parameters are defined similarly.

The above definition and the theorem below are written in terms of a fixed set (or nonrandom sample) s and expectations in the form $\mathrm{E}_{\mathcal{G}}[\ \cdot\ ]$, taken with respect to $Y \sim \mathcal{G}$. An equivalent formulation would be to consider a random S and replace these expectations by $\mathrm{E}_{\mathcal{G}}[\ \cdot\ |S = s]$ provided that Y and S are independent, which means that the design $\mathcal{P}$ is ignorable. If $E_{\mathcal{G}}[(t(s, Y) - \bar{Y})|S = s] = 0$ for an ignorable design $\mathcal{P}$, then we can now take expectation with respect to $S \sim \mathcal{P}$, to obtain $E_{\mathcal{P},\mathcal{G},}(t - \bar{Y}) = 0$. Exchanging the order of the expectations, it is easy to see that a $\mathcal{G}$-unbiased predictor t satisfies $E_{\mathcal{G},\mathcal{P}}(t - \bar{Y}) = 0$ for any *ignorable* design $\mathcal{P}$. These operations cannot be done if $\mathcal{P}$ is informative, that is, if it depends on Y.

On the other hand, for any design $\mathcal{P}$ (ignorable or not), a $\mathcal{P}$-unbiased estimate t of $\bar{Y}$ satisfies $E_{\mathcal{G},\mathcal{P}}(t - \bar{Y}) = 0$ for any $\mathcal{G}$.

Theorem 32 below, which is one of many results on optimality in the class of $\mathcal{G}$- unbiased predictors, appears in Hedayat and Sinha (1991) with further references. A closely related result appears in Royall (1970b). The auxiliary variables x_i, b_i below are assumed to be known constants.

THEOREM 32. *Let $Y \sim \mathcal{G} \in \mathbb{G}_L$, where $\mathbb{G}_L$ is a family of distributions such that $Z_i = \frac{(Y_i - b_i)}{x_i}$ satisfy $E_{\mathcal{G}} Z_i = \mu$, $Var_{\mathcal{G}} Z_i = \sigma^2$, and $Corr_{\mathcal{G}}(Z_i, Z_j) = \rho$ for all $i \neq j$ for some (unknown) $(\mu, \sigma, \rho) \in \Theta$, a parameter space which contains at least two distinct values of μ, and let x_i, b_i be known. Let $s \subset \mathcal{N}$ be fixed and $|s| = n$, and consider the linear predictor*

$$t^* = t^*(s, Y_s) = \frac{n}{N}\bar{Y}_s + \left(1 - \frac{n}{N}\right)\left(\bar{Z}_s\, \bar{x}_{s^c} + \bar{b}_{s^c}\right),$$

where, $\bar{Y}_s = \frac{1}{n}\sum_{i\in s} Y_i$, $\bar{Z}_s = \frac{1}{n}\sum_{i\in s} Z_i$, s^c is the complement of s in $\mathcal{N}$, $\bar{x}_{s^c} = \frac{1}{N-n}\sum_{i\in s^c} x_i$, and $\bar{b}_{s^c} = \frac{1}{N-n}\sum_{i\in s^c} b_i$. Then, t^* is $\mathcal{G}$-unbiased for any $\mathcal{G} \in \mathbb{G}_L$, and

$$E_{\mathcal{G}}(t^*(s, Y_s) - \bar{Y})^2 \leq E_{\mathcal{G}}(t(s, Y_s) - \bar{Y})^2$$

for any linear predictor t of $\bar{Y}$ that is $\mathcal{G}$-unbiased for all $\mathcal{G} \in \mathbb{G}_L$.

THEOREM 33. *Under the conditions of Theorem 32, let now S be a random sample satisfying $S \sim \mathcal{P}$, where $\mathcal{P}$ is any ignorable design. Then,*

$$E_{\mathcal{G},\mathcal{P}}(t^*(S, Y_S) - \bar{Y})^2 \leq E_{\mathcal{G},\mathcal{P}}(t(S, Y_S) - \bar{Y})^2$$

for any linear predictor t of the population mean $\bar{Y}$, that is $\mathcal{G}$-unbiased for all $\mathcal{G} \in \mathbb{G}_L$.

PROOF THEOREM 33. This follows from the inequality of Theorem 32 by taking $\mathcal{P}$ expectation and exchanging the order of expectations as explained above for ignorable designs. □

PROOF THEOREM 32. The proof is almost the same as in Hedayat and Sinha (1991). We can express any linear predictor in the form,

$$t(s, Y_s) = \frac{n}{N}\bar{Y}_s + \Big(1 - \frac{n}{N}\Big)\hat{t}(s, Y_s), \quad \text{where } \hat{t}(s, Y_s) = \sum_{i\in s} c_i Y_i + d.$$

We have $\bar{Y} = \frac{n}{N}\bar{Y}_s + (1-\frac{n}{N})\bar{Y}_{s^c}$, where $\bar{Y}_{s^c} = \frac{1}{N-n}\sum_{i\in s^c} Y_i$, and therefore t is $\mathcal{G}$-unbiased if and only if $E_{\mathcal{G}}(\hat{t} - \bar{Y}_{s^c}) = 0$, which is equivalent to $\sum_{i\in s} c_i(x_i\mu + b_i) + d = \bar{x}_{s^c}\mu + \bar{b}_{s^c}$. The latter equality holds for two distinct values of μ if and only if,

$$\sum_{i\in s} c_i b_i + d = \bar{b}_{s^c}, \quad \text{and} \quad \sum_{i\in s} c_i x_i = \bar{x}_{s^c}. \tag{36}$$

It suffices to minimize

$$E_{\mathcal{G}}(\hat{t} - \bar{Y}_{s^c})^2 = \mathrm{Var}_{\mathcal{G}}(\hat{t} - \bar{Y}_{s^c}) = \mathrm{Var}_{\mathcal{G}}\hat{t} + \mathrm{Var}_{\mathcal{G}}\bar{Y}_{s^c} - 2\,\mathrm{Cov}_{\mathcal{G}}(\hat{t},\ \bar{Y}_{s^c}).$$

Using Eq. (36), it is easy to calculate that $Cov_{\mathcal{G}}(t,\ \bar{Y}_{s^c}) = \rho\sigma^2\bar{x}_{s^c}^2$. Therefore, the above minimization is achieved by finding $\hat{t}$ satisfying Eq. (36), and having a minimal variance. A straightforward expansion of the variance and Eq. (36) lead to

$$\mathrm{Var}_{\mathcal{G}}\hat{t} = \sigma^2\Big[\,\rho\Big(\sum_{i\in s} c_i x_i\Big)^2 + (1-\rho)\sum_{i\in s} c_i^2 x_i^2\,\Big] = \sigma^2\Big[\,\rho\bar{x}_{s^c}^2 + (1-\rho)\sum_{i\in s} c_i^2 x_i^2\,\Big].$$

We can now use the Lagrange method to minimize $\sum_{i\in s} c_i^2 x_i^2$ subject to the constraint $\sum_{i\in s} c_i x_i = \bar{x}_{s^c}$ from Eq. (36). We readily obtain the solution $c_i = \bar{x}_{s^c}/nx_i$. From Eq. (36), we can now obtain d, and putting it all together with some simple calculations, the result follows. □

It is now possible to write an explicit expression of $E_{\mathcal{G}}(t^*(s, Y_s) - \bar{Y})^2$ for any set s, and minimize over s of a given size, thus obtaining an efficient purposive (nonrandom) sample. Such considerations led Royall (1970b) to advocate purposive rather than random sample selection. This approach, and the concept of $\mathcal{G}$-unbiasedness depend on the superpopulation model, unlike man-made randomness and $\mathcal{P}$-unbiasedness, where the statistician controls the randomization procedure. The efficiency of purposive designs constructed in the above manner is sensitive to the choice of the prior or superpopulation model and, therefore, robustness issues arise; see, for example, Scott et al. (1978), Hansen et al. (1983), and references therein. See also Valliant et al. (2000), Mukhopadhyay (1998), and Chaudhuri and Stenger (1992) for further discussion and references on the issues arising here, and in other parts of this chapter.

7. Beyond simple random sampling

We have so far concentrated on relatively simple models, and for many results (but not all) on simple sampling designs, with emphasis on (conditional) simple random sampling. We now discuss a few examples of results on various well-known sampling designs, and more general models. Only parts of the results are proved, and other parts are explained or stated without a proof. Here, as in the whole chapter, the results given

constitute a sample and certainly not a survey. In all examples below, only quadratic loss and the corresponding MSE are considered.

7.1. pps cluster sampling

Related results to Proposition 12, but for **cluster sampling**, with clusters of different sizes and when the estimated parameter is a weighted average (by cluster size) of the cluster means, were given by Scott and Smith (1975), and Scott (1977). They consider ***Bernoulli sampling***, that is, sampling n clusters with replacement, where unit i is drawn with probability p_i in each draw, and in particular the case of ***probability proportional to size (pps) sampling***, where p_i are proportional to cluster size. They show that under certain conditions, the pps strategy minimizes $\sup_{\mathcal{Y}}$MSE for the pps- Horvitz–Thompson estimator in the class of Bernoulli designs with expected sample size n. When the conditions are relaxed, approximate minimaxity is derived.

7.2. Approximate minimax and the Rao–Hartley–Cochran strategy

We now describe results of Cheng and Li (1983, 1987) which extend the results of Section 7.1. Further references can be found in these papers. Consider a population $\mathcal{Y} = (y_1, \ldots, y_N)$ satisfying $y_i = \theta x_i + \varepsilon_i$, $i = 1, \ldots, N$, where $\varepsilon_i = \delta_i g(x_i)$ are nonrandom errors, the x_i's and g are known, and $\delta = (\delta_1, \ldots, \delta_N)$ belongs to some known set L, and $\theta \in \Theta$, some suitable parameter space, is an unknown nuisance parameter.

Given a sample S, a ***linear estimator*** is of the form $t(S, Y) = \sum_{i\in S} r_{si}\, y_i$, that is, a linear combination of the observations with weights that may depend on S. Let r_s^t (r_s) denotes the row (column) vector $r_s^t = (r_{s1}, \ldots, r_{sN})$, where for $i \notin S$ we set $r_{si} = 0$. For $\mathbf{R} = \{r_s : S \in 2^N\}$ we set $t_{\mathbf{R}}(S, Y) = \sum_{i\in S} r_{si}\, y_i = r_s^t Y$. A strategy consists of a pair $(\mathcal{P},\ t_{\mathbf{R}}(S, Y))$. Our goal is to estimate the population mean $\bar{\mathcal{Y}} = \sum_{i=1}^N y_i/N$, using the auxiliary information, and we look for a strategy that minimizes (approximately) the ***risk***

$$\sup_{\theta\in\Theta,\delta\in L} \sum_S \mathcal{P}(S)\Big(\sum_{i\in S} r_{si}\, y_i - \bar{\mathcal{Y}}\Big)^2.$$

Set $\mathbf{x}^t = (x_1, \ldots, x_N)$, $\mathbf{1}$ an N-vector of 1's, $\bar{X} = \sum_{i=1}^N x_i/N$, and let G be the $N \times N$ diagonal matrix $G = diag(g(x_1), \ldots, g(x_N))$. We have

$$\begin{aligned}&\sum_S \mathcal{P}(S)\Big(\sum_{i\in S} r_{si}\, y_i - \sum_{i=1}^N y_i/N\Big)^2\\ &\quad = \sum_S \mathcal{P}(S)(\theta\mathbf{x} + G\delta)^t(r_s - \mathbf{1}/N)(r_s - \mathbf{1}/N)^t(\theta\mathbf{x} + G\delta),\end{aligned}$$

and it is easy to see that if Θ is unbounded then the risk is bounded if and only if $(r_s - \mathbf{1}/N)^t\mathbf{x} = 0$. If we restrict our choice to such r_s's, then we guarantee that $\sum_{i\in S} r_{si}\, x_i = \bar{X}$, so that the linear coefficients are **calibrated** for $\bar{X}$. Such a strategy is called ***representative***.

In order to describe one of the results from Cheng and Li (1983), we now define the Rao–Hartley–Cochran (RHC) strategy. In order to obtain a sample of size n, divide the population at random (all partition having equal probabilities) into n groups of

predetermined sizes N_j such that $\sum_{j=1}^{n} N_j = N$. Let X_j be the sum of the x_i's in group j. Draw one element from each group, so that if the ith unit is in the jth group, it is drawn with probability x_i/X_j, and denote its y-value by y_j and its x-value by x_j. The RHC estimator for the population mean $\mathcal{Y}$ is then $t_{\mathrm{RHC}} = \sum_{j=1}^{n} X_j\, y_j/Nx_j$.

Note that this ($\mathcal{P}$-unbiased) estimator is not of the type $t(S, \mathcal{Y})$ considered in this chapter, which are functions of the sampled set S and the corresponding y-values. The quantities X_j are random since they depend on the random partition, with distribution depending on the data $D[S, \mathcal{Y}]$ (actually, on S). Hence it is a **randomized estimator**.[6] With convex loss, we could replace X_j by $E(X_j|S)$ and by Jensen's inequality the risk is reduced (see Remark 4). This calculation is usually complex and, therefore, it is avoided.

For suitable L and g, and under assumptions relating n, N, and the x_i's which require that n largest x_i's are not too large, Cheng and Li (1983) show that the risk of the RHC strategy is bounded by $1 + \varepsilon$ times the maximal risk $\sup_{\theta\in\Theta,\delta\in L} \sum_S \mathcal{P}(S)(\sum_{i\in S} r_{si}\, y_i - \bar{\mathcal{Y}})^2$, where an explicit bound on ε in terms of the x_i's is given. Thus, the RHC strategy is approximate minimax. The details will not be given here.

Cheng and Li (1987) show interesting relation between models as above and **superpopulation** models where the ε_i's are random variables. Such a superpopulation model is considered in Section 7.3.

7.3. Minimax linear esimation in a superpopulation model

Our next discussion concerns a **superpopulation** model that is closely related to the one given in Theorem 32. The results stated here are from Stenger (2002).

Consider a population $Y = (Y_1, \ldots, Y_N) \sim \mathcal{G}$ (see the notation in Section 6.1) generated according to the superpopulation model $Y_i = \theta x_i + \varepsilon_i$, where ε_i are **random variables** satisfying $\mathrm{E}\varepsilon_i = 0$, $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = \gamma u_{ij}$, the x_i's are known, $i, j \in \mathcal{N}$, γ is unknown and the u_{ij}'s are discussed below. Our goal is to estimate the parameter θ on the basis of a sample of size n. Writing $Z_i := Y_i/x_i = \theta + \varepsilon_i/x_i$, we see the similarity to the model of Theorem 32, where now we allow a more general covariance structure.

Parts of the discussion that follows are similar to that of Section 7.2; however, here we are dealing with a superpopulation model. Given a sample S, a ***linear estimator*** is of the form $t(S, Y) = \sum_{i\in S} r_{si} Y_i$, that is, a linear combination of the observations with weights that may depend on S. For $\mathbf{R} = \{r_s : S \in 2^N\}$ we set $t_{\mathbf{R}}(S, Y) = \sum_{i\in S} r_{si} Y_i = r_s^t \mathbf{Y}$, where r_s^t denotes the row vector $(r_{s1}, \ldots, r_{sN})$, and for $i \notin S$ we set $r_{si} = 0$.

Let U denote the $N \times N$ matrix with entries u_{ij}. We assume that $U \in B$, some (known) class of positive definite matrices. For S fixed, the usual MSE decomposition to variance and bias squared yields

$$E_{\mathcal{G}}\Big(\sum_{i\in S} r_{si} Y_i - \theta\Big)^2 = \gamma r_s^t U r_s + \theta^2\Big(\sum_{i\in S} r_{si} x_i - 1\Big). \tag{37}$$

A strategy consists of a pair $(\mathcal{P}, t_{\mathbf{R}}(S, Y))$. We have

$$E_{\mathcal{G},\mathcal{P}}\Big(\sum_{i\in S} r_{si} Y_i - \theta\Big)^2 = \sum_S \mathcal{P}(S) E_{\mathcal{G}}\Big(\sum_{i\in S} r_{si} Y_i - \theta\Big)^2,$$

[6] Since the probabilities determining the randomization depend on S, t_{RHC} is a *behavioral estimator*.

and here we define a minimax strategy as the strategy $(\mathcal{P}, t_{\mathbf{R}})$ minimizing

$$\sup_{\theta\in\mathbb{R}, U\in B} E_{\mathcal{G},\mathcal{P}}\Big(\sum_{i\in S} r_{si}Y_i - \theta\Big)^2. \tag{38}$$

In view of Eq. (37), the expression under the sup is arbitrarily large for large values of θ, unless we set $\sum_{i\in S} r_{si}x_i - 1 = 0$ for all S in the support of $\mathcal{P}$. This condition is equivalent to $\mathrm{E}_{\mathcal{G}}\sum_{i\in S} r_{si}Y_i = \theta$ for such S, and hence the same holds for $\mathrm{E}_{\mathcal{G},\mathcal{P}}$, and our estimators are unbiased (see Section 6.3). If U is known, that is, $|B| = 1$, the problem reduces to finding a set S that minimizes $\inf_{r_s}\{r_s^t U r_s : \sum_{i\in S} r_{si}x_i - 1 = 0\}$, and we conclude that the minimax strategy is degenerate, concentrated at a minimizing set S. Thus for this problem, random sampling is not required. Clearly the minimizing S depends on the x_i's and U. Degenerate designs are called *purposive sampling*.

Next consider $|B| > 1$, and suppose that the covariance matrix U is in the set of diagonal matrices $B = \{W = diag(w_1, \ldots, w_N) : w_i > 0\ \forall i,\ \sum \beta_i w_i \le 1\}$ for some given $\beta_1, \ldots, \beta_N > 0$. Since the matrices in B are diagonal, the ε_i's are uncorrelated. Assume $\alpha_i := n\beta_i x_i^2 / \sum_{i=1}^N \beta_i x_i^2 \le 1$ for all $i \in \mathcal{N}$. Stenger's (2002) result is

THEOREM 34. *A minimax strategy $(\mathcal{P}_0, t_{\mathbf{R}_0}(S, Y))$, minimizing Eq. (38) among strategies consisting of size n designs and linear estimators, is given by any size n design $\mathcal{P}_0$ having inclusion probabilities $\alpha_i = n\beta_i x_i^2 / \sum_{i=1}^N \beta_i x_i^2$, and $t_{\mathbf{R}_0} = \frac{1}{n}\sum_{i\in S} Y_i/x_i$.*

PROOF. By Eq. (37) and the discussion following Eq. (38), the problem of finding the strategy $(\mathcal{P}, t_{\mathbf{R}})$ minimizing Eq. (38), that is, the minimax strategy, is equivalent to minimizing

$$\sup_{W\in B} \sum_S r_s^t W r_s \mathcal{P}(S) \ \text{ subject to } r_s^t x - 1 = 0. \tag{39}$$

For a given S, W, and $x = (x_1, \ldots, x_N)$, let $\hat\theta_S(W)$ be the linear estimator $\sum_{i\in S} r_{si}Y_i$ derived by minimizing $r_s^t W r_s$ subject to $r_s^t x - 1 = \sum_{i\in S} r_{si}x_i - 1 = 0$. Using Lagrange multipliers, we obtain $r_{si} = \frac{x_i/w_i}{\sum_{i\in S} x_i^2/w_i}$ for $i \in S$, and therefore $\hat\theta_S(W) = \frac{\sum_{i\in S} x_i Y_i/w_i}{\sum_{i\in S} x_i^2/w_i}$. It is easy to see that for any $\mathcal{P}$ the same vectors r_s also minimize $\sum_S r_s^t W r_s \mathcal{P}(S)$ subject to the condition $r_s^t x - 1 = 0$ holding for all S.

By compactness, the sup in Eq. (39) is attained at some $V = diag(v_1, \ldots, v_N) \in B$, and therefore the vectors r_s minimizing Eq. (39) must satisfy $r_{si} = r_{si}(V) = \frac{x_i/v_i}{\sum_{i\in S} x_i^2/v_i}$ for $i \in S$. Note that for this $r_s = r_s(V)$ we have $\gamma r_s^t W r_s = \mathrm{Var}_W \hat\theta_S(V)$, where Var_W is the variance with respect to the model $\mathcal{G}$, when W is the true covariance matrix. It follows that finding the minimax strategy is equivalent to finding a design $\mathcal{P}$ and $V \in B$ minimizing

$$\sup_{W\in B} \sum_S \mathrm{Var}_W \hat\theta_S(V)\mathcal{P}(S). \tag{40}$$

Let $V_0 = diag(v_1, \ldots, v_N)$ where $v_i = x_i^2 / \sum_{i=1}^N \beta_i x_i^2$. Then for $|S| = n$, $\hat\theta_S(V_0) = \frac{1}{n}\sum_{i\in S} Y_i/x_i$, $\mathrm{Var}_U \hat\theta_S(V_0) = \frac{\gamma}{n^2}\sum_{i\in S} u_{ii}/x_i^2$, and $\mathrm{Var}_{V_0}\hat\theta_S(V_0) = \frac{\gamma}{n}\big(1/\sum_{i=1}^N \beta_i x_i^2\big)$, which is independent of S; this will turn out to be useful in Eq. (41) below.

Let $\mathcal{P}_0$ be any design with inclusion probabilities $\alpha_i = n\beta_i x_i^2 / \sum_{i=1}^N \beta_i x_i^2$. See Chaudhuri and Vos (1988, Part B) for a survey of methods for construction of such designs. Since $\alpha_i = \sum_{S: S \ni i} \mathcal{P}(S)$, we have for all $V, U \in B$ and any design $\mathcal{P}$

$$\begin{aligned}\sum_S \mathrm{Var}_U \hat{\theta}_S(V_0)\mathcal{P}_0(S) &= \frac{\gamma}{n^2}\sum_{i=1}^N \alpha_i u_{ii}/x_i^2 = \frac{\gamma}{n}\sum_{i=1}^N \beta_i u_{ii} \Big/ \sum_{i=1}^N \beta_i x_i^2 \\ &\le \frac{\gamma}{n}\Big(1/\sum_{i=1}^N \beta_i x_i^2\Big) = \mathrm{Var}_{V_0}\hat{\theta}_S(V_0) = \sum_S \mathrm{Var}_{V_0}\hat{\theta}_S(V_0)\mathcal{P}(S) \qquad (41)\\ &\le \sum_S \mathrm{Var}_{V_0}\hat{\theta}_S(V)\mathcal{P}(S) \le \sup_{W\in B}\sum_S \mathrm{Var}_W \hat{\theta}_S(V)\mathcal{P}(S),\end{aligned}$$

where the first inequality holds because $U \in B$, and the second because $V = V_0$ minimizes $\mathrm{Var}_{V_0}\hat{\theta}_S(V)$ by definition of $\hat{\theta}_S(W)$. It follows that for any $V \in B$ and any design $\mathcal{P}$

$$\sup_{W\in B}\sum_S \mathrm{Var}_W \hat{\theta}_S(V_0)\mathcal{P}_0(S) \le \sup_{W\in B}\sum_S \mathrm{Var}_W \hat{\theta}_S(V)\mathcal{P}(S), \qquad (42)$$

and the strategy $(\mathcal{P}_0, \mathbf{r}(V_0))$ minimizes the expression in Eq. (40), and hence it is the minimax strategy in the sense defined in Eq. (38). □

Unlike the result of Theorem 32, which concerns estimation or prediction of the population mean $\bar{Y}$, the problem of estimating the regression parameter θ discusses above and the sample selection based on the x_i's may be viewed as belonging to the area of *optimal regression design* rather than sampling. Note in particular that even if the whole population $(Y_1, \ldots, Y_N)$ is observed, the parameter θ is not determined. Stenger (2002) discusses also the problem of predicting $\bar{Y}$, under the same regression model, and proves existence of minimax strategies. Again, purposive sampling suffices when $|B| = 1$, and random sampling is required for $|B| > 1$.

Returning to the problem of estimating the population mean under a **superpopulation** model, we now discuss the seminal work of Aggarwal (1959, 1966). The population $Y = (Y_1, \ldots, Y_N)$ is distributed according to $\mathcal{G} \in \mathbb{H}$, where $\mathbb{H}$ is the class of distributions $\mathcal{G}$ that are concentrated on a hyperplane in $\mathbb{R}^N$ of the form $Y_1 + \ldots Y_N =$ constant, say $N\mu_{\mathcal{G}}$, and subject to

$$E_{\mathcal{G}}\sum_{i=1}^N (Y_i - \mu_{\mathcal{G}})^2 \le M \qquad (43)$$

for some $M > 0$. Setting $\mathrm{E}_{\mathcal{G}} Y_i = \mu_{\mathcal{G},i}$, and $\mathrm{Var}_{\mathcal{G}} Y_i = \sigma^2_{\mathcal{G},i}$, we can express that latter condition as $\sum_{i=1}^N [\sigma^2_{\mathcal{G},i} + (\mu_{\mathcal{G},i} - \mu_{\mathcal{G}})^2] \le M$. The goal is to estimate the population mean $\bar{Y}$, which here equals $\mu_{\mathcal{G}}$. Under $\mathcal{P}_s$, simple random sampling of n observations, consider the problem of finding the minimax estimator, that is, the estimator t minimizing the risk $\sup_{\mathcal{G}\in\mathbb{H}} E_{\mathcal{P}_s,\mathcal{G}}(t(S, Y) - \bar{Y})^2$. Aggarwal (1959) uses Bayesian calculations (see Section 3.5) to show that the minimax estimator is the sample mean.

If the population is divided into given **strata**, and simple random sampling with a given sample size is carried out in each stratum, and if in each stratum a superpopulation model of the above kind holds (with the bound M_i instead of M of Eq. (43) for the ith

stratum), then the usual weighted sum of the strata means is shown to be minimax. For a statistician who can choose the sample sizes, and the cost of sampling is added to the above risk, Aggarwal (1959) provides the minimax strategy,[7] consisting of the same weighted mean, and where naturally the sample sizes in the strata depend on the bounds M_i, and the cost of sampling in each stratum.

Aggarwal (1966) provides similar results for ***two-stage sampling***. Now the population is divided into given subgroups called ***primary units*** (or **clusters**). A simple random sample of primary units (clusters) is selected in the first stage (whereas in stratified sampling all strata are sampled), and a second-stage simple random sampling is carried out in each of the selected clusters. The superpopulation model constrains the cluster means to be on a hyperplane, as well as the Y's within each cluster, with conditions similar to Eq. (43) within and between clusters, and suitable bounds replacing M. With weights computed in terms of these bounds and the sample sizes, a weighted average of the sample means in the sampled clusters is shown to be minimax for given sample sizes. A minimax allocations of sample sizes that depends on the bounds and the sampling costs is also given, which together with the above estimator comprise a minimax strategy.[7]

8. List of main notations

$\mathcal{Y} = (y_1, \ldots, y_N)$, a finite population of size N. $\mathcal{N} = \{1, \ldots, N\}$. $S \subseteq \mathcal{N}$, a sample.

$\mathcal{Y}_S$ – the multiset containing all y_i-values arising from distinct labels $i \in S$.

$Y = (Y_1, \ldots, Y_N)$, a random finite population of size N under a superpopulation model.

$\mathcal{G}$ – distribution of Y (prior or superpopulation model). $\mathbb{G}$ – class of exchangeable priors.

Y_S – the multiset containing all Y_i-values arising from distinct labels $i \in S$.

$\bar{\mathcal{Y}} = \frac{1}{N}\sum_{i=1}^{N} y_i$, the population mean. $\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i$, the population mean under a superpopulation model.

$\bar{\mathcal{Y}}_S = \frac{1}{n}\sum_{i\in S} y_i$, the sample mean, where $n = |S|$. $\bar{Y}_S = \frac{1}{n}\sum_{i\in S} Y_i$, the sample mean under a superpopulation model.

$D = D[S, \mathcal{Y}] = \{(i, y_i) : i \in S\}$, the data.

$t = t(D) = t(\{(i, y_i) : i \in S\})$ – an estimator. We also use $t(D[S, \mathcal{Y}])$, or $t(S, \mathcal{Y})$. Under a superpopulation model, we use $t(\{(i, Y_i) : i \in S\})$ or $t(D[S, Y])$, etc.

$\mathcal{P}$ – a sampling design (probability over subsets S of $\mathcal{N}$). $\alpha_i = \mathcal{P}(\{i \in S\})$, inclusion probabilities.

$\mathrm{SRS} = \mathcal{P}_s$ – simple random sampling without replacement, also SRS.

$t_{\mathrm{HT}} = \sum_{i\in S} y_i/\alpha_i$ – the Horvitz–Thompson estimator.

[7] Among strategies based on simple random sampling.

$L(t, \mathcal{Y})$ – loss when the estimator takes the value t.

$R(\mathcal{P}, t; \mathcal{Y}) := E_{\mathcal{P}} L(t, \mathcal{Y}) = \sum_S \mathcal{P}(S) L(t(D[S, \mathcal{Y}_S]), \mathcal{Y})$ – risk.

$\text{MSE}\,(\mathcal{P}, t; \mathcal{Y}) = E_{\mathcal{P}}(t - \theta)^2$

Acknowledgements

I am grateful to Larry Goldstein, Yakov Malinovsky, Gad Nathan, and Ya'acov Ritov for many illuminating discussions of the subject matter of this chapter. Micha Mandel read parts of the chapter while it was being written and I am indebted to him in many ways, and J. N. K. Rao and Alistair Scott read the first draft. They all made invaluable suggestions, raised important questions, and contributed new ideas. Their corrections definitely reduced the number of errors in the manuscript.

References

Abowd, J.M., Zellner, A. (1985). Estimating gross labor force flows. *Journal of Business and Economic Statistics* **3**, 254–283.

Adam, A., Fuller, W.A. (1992). Covariance estimators for the current population survey. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, Alexandria, VA, 586–591.

Aerts, M., Claeskens, G., Wand, M. (2002). Some theory for penalized spline generalized additive models. *Journal of Statistical Planning and Inference* **103**, 455–470.

Aggarwal, Om.P. (1959). Bayes and minimax procedures in sampling from finite and infinite populations-I. *The Annals of Mathematical Statistics* **30**, 206–218.

Aggarwal, Om.P. (1966). Bayes and minimax procedures for estimating the arithmetic mean of a population with two-stage sampling. *The Annals of Mathematical Statistics* **37**, 1186–1195.

Agresti, A. (2002). *Categorical Data Analysis* (2nd ed.). John Wiley, New York.

Aires, N. (1999). Algorithms to find exact inclusion probabilities for conditional Poisson sampling and Pareto πps sampling designs. *Methodology and Computing in Applied Probability* **1**, 457–469.

Aires, N. (2000). Comparisons between conditional Poisson sampling and Pareto πps sampling designs. *Journal of Statistical Planning and Inference* **88**, 133–147.

Aires, N., Rosén, B. (2005). On inclusion probabilities and relative estimator for Pareto πps sampling. *Journal of Statistical Planning and Inference* **128**, 543–567.

Albert, J.H. (1988). Computational methods using a Bayesian hierarchical generalized linear model. *Journal of the American Statistical Association* **83**, 1037–1044.

Alexander, C.H. (1987). A model-based justification for survey weights. In: *Proceedings of the Section on Survey Research Methods*. American Statistical Association, pp. 183–188.

Alexander, C.H. (2002). Still rolling: Leslie Kish's "rolling samples" and the American community survey. *Survey Methodology* **28**, 39–46.

Anderson, J.A. (1972). Separate sample logistic discrimination. *Biometrika* **59**, 19–35.

Andersson, C., Nordberg, L. (1998). *A User's Guide to CLAN97*. Statistics Sweden, Örebro, Sweden.

Andersson, P.G., Thorburn, D. (2005). An optimal calibration distance leading to the optimal regression estimator. *Survey Methodology* **31**, 95–99.

Anscombe, F.J. (1952). Large sample theory of sequential estimation. *Proceedings of the Cambridge Philosophical Society* **48**, 600–607.

Aragon, Y., Goga, C., Ruiz-Gazen, A. (2006). Estimation non-paramétrique de quantiles en présence d'information auxiliaire. In: Lavellée, P., Rivest, L.-P. (Eds.), *Méthodes d'Enquêtes et Sondages. Pratiques Européenne et Nord-américaine*. Dunod, Paris-Sciences Sup, pp. 377–382.

Arora, V., Lahiri, P. (1997). On the superiority of the Bayesian method over the BLUP in small area estimation problems. *Statistica Sinica* **7**, 1053–1063.

Arratia, R., Goldstein, L., Langholz, B. (2005). Local central limit theorems, the high-order correlations of rejective sampling and logistic likelihood asymptotics. *The Annals of Statistics* **33**, 871–914.

Asok, C., Sukhatme, B.V. (1976). On Sampford's procedure of unequal probability sampling without replacement. *Journal of the American Statistical Association* **71**, 912–918.

Asparouhov, T. (2004). Weighting for unequal probability of selection in multilevel models. MPlus Web Notes, No. 8. Available at: http://www.statmodel.com.

Asparouhov, T. (2006). General multilevel modeling with sampling weights. *Communications in Statistics. Theory and Methods* **35**, 439–460.

Atkinson, A.C. (1985). *Plots, Transformations, and Regression*. Clarendon Press, Oxford.

Australian Bureau of Statistics. (1993). *A Guide to Interpreting Time Series – Monitoring "Trends", An Overview*". Catalogue no. 1348.0, Australian Bureau of Statistics, Canberra, Australia.

Australian Bureau of Statistics. (2007). *Forthcoming Changes to Labour Force Statistics*. Catalogue no. 6292.0, Australian Bureau of Statistics, Canberra, Australia.

Bahadur, R.R. (1966). A note on quantiles in large samples. *The Annals of Mathematical Statistics* **37**, 388–393.

Bailar, B.A. (1975). The effects of rotation group bias on estimates from panel surveys. *Journal of the American Statistical Association* **70**, 23–30.

Bailar, B.A. (1978). Rotation sample biases and their effects on estimates of change. In: Krishan Namboodini, N. (Ed.), *Survey Sampling and Measurement*. Academic Press, New York, pp. 385–407.

Bailey, T.J. (1951). On estimating the size of mobile populations from recapture data. *Biometrika* **38**, 293–306.

Baker, M., McNicholas, A., Garrett, N., Jones, N., Stewart, J., Koberstein, V., Lennon, D. (2000). Household crowding: a major risk factor for epidemic meningo-coccal disease in Cityplace Auckland children. *Pediatric Infectious Disease Journal* **19**, 983–990.

Bankier, M. (2002). Regression estimators for the 2001 Canadian Census. *Presented at the International Conference in Recent Advances in Survey Sampling*.

Basu, D. (1958). On sampling with and without replacement. *Sankhya A* **20**, 287–294.

Basu, D. (1969). Role of sufficiency and likelihood principles in sample survey theory. *Sankhya B* **31**, 441–454.

Basu, D. (1971). An essay on the logical foundations of survey sampling, part 1. In: Godambe, V.P., Sprott, D.A. (Eds.), *Foundations of Statistical Inference*. Holt, Rinehart and Winston, Toronto, Canada, pp. 203–242.

Basu, D. (1980). Randomization analysis of experimental data: the Fisher randomization test (with comments and rejoinder). *Journal of the American Statistical Association* **75**, 575–595.

Battese, G.E., Harter, R.M., Fuller, W.A. (1988). An error components model for prediction of county crop area using survey and satellite data. *Journal of the American Statistical Association* **83**, 28–36.

Beaumont, J.-F. (2005). Calibrated imputation in surveys under a quasi-model-assisted approach. *Journal of the Royal Statistical Society, Series B* **67**, 445–458.

Beaumont, J.-F. (2008). A new approach to weighting and inference in sample surveys. *Biometrika* **95**, 539–553.

Beaumont, J.-F., Alavi, A. (2004). Robust generalized regression estimation. *Survey Methodology* **30**, 195–208.

Bellhouse, D.R., Stafford, J.E. (1999). Density estimation from complex surveys. *Statistica Sinica* **9**, 407–424.

Bellhouse, D.R., Stafford, J.E. (2001). Local polynomial regression techniques in complex surveys. *Survey Methodology* **27**, 197–203.

Bellhouse, D.R., Stafford, J.E. (2003). Graphical displays of complex survey data through kernel smoothing. In: Chambers, R.L., Skinner, C.J. (Eds.), *Analysis of Survey Data*. Wiley, Chichester, West Sussex, pp. 133–150.

Bell, P. (2001). Comparison of alternative labour force survey estimators. *Survey Methodology* **27**, 53–63.

Bell, P.A. (1999). *The Impact of Sample Rotation Patterns and Composite Estimation on Survey Outcomes.* Working Paper. Catalogue no. 1351.0, no99/1, Australian Bureau of Statistics, Canberra, Australia, May 1999.

Belsley, D.A., Kuh, E., Welsch, R.E. (1980). *Regression Diagnostics*, Wiley, New York.

Bell, W. (2001). Discussion with "jackknife in the Fay-Herriott model with an example". *Proceeding of the Seminar on Funding Opportunity in Survey Research*, 98–104.

Bell, W.R. (2004). On RegComponent time series models and thier applications. In: Harvey, A.C., Koopman, S.J., Shephard, N. (Eds.), *State Space and Unobserved Component Models: Theory and Application*. Cambridge University Press, Cambridge, UK, pp. 248–283.

Bell, W.R. (2005). Some consideration of seasonal adjustment variances. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, Alexandria, VA, pp. 2747–2758.

Bell, W.R., Hillmer, S.C. (1990). The time series approach to estimation for repeated surveys. *Survey Methodology* **16**, 195–215.

Bell, W.R., Kramer, M. (1999). Toward variances for X-11 seasonal adjustments. *Survey Methodology* **25**, 13–29.

Bell, W.R., Wilcox, D.W. (1993). The effect of sampling error on the time series behavior of consumption data. *Journal of Econometrics* **55**, 235–265.

Bera, A.K., Bilias, Y., Simlai, P. (2006). Estimating functions and equations: an essay on historical developments with applications to econometrics. *Palmgrave Handbook of Econometrics, Volume 1*.

Berger, J.O. (1985). *Statistical Decision Theory and Bayesian Analysis* (2nd ed.). Springer-Verlag, New York.

Berger, Y.G. (1998a). Rate of convergence to normal distribution for the Horvitz-Thompson estimator. *Journal of Statistical Planning and Inference* **67**, 209–226.

Berger, Y.G. (1998b). Rate of convergence for asymptotic variance of the Horvitz-Thompson estimator. *Journal of Statistical Planning and Inference* **74**, 149–168.

Berger, Y.G. (2004). A simple variance estimator for unequal probability sampling without replacement. *Journal of Applied Statistics* **31**, 305–315.

Berger, Y.G. (2005a). Variance estimation with Chao's sampling scheme. *Journal of Statistical Planning and Inference* **127**, 253–277.

Berger, Y.G. (2005b). Variance estimation with highly stratified sampling designs with unequal probabilities. *Australian & New Zealand Journal of Statistics* **47**, 365–373.

Berger, Y.G. (2007). A jackknife variance estimator for unistage stratified samples with unequal probabilities. *Biometrika* **94**, 953–964.

Berger, Y.G., Skinner, C.J. (2005). A jackknife variance estimator for unequal probability sampling. *Journal of the Royal Statistical Society, Series B* **67**, 79–89.

Bickel, P.J., Freedman, D.A. (1984). Asymptotic normality and the bootstrap in stratified sampling. *The Annals of Statistics* **12**(2), 470–482.

Bickford, C.A., Mayer, C.E., Ware, K.D. (1963). An efficient sampling design for forest inventory: the Northeast Forest Resurvey. *Journal of Forestry* **61**, 826–833.

Bickel, P.J., Lehmann, E.L. (1981). A minimax property of the sample mean in finite populations. *The Annals of Statistics* **9**, 1119–1122.

Bieler, G.S., Williams, R.L. (1995). Cluster sampling techniques in quantal response teratology and developmental toxicity studies. *Biometrics* **51**, 764–776.

Binder, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review* **51**, 279–292.

Binder, D.A. (1992). Fitting Cox's proportional hazards models from survey data. *Biometrika* **79**, 139–147.

Binder, D.A. (1996). Linearization methods for single phase and two phase samples: a cookbook approach. *Survey Methodology* **22**, 17–22.

Binder, D.A., Dick, J.P. (1989a). Modelling and estimation for repeated surveys. *Survey Methodology* **15**, 29–45.

Binder, D.A., Dick, J.P. (1989b). Analysis of seasonal ARIMA models from survey data. In: Sing, A.C., Withridge, P. (Eds.), *Analysis of Data in Time, Proceedings of the 1989 International Symposium.* Statistics Canada, pp. 57–65.

Binder, D.A., Hidiroglou, M.A. (1988). Sampling in time. In: Krishnaiah, P.R., Rao, C.R. (Ed.), *Handbook of Statistics Volume 6: Sampling*. Elsevier/North-Holland, New York; Amsterdam, pp. 187–211.

Binder, D.A., Kovačević, M., Roberts, G. (2004). Design-based methods for survey data: Alternative uses of estimating functions. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, Alexandria, VA, pp. 3301–3312.

Binder, D.A., Kovačević, M.S., Roberts, G. (2005). How important is the informativeness of the sample design? *Proceedings of the Survey Methods Section*, Statistical Society of Canada, Ottawa, pp. 1–11.

Binder, D.A., Patak, Z. (1994). Use of estimating functions for estimation from complex surveys. *Journal of the American Statistical Association* **89**, 1035–1043.

Binder, D.A., Roberts, G.R. (2001). Can informative designs be ignorable? *Newsletter of the Survey Research Methods Section*, American Statistical Association, Issue 12.

Binder, D.A., Roberts, G.R. (2003). Design-based and model-based methods for estimating model parameters. In: Chambers, R.L., Skinner, C.J. (Eds.), *Analysis of Survey Data*. Wiley, Chichester, UK, pp. 29–48.

Bishop, Y.M.M., Fienberg, S.E., Holland, P.W. (1975). *Discrete Multivariate Analysis: Theory and Practice.* Cambridge, MA, MIT Press.

Blackwell, D., Girshick, M.A. (1954). *Theory of Games and Statistical Decisions*. Wiley, New York.

Blight, B.J.N., Scott, A.J. (1973). A stochastic model for repeated surveys. *Journal of the Royal Statistical Society, Series B: Methodological* **35**, 61–66.

Bloznelis, M. (2003). A note on the bias and consistency of the jackknife variance estimator in stratified samples. *Statistics* **37**, 489–504.

Bloznelis, M. (2007). Second-order and resampling approximation of finite population U-statistics based on stratified samples. *Statistics* **41**, 321–332.

Bolfarine, H., Zacks, S. (1992). *Prediction Theory for Finite Populations*. Springer-Verlag, New York.

Boos, D.D. (1992). On generalized score tests. *The American Statistician* **4**, 327–333.

Booth, J.G., Butler, R.W., Hall, P. (1994). Bootstrap methods for finite populations. *Journal of the American Statistical Association* **89**, 1282–1289.

Booth, J.G., Hobert, J.P. (1998). Standard errors of predictors in generalized linear mixed models. *Journal of the American Statistical Association* **93**, 262–272.

Brackstone, G.J. (1987). Small area data: policy issues and technical challenges. In: Platek, R., Rao, J.N.K., Sarndal, C.E., Singh, M.P. (Eds.), *Small Area Statistics*. Wiley, New York, pp. 3–20.

Breckling, J.U., Chambers, R.L., Dorfman, A.H., Tam, S.M., Welsh, A.H. (1994). Maximum likelihood inference from sample survey data. *International Statistical Review* **62**, 349–363.

Breidt, F.J., Claeskens, G., Opsomer, J.D. (2005). Model-assisted estimation for complex surveys using penalised splines. *Biometrika* **92**(4), 831–846.

Breidt, F.J., Opsomer, J.D. (2000). Local polynomial regression estimators in survey sampling. *The Annals of Statistics* **28**, 1026–1053.

Breidt, F.J., Opsomer, J.D., Johnson, A.A., Ranalli, M.G. (2007). Semiparametric model-assisted estimation for natural resource surveys. *Survey Methodology* **33**, 35–44.

Breiman, L., Friedman, J.H., Olshen, R., Stone, C.J. (1984). *Classification and Regression Trees*. Wadsworth, Belmont, CA.

Breslow, N.E. (1996). Statistics in epidemiology: the case-control study. *Journal of the American Statistical Association* **91**, 14–28.

Breslow, N.E. (2005). Case-control studies. In: Aherns, W., Pigeot, I. (Eds.), *Handbook of Epidemiology*. Springer, New York, pp. 287–319.

Breslow, N.E., Cain, K.C. (1988). Logistic regression for two-stage case-control data. *Biometrika* **75**, 11–20.

Breslow, N.E., Chatterjee, N. (1999). Design and analysis of two-phase studies with binary outcome applied to Wilms tumour prognosis. *Applied Statistics* **48**(4), 457–468.

Breslow, N.E., Clayton, D.G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association* **88**, 9–25.

Breslow, N.E., Day, N.E. (1980). *The Analysis of Case-Control Studies*. International Agency for Research on Cancer, Lyon, France.

Breslow, N.E., Holubkov, R. (1997). Maximum likelihood estimation of logistic regression parameters for two-phase outcome-dependent sampling. *Journal of the Royal Statistical Society, Series B* **59**, 447–461.

Breslow, N.E., McNeney, B., Wellner, J.A. (2003). Large sample theory for semiparametric regression models with two-phase, outcome dependent sampling. *The Annals of Statistics* **31**, 1110–1139.

Breslow, N.E., Robins, J.M., Wellner, J.A. (2000). On the semi-parametric efficiency of logistic regression under case-control sampling. *Bernoulli* **6**, 447–455.

Breslow, N.E., Zhao, L.P. (1988). Logistic regression for stratified case-control studies. *Biometrics* **44**, 891–899.

Breu, P., Ernst, L. (1983). Alternative estimators to the current composite estimators. *Proceedings of the Survey Research Methods Section*, American Statistical Association, Alexandria, VA, pp. 397–402.

Breunig, R.V. (2001). Density estimation for clustered data. *Econometric Reviews* **20**(3), 353–367.

Brewer, K.R.W. (1963). Ratio estimation and finite populations: some results deductible from the assumption of an underlying stochastic process. *Australian Journal of Statistics* **5**, 93–105.

Brewer, K.R.W. (1994). Survey sampling inference: some past perspectives and present prospects. *Pakistan Journal of Statistics* **10**(1)**A**, 213–233.

Brewer, K.R.W. (1995). Combining design-based and model-based inference. In: Cox, B.G., Binder, D.A., Chinappa, B.N., Christianson, A., Colledge, M.J., Kott, P.S. (Eds.), *Business Survey Methods*. Wiley, New York, pp. 589–606.

Brewer, K.R.W. (1999). Cosmetic calibration with unequal probability sampling. *Survey Methodology* **25**, 205–212.

Brewer, K.R.W., Donadio, M.E. (2003). The high entropy variance of the Horvitz-Thompson estimator. *Survey Methodology* **29**, 189–196.

Brewer, K.R.W., Hanif, M. (1983). *Sampling with Unequal Probabilities*. Lecture Notes in Statistics, 15. Springer, New York.

Brewer, K.R.W., Mellor, R.W. (1973). The effect of sample structure on analytical surveys. *The Australian Journal of Statistics* **15**, 145–152.

Brewer, K.R.W., Sarndal, C.E. (1983). Six approaches to enumerative survey sampling. In: Madow, W.G., Olkin, I. (Eds.), *Incomplete Data in Sample Survey*. Academic Press, pp. 363–368.

Brogan, D.J., Denniston, M.M., Liff, J.M., Flagg, E.W., Coates, R.J., Brinton, L.A. (2001). Comparison of telephone sampling and area sampling: response rates and within-household coverage. *American Journal of Epidemiology* **153**, 1119–1127.

Brumback, B., Greenland, S., Redman, M., Kiviat, N., Diehr, P. (2003). The intensity-score approach to adjusting for confounding. *Biometrics* **59**, 274–285.

Brunsdon, T.M., Smith, T.M.F. (1998). The time series analysis of compositional data. *Journal of Official Statistics* **14**, 237–253.

Burman, J.P. (1980). Seasonal adjustment by signal extraction. *Journal of the Royal Statistical Society, Series A: General* **143**, 321–337.

Burns, R.M. (1990). Multiple and replicate item imputation in a complex sample survey. *Proceedings of the Sixth Annual Research Conference*, U.S. Burea of the Census, Washington, DC, pp. 655–665.

Burridge, P., Wallis, K.F. (1985). Calculating the variance of seasonally adjusted series. *Journal of the American Statistical Association* **80**, 541–552.

Buskirk, T.D. (1998). Nonparametric density estimation using complex survey data. *ASA Proceedings of the Section on Survey Research Methods*, American Statistical Association, Alexandria, VA, pp. 799–801.

Buskirk, T.D. (1999). *Using nonparametric methods for density estimation with complex survey data*. Ph.D. Thesis, Department of Mathematics, Arizona State University, Tempe, AZ.

Buskirk, T.D., Lohr, S.L. (2005). Asymptotic properties of kernel density estimation with complex survey data. *Journal of Statistical Planning and Inference* **128**, 165–190.

Butar, B.F. (1997). Empirical Bayes methods in survey sampling. Unpublished Ph.D. Thesis, Department of Mathematics and Statistics, University of Nebraska-Lincoln, Lincoln, NE.

Butar, B.F., Lahiri, P. (2003). On measures of uncertainty of empirical Bayes small-area estimators. *Journal of Statistical Planning and Inference* **112**, 63–76.

Butler, N.R., Deotisou, S., Shepherd, P. (1997). *1970 British Cohort Study (BCS70) Ten-year follow-up: A guide to the BCS70 10-year Data available at the Economic and Social Research Council Data Archive*. Social Statistics Research Unit, City University, London.

Cain, K.C., Breslow, N.E. (1988). Logistic regression analysis and efficient design for two-stage studies. *American Journal of Epidemiology* **128**, 1198–1206.

Cantwell, P.J. (1990). Variance formulae for composite estimators in rotation designs. *Survey Methodology* **16**, 153–163.

Cantwell, P.J., Caldwell, C.V. (1998). Examining the revisions in rotation panel design. *Journal of Official Statistics* **14**, 47–54.

Cantwell, P.J., Ernst, L.R. (1993). Short-term changes to the CPS composite estimator in January 1994. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, Alexandria, VA, pp. 724–729.

Canty, A.J., Davison, A.C. (1999). Resampling-based variance estimation for labour force surveys. *The Statistician* **48**, 379–391.

Caplan, D., Haworth, M., Steel, D.G. (1999). UK labour market statistics: combining continuous survey data into monthly reports. *Bulletin of the International Statistical Institute,* 52nd Session, contributed papers, Book 2, 25–26.

Carlin, B.P., Louis, T.A. (1996). *Bayes and Empirical Bayes Methods for Data Analysis*. Chapman and Hall, New York.

Carlin, B.P., Louis, T.A. (2000). *Bayes and Empirical Bayes Methods for Data Analysis* (2nd ed.). Chapman and Hall, New York.

Caron, N., Sautory, O. (2004). Calages simultanés pour différentes unités d'une même enquête. Document de travail Méthodologie statistique n° 0403, INSEE.

Carroll, R.J., Ruppert, D. (1988). *Transformation and Weighting in Regression*. Chapman and Hall, New York.

Cassel, C.M., Särndal, C.E., Wretman, J.H. (1977). *Foundations of Inference in Survey Sampling*. Wiley, New York.

Ceppellini, R., Siniscalco, M., Smith, C.A.B. (1955). The estimation of gene frequencies in a random mating population. *Annals of Human Genetics* **20**, 97–115.

Chambers, J.M., Cleveland, W.S., Kleiner, B., Tukey, P.A. (1983). *Graphical Methods for Data Analysis*. Wadsworth International Group, Belmont, CA.

Chambers, R.L. (1986). Design-adjusted parameter estimation. *Journal of the Royal Statistical Society, Series A* **149**, 161–173.

Chambers, R.L. (2003). Introduction to part A. In: Chambers, R.L., Skinner, C.J. (Eds.), *Analysis of Survey Data*. John Wiley & Sons, Chichester, UK, pp. 13–28.

Chambers, R.L., Dorfman, A.H., Hall, P. (1992). Properties of estimators of the finite population distribution function. *Biometrika* **79**(3), 577–582.

Chambers, R.L., Dorfman, A.H., Sverchkov, M.Y. (2003). Nonparametric regression with complex survey data. In: Chambers, R.L., Skinner, C.J. (Eds.), *Analysis of Survey Data*. John Wiley & Sons, Chichester, UK, pp. 151–174.

Chambers, R.L., Dorfman, A.H., Wang, S. (1998). Limited information likelihood analysis of survey data. *Journal of the Royal Statistical Society, Series B* **60**, 397–411.

Chambers, R.L., Dorfman, A.H., Wehrly, T.E. (1993). Bias robust estimation in finite populations using nonparametric calibration. *Journal of the American Statistical Association* **88**(421), 268–277.

Chambers, R.L., Dunstan, R. (1986). Estimating distribution functions from survey data. *Biometrika* **73**, 597–604.

Chambers, R.L., Skinner, C.J. (2003). *Analysis of Survey Data*. John Wiley & Sons, Chichester, UK.

Chambles, L.E., Boyle, K.E. (1985). Maximum likelihood methods for complex sample data: logistic regression and discrete proportional hazards models. *Communications in Statistics-Theory and Methods* **14**, 1377–1392.

Chandola, T., Clarke, P., Morris, J.N., Blane, D. (2006). Pathways between education and health: a causal modelling approach. *Journal of the Royal Statistical Society* **A169**, 337–359.

Chang, T., Kott, P.S. (2008). Using calibration weighting to adjust for nonresponse under a plausible model. *Biometrika* **95**, 555–571. Available at: http://www.nass.usda.gov/research/reports/cal_paper_rev3.pdf.

Chao, M.T. (1982). A general purpose unequal probability sampling plan. *Biometrika* **69**, 653–656.

Chao, M.T., Lo, S.H. (1985). A bootstrap method for finite populations. *Sankhya A* **47**, 399–405.

Chao, C.-T., Thompson, S.K. (2001). Optimal adaptive selection of sampling sites. *Environmetrics* **12**, 517–538.

Chapman, D.G. (1951). Some properties of hypergeometric distribution with applications to zoological sample census. *University of California Publications in Statistics* **1**, 131–160.

Chapman, D.G. (1952). Inverse, multiple and sequential sample censuses. *Biometrics* **8**, 286–306.

Chatterjee, S., Lahiri, P. (2007). A simple computational method for estimating mean squared prediction error in general small-area model, *Proceedings of the Section on Survey Research Methods*, American Statistical Association, Alexandria VA, pp. 3486–3493.

Chatterjee, S., Lahiri, P., Li, H. (2008). On small area prediction interval problems. *The Annals of Statistics* **36**, 1221–1245.

Chattopadhyay, M., Lahiri, P., Larsen, M., Reimnitz, J. (1999). Composite estimation of drug prevalances for sub-state areas. *Survey Methodology* **25**, 81–86.

Chaudhary, M.A., Sen, P.K. (1998). Rescaling bootstrap inference from complex surveys. *Pakistan Journal of Statistics* **14**, 149–159.

Chaudhary, M.A., Sen, P.K. (2002). Reconciliation of asymptotics for unequal probability sampling without replacement. *Journal of Statistical Planning and Inference* **102**, 71–81.

Chaudhuri, A., Stenger, H. (1992). *Survey Sampling: Theory and Methods*. Marcel Dakker Inc, New York.

Chaudhuri, A., Vos, J. (1988). *Unified Theory and Strategies of Survey Sampling*. North-Holland, Amsterdam.

Chauvet, G., Tillé, Y. (2006). A fast algorithm of balanced sampling. *Computational Statistics* **21**, 53–62.

Cheng, C.S., Li, K.C. (1983). A minimax approach to sample surveys. *The Annals of Statistics* **11**, 552–563.

Cheng, C.S., Li, K.C. (1987). Optimality criteria in survey sampling. *Biometrika* **74**, 337–345.

Chen, J., Chen, S.Y., Rao, J.N.K. (2003). Empirical likelihood confidence intervals for the mean of a population containing many zero values. *The Canadian Journal of Statistics* **31**, 53–68.

Chen, J., Qin, J. (1993). Empirical likelihood estimation for finite population and the effective usage of auxiliary information. *Biometrika* **80**, 107–116.

Chen, J., Rao, J.N.K. (2007). Asymptotic normality under two-phase sampling designs. *Statistica Sinica* **17**, 1047–1064.

Chen, J., Sitter, R.R. (1993). Edgeworth expansion and the bootstrap for stratified sampling without replacement from a finite population. *The Canadian Journal of Statistics* **21**, 347–357.

Chen, J., Sitter, R.R. (1999). A pseudo empirical likelihood approach to the effective use of auxiliary information in complex surveys. *Statistica Sinica* **9**, 385–406.

Chen, J., Sitter, R.R., Wu, C. (2002). Using empirical likelihood methods to obtain range restricted weights in regression estimators for surveys. *Biometrika* **89**, 230–237.

Chen, J., Wu, C. (2002). Estimation of distribution function and quantiles using the model-calibrated pseudo empirical likelihood method. *Statistica Sinica* **12**, 1223–1239.

Chen, P., Penne, M.A., Singh, A.C. (2000). Experience with the general exponential model for weight calibration for the National Household Survey on Drug Abuse. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, Washington, DC, pp. 604–609.

Chen, S., Lahiri, P. (2003). A comparison of different MSPE estimators of EBLUP for the Fay-Herriot model. *Proceedings of the Survey Research Methods Section of the ASA*.

Chen, S., Lahiri, P. (2008). On mean squared prediction error estimation in small area estimation problems. *Communications in Statistics – Theory and Methods* **37**(11), 1792–1798.

Chen, S., Lahiri, P., Rao, J.N.K. (2007). Robust mean squared prediction error estimators of EBLUP of a small area total under the Fay-Herriot model. *Proceedings of the Statistics Canada Symposium*.

Chen, S.X. (2000). General properties and estimation of conditional Bernoulli models. *Journal of Multivariate Analysis* **74**, 67–87.

Chen, S.X., Dempster, A.P., Liu, J.S. (1994). Weighted finite population sampling to maximize entropy. *Biometrika* **81**, 457–469.

Chen, Z.-G., Wong, P., Morry, M., Fung, H. (2003). *Variance Estimation for X-11 Seasonal Adjustment Procedure: Spectrum Approach and Comparison.* Working paper, No. BSMD-2003-001E, Statistics Canada Ottawa, Canada.

Chernoff, H., Lehmann, E.L. (1954). The use of maximum likelihood estimates in χ^2 test for goodness of fit. *The Annals of Mathematical Statistics* **25**, 579–586.

Cholette, P.A., Dagum, E.B. (1994). Benchmarking time series with autocorrelated survey errors. *International Statistical Review* **62**, 365–377.

Chromy, J.R. (1987). Design optimization with multiple objectives. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, Washington, DC, pp. 194–199.

Chua, T., Fuller, W.A. (1987). A model for multivariate response error applied to labor flows. *Journal of the American Statistical Association* **82**, 46–51.

Cleveland, W.P., Tiao, G.C. (1976). Decomposition of seasonal time series: a model for the X-11 program. *Journal of the American Statistical Association* **71**, 581–587.

Cleveland, W.S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association* **74**, 829–836.

Cleveland, W.S., McGill, R. (1984). The many faces of a scatterplot. *Journal of the American Statistical Association* **79**, 807–822.

Cochran, W.G. (1939). The use of the analysis of variance in enumeration by sampling. *Journal of the American Statistical Association* **34**, 492–510.

Cochran, W.G. (1963). *Sampling Techniques*. Wiley, New York.

Cochran, W.G. (1977). *Sampling Techniques* (3rd ed.). John Wiley, New York.

Cohen, M.P., Kuo, L. (1985a). The admissibility of the empirical distribution function. *The Annals of Statistics* **13**(1), 262–271.

Cohen, M.P., Kuo, L. (1985b). Minimax sampling strategies for estimating a finite population distribution function. *Statistics and Decisions* **3**, 205–224.

Cohen, B.B., Barbano, H.E., Cox, C.S., Feldman, J.J., Finucane, F.F., Kleinman, J.C., Madans, J.H. (1987). Plan and operation of NHANES I Epidemiologic Followup Study, 1982–84. *Vital and Health Statistics Series 1*, No. 22 (DHHS Pub. No. PHS 87-1324). US Government Printing Office, National Center for Health Statistics, Washington, DC.

Cook, R.D., Weisberg, S. (1994). *An Introduction to Regression Graphics*. Wiley, New York.

Cosslett, S.R. (1981). Maximum likelihood estimation for choice-based samples. *Econometrica* **49**, 1289–1316.

Cowling, A., Chambers, R., Lindsay, R., Parameswaran, B. (1996). Applications of spatial smoothing to survey data. *Survey Methodology* **22**, 175–183.

Cox, D.R. (1975). Prediction intervals and empirical Bayes confidence intervals. In: Gani, J. (Ed.), *Perspectives in Probability and Statistics (Papers in Honor of M.S. Barlett)*. National Academic Press, Washington, DC, pp. 47–55.

Cox, D.R., Hinkley, D.V. (1974). *Theoretical Statistics*. Chapman and Hall, London.

Cressie, N.A.C. (1992). REML estimation in empirical Bayes smoothing of census undercount. *Survey Methodology* **18**, 75–94.

D'Alo, M., Di Consiglio, L., Falorsi, S., Solari, F. (2006). Small area estimation of the Italian poverty rate. *Statistics in Transition* **7**, 771–784.

Dagum, E.B. (1988). *The X11ARIMA/88 Seasonal Adjustment Method - Foundations and User's Manual.* Statistics Canada, Ottawa, ON.

Dagum, E.B., Chhab, N., Chiu, K. (1996). Derivation and properties of the X11 ARIMA and Census X11 linear filters. *Journal of Official Statistics* **12**, 329–347.

Darling, D.A., Robbins, H. (1967). Finding the size of a finite population. *The Annals of Mathematical Statistics* **38**, 1392–1398.

Darroch, J.N. (1958). The multiple recapture census, I: estimation of closed population. *Biometrika* **45**, 343–359.

Da Silva, D.N., Opsomer, J.D. (2004). Properties of the weighting cell estimator under a nonparametric response mechanism. *Survey Methodology* **30**, 45–55.

Da Silva, D.N., Opsomer, J.D. (2006). A kernel smoothing method to adjust for unit nonresponse in sample surveys. *The Canadian Journal of Statistics* **34**, 563–579.

Das, K., Jiang, J., Rao, J.N.K. (2004). Mean squared error of empirical predictor. *The Annals of Statistics* **32**, 818–840.

Datta, G.S. (1992). A unified Bayesian prediction theory for mixed linear models with application. *Statistics and Decisions* **10**, 337–365.

Datta, G.S., Day, B., Basawa, I.V. (1999a). Empirical best linear unbiased and empirical Bayes prediction in multivariate small area estimation. *Journal of Statistical Planning and Inference* **75**, 269–279.

Datta, G.S., Day, B., Maiti, T. (1998). Multivariate Bayesian small area estimation: an application to survey and satellite data. *Sankhya A* **60**, 344–362.

Datta, G.S., Fay, R.E., Ghosh, M. (1991). Hierarchical and empirical multivariate Bayes analysis in small area estimation. *Proceedings of the Seventh Annual Research Conference of the Bureau of the Census*, Arlington, VA, pp. 63–79.

Datta, G.S., Ghosh, M. (1991). Bayesian prediction in linear models: applications to small area estimation. *The Annals of Statistics* **19**, 1748–1770.

Datta, G.S., Ghosh, M., Huang, E., Isaki, C., Schultz, L., Tsay, J. (1992). Hierarchical and empirical Bayes methods for adjustment of census undercount: the 1988 Missouri dress rehearsal data. *Survey Methodology* **18**, 95–108.

Datta, G.S., Ghosh, M., Nangia, N., Natarajan, K. (1996). Estimation of median income of four-person families: a Bayesian approach. In: Berry, D.A., Chaloner, K.M., Geweke, J.K. (Eds.), *Bayesian Analysis in Statistics and Econometrics*. Wiley, New York, pp. 129–140.

Datta, G.S., Ghosh, M., Smith, D., Lahiri, P. (2002b). On an asymptotic theory of conditional and unconditional coverage probabilities of empirical Bayes confidence intervals. *Scandinavian Journal of Statistics* **29**, 139–152.

Datta, G.S., Ghosh, M., Waller, L. (2000). Hierarchical and empirical Bayes methods for environmental risk assessment. In: Sen, P.K., Rao, C.R. (Eds.), *Handbook of Statistics Bioenvironmental and Public Health Statistics,* vol. 18. North-Holland, pp. 223–245.

Datta, G.S., Lahiri, P. (2000). A unified measure of uncertainty of estimated best linear unbiased predictors in small area estimation problems. *Statistica Sinica* **10**, 613–627.

Datta, G.S., Lahiri, P., Maiti, T. (2002a). Empirical Bayes estimation of median income of four-person families by state using time series and cross-sectional data. *Journal of Statistical Planning and Inference* **102**, 83–97.

Datta, G.S., Lahiri, P., Maiti, T., Lu, K.L. (1999b). Hierarchical Bayes estimation of unemployment rates for the states of the U.S. *Journal of the American Statistical Association* **94**, 1074–1082.

Datta, G.S., Mukerjee, R. (2004). *Probability Matching Priors: Higher Order Asymptotics*. Springer.

Datta, G.S., Rao, J.N.K., Smith, D.D. (2005). On measuring the variability of small area estimators under a basic area level model. *Biometrika* **92**, 183–196.

de Boor, C. (2001). *A Practical Guide to Splines* (Revised ed.). Springer-Verlag, New York.

Dellaportas, P., Smith, A.F.M. (1993). Bayesian inference for generalized linear and proportional hazards models via Gibbs sampling. *Applied Statistics* **42**, 443–459.

DeMets, D., Halperin, M. (1977). Estimation of simple regression coefficients in samples arising from subsampling procedures. *Biometrics* **33**, 47–56.

Deming, W.E. (1956). On simplifications of sampling design through replication with equal probabilities and without stages. *Journal of the American Statistical Association* **51**, 24–53.

Deming, W.E., Stephan, F.F. (1940). On a least squares adjustment of a sample frequency table when the expected marginal total are known. *The Annals of Mathematical Statistics* **11**, 427–444.

Demnati, A., Rao, J.N.K. (2004). Linearization variance estimators for survey data (with discussion). *Survey Methodology* **30**, 17–34.

Dempster, A.P., Rubin, D.B., Tsutakawa, R.K. (1981). Estimation in covariance components models. *Journal of the American Statistical Association* **76**, 341–353.

Dempster, A.P., Tomberlin, T.J. (1980). The analysis of census undercount from a post-enumeration survey. *Proceedings of the Conference on Census Undercount*, pp. 88–94.

Deville, J.-C. (1999). Variance estimation for complex statistics and estimators: linearization and residual techniques. *Survey Methodology* **25**, 193–203.

Deville, J.-C., Goga, C. (2004). Estimation par régression par polynômes locaux dans des enquêtes sur plusieurs échantillons. In: Ardilly, P. (Ed.), *Echantillonnage et Méthodes d'Enquêtes*, Dunod, Paris-Sciences Sup, pp. 156–162.

Deville, J.-C., Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association* **87**, 376–382.

Deville, J.-C., Särndal, C.-E., Sautory, O. (1993). Generalized raking procedures in survey sampling. *Journal of the American Statistical Association* **88**, 1013–1020.

Diaconis, P., Ylvisaker, D. (1979). Conjugate priors for exponential families. *The Annals of Statistics* **7**, 269–281.

DiCiccio, T.J., Romano, J.P. (1988). A review of bootstrap confidence intervals. *Journal of the Royal Statistical Society, Series B* **50**, 338–354.

DiCiccio, T.J., Romano, J.P. (1990). Nonparametric confidence limits by resampling methods and least favorable families. *International Statistical Review* **58**, 59–76.

DiGaetano, R., Waksberg, J. (2002). Trade-offs in the development of a sample design for case-control studies. *American Journal of Epidemiology* **155**, 771–775.

Diggle, P.J., Liang, K.Y., Zeger, S.L. (1994). *Analysis of Longitudinal Data*. Clarendon Press, Oxford.

Diggle, P., Kenward, M.G. (1994). Informative drop-out in longitudinal data analysis. *Applied Statistics* **43**, 49–93.

Dorfman, A.H. (1992). Non-parametric regression for estimating totals in finite populations. *ASA Proceedings of the Section on Survey Research Methods*, American Statistical Association, Alexandria, VA, pp. 622–625.

Dorfman, A.H. (1993). A comparison of design-based and model-based estimators of the finite population distribution function. *Australian Journal of Statistics* **35**(1), 29–41.

Dorfman, A.H., Hall, P. (1993). Estimators of the finite population distribution function using nonparametric regression. *The Annals of Statistics* **21**(3), 1452–1457.

Dorfman, A.H., Valliant, R. (1993). Quantile variance estimators in complex surveys. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pp. 866–871.

DuMouchel, W.H., Duncan, G.L. (1983). Using sample survey weights in multiple regression analysis of stratified samples. *Journal of the American Statistical Association* **78**, 535–543.

Duncan, G.J., Kalton, G. (1987). Issues of design and analysis of surveys across time. *International Statistical Review* **55**, 97–117.

Dunstan, R., Chambers, R.L. (1989). Estimating distribution functions from survey data with limited benchmark information. *Australian Journal of Statistics* **31**(1), 1–11.

Durbin, J. (1958). Sampling theory for estimates based on fewer individuals than the number selected. *Bulletin of the International Statistical Institute* **36**, 113–119.

Durbin, J. (1959). A note on the application of Quenouille's method of bias reduction to the estimation of ratios. *Biometrika* **46**(3/4), 477–480.

Durbin, J. (1960). Estimation of parameters in time series regression models. *Journal of the Royal Statistical Society, Series B* **22**, 139–153.

Durbin, J., Quenneville, B. (1997). Benchmarking by state-space models. *International Statistical Review* **65**, 23–48.

Durrant, G.B., Skinner, C. (2006). Using missing data methods to correct for measurement error in a distribution function. *Survey Methodology* **32**, 25–36.

Durrelman, S., Simon, R. (1989). Flexible regression models with cubic splines. *Statistics in Medicine* **8**, 551–561.

Ecker, A.R. (1955). Rotation sampling. *The Annals of Mathematical Statistics* **26**, 664–685.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *The Annals of Statistics* 7, 1–26.

Efron, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans.* SIAM, Philadelphia, PA.

Efron, B. (1994). Missing data, imputation, and the bootstrap. *Journal of the American Statistical Association* **89**, 463–475 (discussion 475–479).

Efron, B., Hinkley, D.V. (1978). Assessing the accuracy of the maximum likelihood estimator: observed versus expected Fisher information (with discussion). *Biometrika* **65**, 457–487.

Eideh, A.H., Nathan, G. (2006). Fitting time series models for longitudinal survey data under informative sampling. *Journal of Statistical Planning and Inference* **136**, 3052–3069.

Eideh, A.H., Nathan, G. (2009). Joint treatment of nonignorable dropout and informative sampling for longitudinal survey data. In: Lynn, P. (Ed.), *Methodology of Longitudinal Surveys.* Wiley, New York, pp. 251–263.

Engels, E.A., Chen, J., Hartge, P., Cerhan, J.R., Davis, S., Severson, R.K., Cozen, W., Viscidi, R.P. (2005). Antibody responses to simian virus 40 T antigen: a case-control study of non-Hodgkin lymphoma. *Cancer Epidemiology, Biomarkers & Prevention* **14**, 521–525.

Erdös, P., Rényi, A. (1959). On the central limit theorem for samples from a finite population. *Publications of the Mathematical Institute of the Hungarian Academy of Sciences* **4**, 49–61.

Ericksen, E.P., Kadane, J.B. (1985). Estimating the population in a census year: 1980 and beyond. *Journal of the American Statistical Association* **80**, 98–109.

Ericson, W.A. (1965). Optimum stratified sampling using prior information. *Journal of the American Statistical Association* **60**, 750–771.

Ericson, W.A. (1969a). Subjective Bayesian models in sampling finite populations (with discussion). *Journal of the Royal Statistical Society, Series B* **31**, 195–233.

Ericson, W.A. (1969b). A note on the posterior mean. *Journal of the Royal Statistical Society, Series B* **31**, 332–334.

Ericson, W.A. (1988). Bayesian inference in finite populations. In: Krishnaiah, P.R., Rao, C.R. (Eds.), *Handbook of Statistics*, vol. 6. Elsevier Science Publishers, pp. 213–246.

Estevao, V.M., Hidiroglou, M.A., Särndal, C.-E. (1995). Methodological principles for a generalized estimation system at Statistics Canada. *Journal of Official Statistics* **11**, 181–204.

Estevao, V.M., Särndal, C.-E. (1999). The use of auxiliary information in design-based estimation for domains. *Survey Methodology* **2**, 213–221.

Estevao, V.M., Särndal, C.-E. (2000). A functional form approach to calibration. *Journal of Official Statistics* **16**, 379–399.

Estevao, V.M., Särndal, C.-E. (2004). Borrowing strength is not the best technique within a wide class of design-consistent domain estimators. *Journal of Official Statistics* **20**, 645–669.

Estevao, V.M., Särndal, C.-E. (2006). Survey estimates by calibration on complex auxiliary information. *International Statistical Review* **74,** 127–147.

Estrella, A. (1998). A new measure of fit for equations with dichotomous dependent variables. *Journal of Business and Economic Statistics* **10**, 198–205.

Eubank, R.L. (1999). *Nonparametric Regression and Smoothing Splines* (2nd ed.). Marcel Dekker, New York.

Fabrizi, E., Ferrante, M.R., Pacei, S. (2007). Small area estimation of average household income based on unit level models for panel data. *Survey Methodology* **33**, 187–198.

Fahrmeier, L., Tutz, G. (2001). *Multivariate Statistical Modeling Based on Generalized Linear Models* (2nd ed.). Springer, New York.

Falorsi, P.D., Falorsi, S., Russo, A. (2000). A conditional analysis of some small area estimators in sampling with two primary units selected in each stratum. *Statistics in Transition* **4**, 565–585.

Falorsi, P.D., Orsini, D., Righi, P. (2006). Balanced and coordinated sampling designs for small domain estimation. *Statistics in Transition* **7**, 805–829.

Farrell, P.J. (2000). Bayesian inference for small area proportions. *Sankhya B* **62**, 402–416.

Farrell, P.J., MacGibbon, B., Tomberlin, T.J. (1997). Empirical Bayes estimators of small area proportions in multistage designs. *Statistica Sinica* **7**, 1065–1083.

Fay, R.E. (1985). A jackknifed chi-square test for complex samples. *Journal of the American Statistical Association* **80**, 148–157.

Fay, R.E. (1987). Application of multivariate regression to small domain estimation. In: Platek, R., Rao, J.N.K., Sarndal, C.E., Singh, M.P. (Eds.), *Small Area Statistics*. Wiley, New York, pp. 91–102.

Fay, R.E., Herriot, R.A. (1979). Estimates of income for small places: an application of James-Stein procedures to census data. *Journal of the American Statistical Association* **74**, 269–277.

Fears, T.R., Brown, C.C. (1986). Logistic regression methods for retrospective case control studies using complex sampling procedures. *Biometrics* **42**, 955–960.

Fears, T.R., Gail, M.H. (2000). Analysis of a two-stage case-control study with cluster sampling of controls: application to nonmelanoma skin cancer. *Biometrics* **56**, 190–198.

Feder, M. (2001). Time series analysis of repeated surveys: the state-space approach. *Statistica Neerlandica* **55**, 182–199.

Federal Committee on Statistical Methodology (1993). Indirect Estimators in Federal Programs. US Office of Management and Budget, Statistical Policy Working Paper No. 21.

Fellegi, I.P. (1980). Approximate tests of independence and goodness of fit based on stratified multistage samples. *Journal of the American Statistical Association* **75**, 261–268.

Ferguson, T.S. (1967). *Mathematical Statistics: A Decision Theoretic Approach*. Academic Press, Inc, New York.

Fieller, E.C. (1932). The distribution of the index in a normal bivariate population. *Biometrika* **24**, 428–440.

Fienberg, S.E. (1980). *The Analysis of Cross-Classified Categorical Data* (2nd ed.). MIT Press, reprinted by Springer.

Findley, D.F., Monsell, B.C., Bell, W.R., Otto, M.C., Chen, B. (1998). New capabilities and methods of the X-12-ARIMA seasonal-adjustment program. *Journal of Business and Economic Statistics* **16**, 127–177.

Firth, D., Bennett, K.E. (1998). Robust models in probability sampling. *Journal of the Royal Statistical Society, Series B* **60**, 3–21.

Folsom, R.E., Singh, A.C. (2000). The generalized exponential model for sampling weight calibration for extreme values, nonresponse, and poststratification. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, Washington, DC, pp. 598–603.

Francesconi, M. (2005). An evaluation of the childhood family structure measures from the sixth wave of the British Household panel survey. *Journal of the Royal Statistical Society* **A168**, 539–566.

Francisco, C.A., Fuller, W.A. (1991). Quantile estimation with a complex survey design. *The Annals of Statistics* **19**, 454–469.

Freedman, D.A., Navidi, W.C. (1986). Regression methods for adjusting the 1980 census (with discussion). *Statistical Science* **18**, 75–94.

Friedman, J.H. (1991). Multivariate adaptive regression splines (with discussion). *The Annals of Statistics* **19**, 1–141.

Fuller, W.A. (1975). Regression analysis for sample surveys. *Sankhya, Series C* **37**, 117–132.

Fuller, W.A. (1976). *Introduction to Statistical Time Series*. Wiley, New York.

Fuller, W.A. (1984). Least squares and related analyses for complex survey designs. *Survey Methodology* **10**, 97–118.

Fuller, W.A. (1989). Prediction of true values for the measurement error model. *Conference on Statistical Analysis of Measurement Error Models and Applications*, Humboldt State University.

Fuller, W.A. (1990a). Analysis of repeated surveys. *Survey Methodology* **16**, 167–180.

Fuller, W.A. (1990b). Prediction of true values for the measurement error model. In: Brown, P.J., Fuller, W.A. (Eds.), *Statistical Analysis of Measurement Error Models and Applications, Contemporary Mathematics,* **112**. AMS, Providence, RI, pp. 41–57.

Fuller, W.A. (1998). Replication variance estimation for two-phase samples. *Statistica Sinica* **8**, 1153–1164.

Fuller, W.A. (2002). Regression estimation for survey samples. *Survey Methodology* **28**, 5–23.

Fuller, W.A., Harter, R.M. (1987). The multivariate components of variance model for small area estimation. In: Platek, R., Rao, J.N.K., Sarndal, C.E., Singh, M.P. (Eds.), *Small Area Statistics*. Wiley, New York, pp. 103–123.

Fuller, W.A., Louglin, M.M., Baker, H.D. (1994). Regression weighting for the 1987–88 National Food Consumption Survey. *Survey Methodology* **20**, 75–85.

Fuller, W.A., Rao, J.N.K. (2001). A regression composite estimator with application to the Canadian labour force. *Survey Methodology* **27**, 45–51.

Funaoka, F., Saigo, H., Sitter, R.R., Toida, T. (2006). Bernoulli bootstrap for stratified multistage sampling. *Survey Methodology* **32**(2), 151–156.

Gabler, S. (1990). *Minimax Solutions in Sampling from Finite Populations*. Springer-Verlag, Berlin, Germany.

Gabler, S., Stenger, H. (2000). Minimax strategies in survey sampling. *Journal of Statistical Planning and Inference* **90**, 305–321.

Gabriel, K.R. (1962). Ante-dependence analysis of an ordered set of variables. *The Annals of Mathematical Statistics* **33**, 201–212.

Gambino, J., Kennedy, B., Singh, M.P. (2001). Regression composite estimation for the Canadian labour force survey: evaluation and implementation. *Survey Methodology* **27**, 65–74.

Ganesh, N. (2007). Small area estimation and prediction problems: spatial models, Bayesian multiple comparisons, and robust MSE estimation. Ph.D. dissertation, Department of Mathematics, University of Maryland, College Park, MD.

Ganesh, N., Lahiri, P. (2008). A new class of average moment matching prior. To appear in *Biometrika* **95**.

Gelfand, A.E., Smith, A.F.M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association* **85**, 398–409.

Gelman, A. (2007). Struggles with survey weighting and regression modeling (with discussion). *Statistical Science* **22**, 153–164.

Gelman, A., Carlin, J.B., Stern, H.S., Rubin, D.B. (2003). *Bayesian Data Analysis* (2nd ed.). London: CRC Press.

Gelman, A., Carlin, J.B., Stern, H.S., Rubin, D.B. (2004). *Bayesian Data Analysis* (2nd ed). CRC Press, Boca Raton, FL.

Gentle, J.E. (1998). *Numerical Linear Algebra for Applications in Statistics*. Springer-Verlag, New York.

Gershunskaya, J.B., Lahiri, P. (2005). Variance estimation for domains in the U.S. Current Employment Statistics Program. *Proceedings of the Survey Research Methods*, American Statistical Association, pp. 3044–3051.

Ghosh, M. (1992). Constrained Bayes estimation with applications. *Journal of the American Statistical Association* **87**, 533–540.

Ghosh, M., Maiti, T. (2004). Small area estimation based on natural exponential family quadratic variance function models and survey weights. *Biometrika* **91**, 95–112.

Ghosh, M., Maiti, T. (2008). Empirical Bayes confidence intervals for means of natural exponential family quadratic variance function distributions with application to small area estimation. To appear in *Scandinavian Journal of Statistics*.

Ghosh, M., Meeden, G. (1997). *Bayesian Methods for Finite Population Sampling*. Chapman & Hall, London.

Ghosh, M., Meeden, G. (1986). *Empirical Bayes Estimation in Finite Population Sampling*. *Journal of the American Statical Association* **81**, 1058–1063.

Ghosh, M., Nangia, N., Kim, D. (1996). Estimation of median income of four-person families: a Bayesian time series approach. *Journal of the American Statistical Association* **91**, 1423–1431.

Ghosh, M., Natarajan, K. (1999). Small area estimation: a Bayesian perspective. In: Ghosh, S. (Ed.), *Multivariate Analysis, Design of Experiments and Survey Sampling*. Marcel Dekker, New York, pp. 69–92.

Ghosh, M., Natarajan, K., Stroud, T.W.F., Carlin, B.P. (1998). Generalized linear models for small-area estimation. *Journal of the American Statistical Association* **93**, 273–282.

Ghosh, M., Rao, J.N.K. (1994). Small area estimation: an appraisal (with discussion). *Statistical Science* **9**, 55–93.

Ghosh, M., Sinha, B.K. (1990). On the consistency between model and design based estimators in survey sampling. *Communications in Statistics* **20**, 689–702.

Ghosh, M., Sinha, K. (2007). Empirical Bayes estimation in finite population sampling under functional measurement error models. *Journal of Statistical Planning and Inference* **137**, 2759–2773.

Ghosh, M., Sinha, K., Kim, D. (2006). Empirical and hierarchical Bayesian estimation in finite population sampling under structural measurement error models. *Scandinavian Journal of Statistics* **33**, 591–608.

Ghysels, E., Perron, P. (1993). The effect of seasonal adjustment filters and tests for a unit root. *Journal of Econometrics* **55**, 57–98.

Gilks, W.R., Wild, P. (1992). Adaptive rejection sampling for Gibbs sampling. *Applied Statistics* **41**, 337–348.

Gill, R.D., Robins, J.M. (2001). Causal inference for complex longitudinal data: the continuous case. *The Annals of Statistics* **29**, 1785–1811.

Giommi, A. (1984). On the estimation of the probability of response in finite population sampling (Italian). *Societa Italiana di Statistica, Atti della Riunione Scientifica della Societa Italiana* **32**(1), 275–284.

Giommi, A. (1987). Nonparametric methods for estimating individual response probabilities. *Survey Methodology* **13**, 127–134.

Godambe, V.M. (1955). A unified theory of sampling from finite populations. *Journal of the Royal Statistical Society, Series B* **17**, 269–278.

Godambe, V.P., Joshi, V.M. (1965). Admissibility and Bayes estimation in sampling finite populations I. *The Annals of Mathematical Statistics* **36**, 1707–1722.

Godambe, V.P. (1955). A unified theory for sampling from finite populations. *Journal of the Royal Statistical Society*, *Series B* **17**, 269–278.

Godambe, V.P. (1960). An optimal property of regular maximum likelihood estimator. *The Annals of Mathematical Statistics* **31**, 1208–1212.

Godambe, V.P. (1966a). A new approach to sampling from finite populations, I. sufficiency and linear estimation. *Journal of the Royal Statistical Society, Series B* **28**, 310–319.

Godambe, V.P. (1966b). A new approach to sampling from finite populations, I and II. *Journal of the Royal Statistical Society, B* **28**, 310–328.

Godambe, V.P. (1989). Estimation of Cumulative Distribution of a Survey Population. Technical report, University of Waterloo.

Godambe, V.P. (1991). Orthogonality of estimating functions and nuisance parameters. *Biometrika* **78**, 143–151.

Godambe, V.P., Heyde, C.C. (1987). Quasi likelihood and optimal estimation. *International Statistical Review* **55**, 231–244.

Godambe, V.P., Joshi, V.M. (1965). Admissibility and Bayes estimation in sampling finite populations – I. *The Annals of Mathematical Statistics* **36**, 1707–1723.

Godambe, V.P., Kale, B.K. (1991). Estimating functions: an overview. In: Godambe, V.P. (Ed.), *Estimating Functions.* Oxford University Press, Oxford, pp. 3–20.

Godambe, V.P., Thompson, M.E. (1986a). Parameters of superpopulation and survey population: their relationships and estimation. *International Statistical Review/Revue Internationale de Statistique* **54**(2), 127–138.

Godambe, V.P., Thompson, M.E. (1986b). Some optimality results in the presence of non-response. *Survey Methodology* **12**, 29–36.

Godambe, V.P., Thompson, M.E. (1989). An extension of quasi-likelihood estimation (with discussion). *Journal of Statistical Planning and Inference* **12**, 137–172.

Godambe, V.P., Thompson, M.E. (1999). A new look at confidence intervals in survey sampling. *Survey Methodology* **25**, 161–173.

Godambe, V.P., Thompson, M.E. (2006). Reflections on missing data. Working paper.

Goebel, J.J., Schmude, K.O. (1982). Planning the SCS National Resources Inventory. Arid Land Resource Inventories' Workshop, pp. 148–153. USDA, Forest Service General Technical Report, WO-28, Washington, DC.

Goga, C. (2004). Estimation de l'évolution d'un total en présence d'information auxiliaire: une approche par splines de régression. *Comptes Rendus de l'Académie des Sciences Paris Ser. I* **339**, 441–444.

Goga, C. (2005). Réduction de la variance dans les sondages en présence d'information auxiliaire: une approche non paramétrique par splines de régression. *The Canadian Journal of Statistics* **33**, 163–180.

Goldstein, H. (1986). Multi-level mixed linear model analysis using iterative generalized least squares. *Biometrika* **73**, 43–56.

Goldstein, H. (1995). *Multilevel Statistical Models*. Edward Arnold, London.

Goldstein, H. (2002). *Multilevel Statistical Models* (3rd ed.). Edward Arnold, London.

Goldstein, H., Healy, M.J.R., Rasbash, J. (1994). Multilevel time series models with applications to repeated measures data. *Statistics in Medicine* **13**, 1643–1655.

Goldstein, M. (1975a). Approximate Bayes solutions to some nonparametric problems. *The Annals of Statistics* **3**, 512–517.

Goldstein, M. (1975b). A note on some Bayesian nonparametric estimates. *The Annals of Statistics* **3**, 736–740.

Gomez, V., Maravall, A. (1997). Programs TRAMO and SEATS: instructions for the user (beta version: June 1997). Working paper 97001, Ministerio de Economia y Hacienda, Dorrection General de Analisis y Programacion Presupuestraria, Madrid, Spain.

Gong, T, Little, R.J.A., Ragmunathan, T.E. (2003). Analysis of multivariate missing data with nonignorable response. *Biometrika* **90**, 747–764.

Gonzalez, M.E. (1973). Use and evaluation of synthetic estimates. *Proceedings of the Social Statistics Section*, American Statistical Association, 33–36.

Goodman, L.A. (1953). Sequential sampling tagging for population size problems. *The Annals of Mathematical Statistics* **24**, 56–69.

Graubard, B.I., Fears, T.R., Gail, M.H. (1989). Effects of cluster sampling on epidemiologic analysis in population-based case-control sampling. *Biometrics* **45**, 1053–1071.

Graubard, B.I., Korn, E.L. (2002). Inference for superpopulation parameters using sample surveys. *Statistical Science* **17**, 73–96.

Graubard, B.I., Korn, E.L., Midthune, D. (1997). Testing goodness of fit for logistic regression with survey data. *Proceedings of the Section on Survey Research Methods,* American Statistical Association, Washington, DC, 170–174.

Grilli, L., Pratesi, M. (2004). Weighted estimation in multi-level ordinal and binary models in the presence of informative sampling designs. *Survey Methodology* **30**, 93–103.

Grizzle, J.E., Starmer, C.F., Koch, G.G. (1969). Analysis of categorical data by linear models. *Biometrics* **25**, 489–504.

Gross, S. (1980). Median estimation in sample surveys. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, Alexandria, VA, 181–184.

Gunlicks, C.l.A., Corteville, J.S., Mansur, K. (1997). Current population survey variance properties. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, Alexandria, VA, pp. 558–563.

Guo, S., Roche, A.F., Baumgartner, R.N., Chumlea, W.C., Ryan, A.S. (1990). Kernel regression for smoothing percentile curves: reference data for calf and subscapular skinfold thicknesses in Mexican Americans. *American Journal of Clinical Nutrition* **51**, 908S–916S.

Gurney, M., Daley, J.F. (1965). A multivariate approach to estimation in periodic sample surveys. *Proceedings of the Section of Survey and Research Methods*, American Statistical Association, Alexandria, VA, pp. 247–257.

Haberman, S.J. (1974). *The Analysis of Frequency Data.* University of Chicago Press, Chicago, IL.

Hájek, J. (1959). Optimum strategy and other problems in probability sampling. *Časopis pro pěstování matematiky* **84**, 387–423.

Hájek, J. (1960). Limiting distributions in simple random sampling from a finite population. *Publications of the Mathematical Institute of the Hungarian Academy of Sciences* **5**, 361–374.

Hájek, J. (1961). Some extensions of the Wald-Wolfowitz-Noether theorem. *The Annals of Mathematical Statistics* **32**, 506–523.

Hájek, J. (1964). Asymptotic theory of rejective sampling with varying probabilities from a finite population. *The Annals of Mathematical Statistics* **35**, 1491–1523.

Hájek, J. (1971). Discussion of 'an essay on the logical foundations of survey sampling, part one' by D. Basu. In: Godambe, V.P., Sprott, D.A. (Eds.), *Foundations of Statistical Inference*. Holt, Rinehart and Winston, Toronto, ON, Canada, p. 236.

Hájek, J. (1981). *Sampling from a Finite Population*. Dekker, New York.

Hájek, J., Šidák, Z. (1967). *Theory of Rank Tests*. Academic Press, New York.

Hall, P., Maiti, T. (2006a). Nonparametric estimation of mean squared prediction error in the nested-error regression models. *The Annals of Statistics* **34**, 1733–1750.

Hall, P., Maiti, T. (2006b). On parametric bootstrap methods for small area prediction. *Journal of the Royal Statistical Society, Series B: Statistical Methodology* **68**(2), 221–238.

Hansen, M.H., Hurvitz, W.N. (1943). On the theory of sampling from finite populations. *The Annals of Mathematical Statistics* **14**, 333–362.

Hansen, M.H, Hurvitz, W.N., Madow, W.G. (1978). On inference and estimation from sample surveys (with discussion). *Proceedings of the Survey Research Methods Section*, American Statistical Association, pp. 82–107.

Hansen, M.H., Hurwitz, W.N., Madow, W.G. (1953). *Sample Survey Methods and Theory*. John Wiley & Sons, New York; Chichester.

Hansen, M.H., Madow, W.G., Tepping, B.J. (1983). An evaluation of model-dependent and probability-sampling inferences in sample surveys (with discussion). *Journal of the American Statistical Association* **78**, 776–807.

Härdle, W. (1990). *Applied Nonparametric Regression*. Cambridge University Press, Cambridge.

Harms, T., Duchesne, P. (2006). On calibration estimation for quantiles. *Survey Methodology* **32**(1), 37–52.

Hartge, P., Brinton, L.A., Cahill, J.I., West, D., Hauk, M., Austin, D., Silverman, D., Hoover, R.N. (1984a). Design and methods in a multi-center case-control interview study. *American Journal of Public Health* **74**, 52–56.

Hartge, P., Brinton, L.A., Rosenthal, J.F., Cahill, J.I., CityplaceHoover, R.N., Waksberg, J. (1984b). Random digit dialing in selecting a population-based control group. *American Journal of Epidemiology* **120**, 825–833.

Hartigan, J.A. (1969). Linear Bayesian methods. *Journal of the Royal Statistical Society, Series B* **31**, 446–454.

Hartley, H.O., Rao, J.N.K. (1962). Sampling with unequal probabilities and without replacement. *The Annals of Mathematical Statistics* **33**, 350–374.

Hartley, H.O. (1959). Analytical studies of survey data. In: *Volume in Honor of Corrado Gini*. Instituto di Statistica, Rome, Italy.

Hartley, H.O. (1966). Systematic sampling with unequal probability and without replacement. *Journal of the American Statistical Association* **61**, 739–748.

Hartley, H.O., Rao, J.N.K. (1968). A new estimation theory for sample surveys. *Biometrika* **55**, 547–557.

Hartley, H.O., Rao, J.N.K. (1969). A new estimation theory for sample surveys, II. In: *New Developments in Survey Sampling*. Wiley Interscience, New York, pp. 147–169.

Harvey, A. (1989). Forecasting, Structural Time Series Models, and the Kalman Filter. Cambridge University Press.

Harvey, A., Chung, C. (2000). Estimating the underlying change in unemployment in the UK. *Journal of the Royal Statistical Society, Series A: Statistics in Society* **163**, 303–328.

Harville, D.A. (1985). Decomposition of prediction error. *Journal of the American Statistical Association* **80**, 132–138.

Hastie, T.J., Tibshirani, R.J. (1990). *Generalized Additive Models*. Chapman and Hall, Washington, DC.

Hausman, J.A., Watson, M.W. (1985). Errors in variables and seasonal adjustment procedures. *Journal of the American Statistical Association* **80**, 531–540.

Haziza, D., Rao, J.N.K. (2006). A nonresponse model approach to inference under imputation for missing survey data. *Survey Methodology* **32**, 53–64.

Heckman, J.J., Robb, R. (1985). Alternative methods for evaluating the impact of interventions: an overview. *Journal of Econometrics* **30**, 239–267.

Heckman, J.J., Robb, R. (1989).The value of longitudinal data for solving the problem of selection bias in evaluating the impact of treatments on outcomes. In: Kasprzyk, D., Duncan, G., Kalton, G., Singh, M.P. (Eds.), *Panel Surveys*. John Wiley & Sons, Chichester, New York, pp. 512–538.

Hedayat, A.S., Sinha, B.K. (1991). *Design and Inference in Finite Population Sampling*. Wiley, New York.

Hedlin, D., Falvey, H., Chambers, R., Kokic, P. (2001). Does the model matter for GREG estimation? A business survey example. *Journal of Official Statistics* **17**, 527–544.

Helland, I.S. (1995). Simple counterexamples against the conditionality principle. *The American Statistician* **49**, 351–356.

Helmers, R., Wegkamp, M. (1998). Wild bootstrapping in finite populations with auxiliary information. *Scandinavian Journal of Statistics* **25**, 383–399.

Henderson, C.R. (1975). Best linear unbiased estimation and prediction under a selection model. *Biometrics* **31**, 423–447.

Henderson, C.R., Kempthorne, O., Searle, S.R., von Krogisk, C.N. (1959). Estimation of environmental and genetic trends from records subject to culling. *Biometrics* **13**, 192–218.

Herson, J. (1976). An investigation of relative efficiency of least-squares prediction to conventional probability sampling plans. *Journal of the American Statistical Association* **71**, 700–703.

He, Z., Sun, D. (1998). Hierarchical Bayes estimation of hunting success rates. *Environmental and Ecological Statistics* **5**, 97–100.

Hidiroglou, M.A., Patak, Z. (2004). Domain estimation using linear regression. *Survey Methodology* **30**, 67–78.

Hidiroglou, M.A., Patak, Z. (2006). Raking ratio estimation: an application to the Canadian Retail Trade Survey. *Journal of Official Statistics* **22**, 71–80.

Hidiroglou, M.A., Rao, J.N.K. (1987). Chi-squared tests with categorical data from complex surveys: parts I and II. *Journal of Official Statistics* **3**, 117–132 (133–140).

Hidiroglou, M.A., Särndal, C.-E. (1985). An empirical study of some regression estimators for small domains. *Survey Methodology* **11**, 65–77.

Hill, B.M. (1968). Posterior distribution of percentiles: Bayes's theorem for sampling from a finite population. *Journal of the American Statistical Association* **63**, 677–691.

Hillmer, S.C., Tiao, G.C. (1982). An ARIMA-model-based approach to seasonal adjustment. *Journal of the American Statistical Association* **77**, 63–70.

Hillmer, S.C., Trabelsi, A. (1987). Benchmarking of economic time series. *Journal of the American Statistical Association* **82**, 1064–1071.

Hinkins, S., Oh, H.L., Scheuren, F. (1994). Inverse sampling design algorithms. *Proceedings of the Section on Survey Research Methods,* American Statistical Association, 626–631.

Hinrichs, P.E. (2003). Consumer expenditure estimation incorporating generalized variance functions in hierarchical Bayes models. Unpublished Ph.D. dissertation, University of Nebraska-Lincoln, Lincoln, NE.

Hoaglin, D.A., Welsch, R.E. (1978). The hat matrix in regression and ANOVA. *The American Statistician* **32**, 17–22, and *Corrigenda* **32**, 146.

Hodges, Jr. J.L., Lehmann, E.L. (1982). Minimax estimation in simple random sampling. In: Kallianpur, G., Krishnaiah, P.R., Ghosh, J.K. (Eds.), *Statistics and Probability: Essays in Honor of C. R. Rao*. North-Holland, Amsterdam, pp. 325–327.

Hoeffding, W. (1948). On a class of statistics with asymptotically normal distribution. *The Annals of Mathematical Statistics* **19**, 293–325.

Hogan, J.W., Lee, J.Y. (2004). Marginal structural quantile models for longitudinal observational studies with time-varying treatment. *Statistica Sinica* **14**, 927–944.

Holt, D., Scott, A.J., Ewings, P.D. (1980). Chi-square test with survey data. *Journal of the Royal Statistical Society, Series A* **143**, 303–320.

Holt, D., Skinner, C.J. (1989). Components of change in repeated surveys. *International Statistical Review* **57**, 1–18.

Holt, D., Smith, T.M.F. (1979). Post-stratification. *Journal of the Royal Statistical Society, Series A* **142**, 33–46.

Holt, D., Smith, T.M.F., Tomberlin, T.J. (1979). A model-based approach to estimation for small subgroups of a population. *Journal of the American Statistical Association* **74**, 405–410.

Holt, D., Smith, T.M.F., Winter, P.D. (1980). Regression analysis of data from complex surveys. *Journal of the Royal Statistical Society, Series A* **143**, 474–487.

Horton, N.J., Bebechuk, J.D., Jones, C.L., Lipsitz, S.R., Caralano, P.J., Zaher, G.E.P., Fitzmaurice, G.M. (1999). Goodness of fit for GEE: an example with mental health service utilization. *Statistics in Medicine* **18**, 213–222.

Horvitz, D.G., Thompson, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* **47**, 663–685.

Hosmer, D.W., Lemeshow, S. (1980). A goodness of fit test for the multiple logistic regression. *Communications in Statistics* **A10**, 1043–1069.

Houseman, E.A., Ryan, L.M., Coull, B.A. (2004). Cholesky residuals for assessing normal errors in a linear model with correlated outcomes. *Journal of the American Statistical Association* **99**, 383–394.

Huang, E.T., Ernst, L.R. (1981). Comparison of an alternative estimator to the current composite estimator in CPS. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, Alexandria, VA, 303–308.

Huang, E.T., Fuller, W.A. (1978). Nonnegative regression estimation for sample survey data. *Proceedings of the Section on Social Statistics*, American Statistical Association, Washington, DC, 300–305.

Hu, F., Kalbfleisch, J.D. (2000). The estimating function bootstrap. *The Canadian Journal of Statistics/La Revue Canadienne de Statistique* **28**, 449–481.

Hyndman, R.J., Fan, Y. (1996). Sample quantiles in statistical packages. *The American Statistician* **21**(4), 361–365.

Iachan, R. (1983). Asymptotic theory of systematic sampling. *The Annals of Statistics* **11**, 959–969.

Ireland, C.T., Kullback, S. (1968). Contingency tables with given marginals. *Biometrika* **55**, 179–188.

Isaki, C.T., Fuller, W.A. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association* **77**, 89–96.

Isaki, C.T., Huang, E.T., Tsay, J.H. (1991). Smoothing adjustment factors from the 1990 post enumeration survey. *Proceedings of the Social Statistics Section*, American Statistical Association, Washington, DC, 338–343.

Jayasuriya, B.R., Valliant, R. (1996). An application of regression and calibration estimation to post-stratification in a household survey. *Survey Methodology* **22**, 127–137.

Jensen, R.J. (1942). Statistical investigtion of a sample survey for obtaining farm facts. *Iowa Agricultural Experiment Station Research Bulletin* **304**, 54–59.

Jiang, J. (1996). REML estimation: asymptotic behavior and related topics. *The Annals of Statistics* **24**, 255–286.

Jiang, J. (1998). Consistent estimators in generalized linear mixed models. *Journal of the American Statistical Association* **93**, 720–729.

Jiang, J., Lahiri, P.S. (2006). Estimation of finite population domain means: a model - assisted empirical best prediction approach. *Journal of the American Statistical Association* **101**, 301–311.

Jiang, J., Lahiri, P.S. (2001). Empirical best prediction for small area inference with binary data. *Annals of the Institute of Statistical Mathematics* **53**, 217–243.

Jiang, J., Lahiri, P.S. (2006). Mixed model prediction and small area estimation. *Test* **15**, 1–96.

Jiang, J., Lahiri, P.S., Wan, S.-M. (2002). A unified Jackknife theory for empirical best prediction with M-estimation. *The Annals of Statistics* **30**, 1782–1810.

Jiang, J., Zhang, W. (2001). Robust estimation in generalized linear mixed models. *Biometrika* **88**, 753–765.

Johnson, A.A., Breidt, F.J., Opsomer, J.D. (2008). Estimating distribution functions from survey data using nonparametric regression. *Journal of Statistical Theory and Practice* **2**, 419–431.

Johnson, A.A., Breidt, F.J., Opsomer, J. (2004). Estimating distribution functions from survey data using nonparametric regression. Preprint Series #04-07. Department of Statistics, Iowa State University.

Jones, R.G. (1980). Best linear unbiased estimators for repeated surveys. *Journal of the Royal Statistical Society, Series B: Methodological* **42**, 221–226.

Joshi, V.M. (1965). Admissibility and Bayes estimation in sampling finite populations III. *The Annals of Mathematical Statistics* **36**, 1730–1742.

Joshi, V.M. (1966). Admissibility and Bayes estimation in sampling finite populations IV. *The Annals of Mathematical Statistics* **37**, 1658–1670.

Jowell, N.P. (1985). Least squares regression with data arising from stratified samples of the dependent variable. *Biometrika* **72**, 11–21.

Judkins, D.R. (1990). Fay's method for variance estimation. *Journal of Official Statistics* **6**(3), 223–239.

Kackar, R.N., Harville, D.A. (1984). Approximations for standard errors of estimators of fixed and random effects in mixed linear models. *Journal of the American Statistical Association* **79**, 853–862.

Kalton, G. (1983). Models in the practice of survey sampling. *International Statistical Review* **51**, 175–188.

Kalton, G. (1986). Handling wave nonresponse in panel surveys. *Journal of Official Statistics* **2**, 303–314.

Kalton, G., Citro, C.F. (1993). Panel surveys: adding the fourth dimension. *Survey Methodology* **19,** 205–215.

Kasprzyk, D., Duncan, G., Kalton, G., Singh, M.P. (Eds.). (1989). *Panel Surveys*. John Wiley & Sons, New York.

Kaufmann, P.R., Herlihy, A.T., Elwood, J.W., Mitch, M.E., Overton, W.S., Sale, M.J., Cougan, K.A., Peck, D.V., Reckhow, K.H., Kinney, A.J., Christie, S.J., Brown, D.D., Hagley, C.A., Jager, H.I. (1988). Chemical Characteristics of Streams in the Mid-Atlantic and Southeastern United States. Volume I: Population Descriptions and Physico-Chemical Relationships. EPA/600/3-88/021a

Kazemian, L., Farrington, D.P. (2005). Comparing the validity of prospective, retrospective, and official onset for different offending categories. *Journal of Quantitative Criminology* **21**, 127–147.

Keiding, N. (1999). Event history analysis and inference from observational epidemiology. *Statistics in Medicine* **18**, 2353–2363.

Kiefer, J. (1957). Invariance, minimax sequential estimation, and continuous time processes. *The Annals of Mathematical Statistics* **28**, 573–601.

Kim, J.-K., Fuller, W.A. (1999). Jackknife variance estimation after hot deck imputation. *1999 Proceedings of the Section on Survey Research Methods*, American Statistical Association, Alexandria, VA, 825–830.

Kim, J.-K., Fuller, W.A. (2004). Fractional hot-deck imputation. *Biometrika* **91**, 559–578.

Kim, J.-K., Kim, J.J. (2007). Nonresponse weighting adjustment using estimated response probability. *The Canadian Journal of Statistics* **35**, 501–514.

Kim, J.-K., Navarro, A., Fuller, W.A. (2006). Replication variance estimation for two-phase stratified sample. *Journal of the American Statistical Association* **101**, 312–320.

Kim, J.-K., Sitter, R.R. (2003). Efficient replication variance estimation for two-phase sampling. *Statistica Sinica* **13**, 641–653.

Kim, K.W., Park, Y.S., Kim, N.Y. (2005). l-step generalized composite estimator under 3-way balanced rotation design. *Journal of the Korean Statistical Society* **34**, 219–233.

Kish, L. (1980). Design and estimation for domains. *The Statistician* **29**, 209–222.

Kish, L. (1987). *Statistical Design for Research*. John Wiley & Sons, New York.

Kish, L. (1998). Space/time variations and rolling samples. *Journal of Official Statistics* **14**, 31–46.

Kish, L., Frankel, M.R. (1974). Inference from complex samples (with discussion). *Journal of the Royal Statistical Society, Series B: Statistical Methodology* **36**, 1–37.

Knowles, J. (1997). *Trend Estimation Practices of National Statistical Institutes*. Office for National Statistics, Methods and Quality Division, UK, MQ043.

Knowles, J., Kenny, P. (1997). *An Investigation of Trend Estimation Methods*. Office for National Statistics, Methods and Quality Division, UK, MQ043.

Koch, G.G., Freeman, D.H., Freeman, J.L. (1975). Strategies in the multivariate analysis of data from complex surveys. *International Statistical Review* **93**, 59–78.

Koehler, K., Larnz, K. (1986). An empirical investigation of goodness-of-fit statistics for sparse multinomials. *Journal of the American Statistical Association* **75**, 336–344.

Ko, H., Hogan, J.W., Mayer, K.H. (2003). Estimating causal treatment effects from longitudinal HIV natural history studies using marginal structural models. *Biometrics* **59**, 152–162.

Koopman, S.J., Harvey, A.C., Doornik, J.A., Shephard, N. (2000). *STAMP 6.0: Structural Time Series Analyser, Modeller and Predictor*. Timebrlake Consultants, London.

Korn, E.L., Graubard, B.I. (1990). Simultaneous testing of regression coefficients with complex survey data. *The American Statistician* **44**, 270–276.

Korn, E.L., Graubard, B.I. (1995). Analysis of large health surveys: accounting for the sample design. *Journal of the Royal Statistical Society, Series A* **158**, 263–295.

Korn, E.L., Graubard, B.I. (1998a). Variance estimation for superpopulation parameters. *Statistica Sinica* **8**, 1131–1151.

Korn, E.L., Graubard, B.I. (1998b). Confidence intervals for proportions with small expected number of positive counts estimated from survey data. *Survey Methodology* **24**, 193–201.

Korn, E.L., Graubard, B.I. (1998c). Scatterplots with survey data. *American Statistician* **52**, 58–69.

Korn, E.L., Graubard, B.I. (1999). *Analysis of Health Surveys*. John Wiley & Sons, New York.

Korn, E.L., Graubard, B.I. (2003). Estimating variance components by using survey data. *Journal of the Royal Statistical Society, Series B* **65**, 175–190.

Korn, E.L., Midthune, D., Graubard, B.I. (1997). Estimating interpolated percentiles from grouped data with large samples. *Journal of Official Statistics* **13**, 385–399.

Kott, P.S. (1989). Robust small domain estimation using random effects modeling. *Survey Methodology* **15**, 3–12.

Kott, P.S. (1990). The design consistent regression estimator and its conditional variance. *Journal of Statistical Planning and Inference* **24**, 287–296.

Kott, P.S. (1994). A note on handling nonresponse in surveys. *Journal of the American Statistical Association* **89**, 693–696.

Kott, P.S. (2003). A practical use for instrumental-variable calibration. *Journal of Official Statistics* **19**, 265–272.

Kott, P.S. (2005). Randomization-assisted model-based survey sampling. *Journal of Statistical Planning and Inference* **48**, 263–277.

Kott, P.S. (2006a). Using calibration weighting to adjust for nonresponse and coverage errors. *Survey Methodology* **32**, 133–142.

Kott, P.S. (2006b). Delete-a-group variance estimation for the general regression estimator under Poisson sampling. *Journal of Official Statistics* **22**, 759–767.

Kott, P.S. (2007). Clarifying some issues in the regression analysis of survey data. *Survey Research Methods* **1**, 11–18. Available at: http://w4.ub.uni-konstanz.de/srm/article/view/47/46.

Kott, P.S., Bailey, J.T. (2000). The theory and practice of maximal Brewer selection. *Proceedings of the Second International Conference on Establishment Surveys, invited papers*, 269–278.

Kott, P.S., Brewer, K.R.W. (2001). Estimating the model variance of a randomization-consistent regression estimator. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, Washington, DC.

Kott, P.S., Stukel, D.M. (1997). Can the jackknife be used with a two-phase sample? *Survey Methodology* **23**, 81–89.

Kott, P.S., Swensson, B., Särndal, C.-E., Wretman, J. (2005). An interview with the authors of the book *Model Assisted Survey Sampling*. *Journal of Official Statistics* **21,** 171–182.

Kovacevic, M.S. (1997). Calibration estimation of cumulative distribution function and quantile from survey data. *Proceedings of the Survey Methods Section, Statistical Society of Canada Meeting*, 1–7.

Kovačević, M.S., Rai, S.N. (2003). A pseudo maximum likelihood approach multilevel modelling of survey data. *Communications in Statistics. Theory and Methods* **32**, 103–121.

Kovar, J.G., Rao, J.N.K., Wu, C.F.J. (1988). Bootstrap and other methods to measure errors in survey estimates. *The Canadian Journal of Statistics/La Revue Canadienne de Statistique* **16**, 25–45.

Krewski, D. (1978). Jackknifing U-statistics in finite populations. *Communications in Statistics: Theory and Methods A* **7**, 1–12.

Krewski, D., Rao, J.N.K. (1981). Inference from stratified samples: properties of the linearization, jackknife and balanced repeated replication methods. *The Annals of Statistics* **9**, 1010–1019.

Krieger, A.M., Pfeffermann, D. (1997). Testing of distribution functions from complex sample surveys. *Journal of Official Statistics* **13**, 123–142.

Krishnaiah, P.R., Rao, C.R. (Eds.). (1988). *Handbook of Statistics*, vol. 6. Elsevier Science, Amsterdam.

Kuk, A.Y.C. (1988). Estimation of distribution functions and medians under sampling with unequal probabilities. *Biometrika* **75**(1), 97–103.

Kuk, A.Y.C. (1993). A Kernel method for estimating finite population distribution functions using auxiliary information. *Biometrika* **80**(2), 385–392.

Kuk, A.Y.C., Mak, T.K. (1989). Median estimation in the presence of auxiliary information. *Journal of the Royal Statistical Society, Series B* **51**(2), 261–269.

Kuk, A.Y.C., Mak, T.K. (1994). A functional approach to estimating finite population distribution functions. *Communications in Statistics, Theory and Methods* **23**(3), 883–896.

Kulperger, R.J., Singh, A.C. (1982). On random grouping in goodness-of-fit tests of discrete distributions. *Journal of Statistical Planning and Inference* **7**, 109–115.

Kumar, S., Lee, H. (1983). Evaluation of composite estimation for the Canadian labour force survey. *Survey Methodology* **9**, 178–201.

Kumar, S., Singh, A.C. (1987). On efficient estimation of unemployment rates from Labour Force Survey Data. *Survey Methodology* **13**, 75–83.

Kuo, L. (1988). Classical and prediction approaches to estimating distribution functions from survey data. *ASA Proceedings of the Section on Survey Research Methods*, American Statistical Association, Alexandria, VA, 280–285.

Lahiri, D.B. (1951). A method of sample selection providing unbiased ratio estimates. *Bulletin of the International Statistics Institute* **33**, 133–140.

Lahiri, P. (1995). A jackknife measure of uncertainty of linear empirical Bayes estimators. Unpublished manuscript.

Lahiri, P. (2003a). On the impact of bootstrap in survey sampling and small-area estimation. *Statistical Science* **18**, 199–210.

Lahiri, P. (2003b). A review of empirical best linear unbiased prediction for the Fay-Herriot small-area model. *The Philippine Statistician* **52**, 1–15.

Lahiri, P. (2008). Some thoughts on resampling methods in sample surveys. Unpublished manuscript.

Lahiri, P., Rao, J.N.K. (1995). Robust estimation of mean squared error of small area estimators. *Journal of the American Statistical Association* **90**, 758–766.

Lahiri, S.N., Maiti, T, Katzoff, M., Parson, V. (2007). Resampling based empirical prediction: an application. *Biometrika* **94**, 469–485.

Laird, N.M., Louis, T.A. (1987). Empirical Bayes confidence intervals based on bootstrap samples. *Journal of the American Statistical Association* **82**, 739–750.

Lam, K.F., Yu, P.L.H., Lee, C.F. (2002). Kernel method for the estimation of the distribution function and the mean with auxiliary information in ranked set sampling. *Environmetrics* **13,** 397–406.

Laniel, N. (1987). Variances for a rotating sample from a changing population. *Proceeding of the Survey Research Methods Section*. American Statistical Association, Alexandria, VA, 496–500.

Larsen, D.P., Thornton, K.W., Urquhart, N.S., Paulsen, S.G. (1994). The role of sample surveys for monitoring the condition of the Nation's lakes. *Environmental Monitoring and Assessment* **32**, 101–134.

Lawless, J.F. (2003a). Censoring and weighting in survival estimation from survey data. *Proceedings of the Survey Methods Section*, Statistical Society of Canada, Ottawa, 31–36.

Lawless, J.F. (2003b). Event history analysis and longitudinal surveys. In: Skinner, C., Chambers, R. (Eds.), *Analysis of Survey Data*. Wiley, New York, pp. 221–243.

Lawless, J.F., Kalbfleisch, J.D., Wild, C.J. (1999). Semiparametric methods for response-selective and missing data problems in regression. *Journal of the Royal Statistical Society, Series B* **61**, 413–438.

Lawless, J.F., Wild, C.J., Kalbfleisch, J.D. (1999). Estimation for response-selective and missing data problems in regression. *Journal of the Royal Statistical Society, Series B* **61**, 413–438.

Lee, A.J., Scott, A.J., Wild, C.J. (2006). Fitting binary regression models with case-augmented samples. *Lifetime Data Analysis* **13**, 385–397.

Lee, A.J., Scott, A.J., Wild, C.J. (2007). On the Breslow-Holubkov estimatot. *Biometrika* **93**, 545–563.

Lee, H. (1990). Estimation of panel correlations for the Canadian labour force survey. *Survey Methodology* **16**, 283–292.

Le Guennec, J., Sautory, O. (2003). *La macro Calmar2*. Manuel d'utilisation. Internal document of INSEE.

Lehmann, E.L., Casella, G. (1998). *Theory of Point Estimation* (2nd ed.). Springer-Verlag, New York.

Lehtonen, R., Pahkinen, E. (2004). *Practical Methods for Design and Analysis of Complex Surveys* (2nd ed.). John Wiley & Sons, Chichester, UK.

Lehtonen, R., Särndal, C.-E., Veijanen, A. (2003). The effect of model choice in estimation for domains, including small domains. *Survey Methodology* **29**, 33–44.

Lehtonen, R., Särndal, C.-E., Veijanen, A. (2005). Does the model matter? Comparing model-assisted and model-dependent estimators of class frequencies for domains. *Statistics in Transition* **7**, 649–673.

Lehtonen, R., Särndal, C.-E., Veijanen, A. (2008). Generalized regression and model-calibration estimation for domains: Accuracy comparison. Paper presented at workshop on Survey Sampling Theory and Methodology, 25–29 August 2008, Kuressaare, Estonia. Available at: http://www.ms.ut.ee/samp2008/present.html.

Lehtonen, R., Veijanen, A. (1999). Domain estimation with logistic generalized regression and related estimators. *Proceedings, IASS Satellite Conference on Small Area Estimation,* Riga, Latvian Council of Science, 121–128.

Lehtonen, R., Veijanen, A. (1998). Logistic generalized regression estimators. *Survey Methodology* **24**, 51–55.

Lent, J., Miller, S., Cantwell, P. (1996). Effect of composite weights on some estimates for the current population survey. *Proceeding of the Survey Research Methods Section,* American Statistical Association, Alexandria, VA, 130–139.

Lepkowski, J.M. (1989). Treatment of wave nonresponse in panel surveys. In: Kasprzyk, D., Duncan, G., Kalton, G., Singh, M.P. (Eds.), *Panel Surveys*. Wiley, New York, pp. 348–374.

Liang, K.-Y., Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13–22.

Li, H. (2007). Small area estimation: an empirical best linear unbiased prediction approach. Ph.D. dissertation, Department of Mathematics, University of Maryland, College Park, MD.

Linacre, S., Zarb, J. (1991). Picking turning points in the Economy. *Australian Economic Indicators*. Catalogue no. 1350. Australian Bureau of Statistics, Canberra, Australia.

Lin, D.Y. (2000). On fitting Cox's proportional hazards models to survey data. *Biometrika* **87**, 37–47.

Lin, H.-M., Hughes, M.D. (1997). Assessing the effects of interventions using longitudinal data with samples subject to selection. *Biometrics* **53**, 924–936.

Little, R.J. (1991). Inference with survey weights. *Journal of Official Statistics* **7**, 405–424.

Little, R.J.A. (1982). Models for non-response in sample surveys. *Journal of the American Statistical Association* **77**, 237–249.

Little, R.J.A. (1995). Modeling the dropout mechanism in repeated measures studies. *Journal of the American Statistical Association* **90**, 1112–1121.

Little, R.J.A. (2004). To model or not to model? Competing modes of inference for finite population sampling. *Journal of the American Statistical Association* **99**, 546–556.

Little, R.J.A., Rubin, D.B. (2002). *Statistical Analysis with Missing Data* (2nd ed.). Wiley, New York.

Lohr, S.L. (1999). *Sampling: Design and Analysis*. Duxbury Press, Pacific Grove, CA.

Lohr, S., Rao, J.N.K. (2003). Resampling methods for MSE estimation with nonlinear small area models. *Proceedings of Statistics Canada Symposium*.

Lombardía, M.-J., González-Manteiga, W., Prada-Sanchez, J.-M. (2003). Bootstrapping the Chambers-Dunstan estimate of a finite population distribution function. *Journal of Statistical Planning and Inference* **116**, 367–388.

Lombardía, M.-J., González-Manteiga, W., Prada-Sanchez, J.-M. (2004). Bootstrapping the Dorfman-Hall-Chambers-Dunstan estimate of a finite population distribution function. *Journal of Nonparametric Statistics* **16**(1–2), 63–90.

Lombardía, M.-J., González-Manteiga, W., Prada-Sanchez, J.-M. (2005). Estimation of a finite population distribution function based on a linear model with unknown errors. *The Canadian Journal of Statistics* **33**(2), 181–200.

Longford, N.T. (2005). *Missing Data and Small-Area Estimation*. Springer.

Lundström, S., Särndal, C.-E. (1999). Calibration as a standard method for the treatment of nonresponse, *Journal of Official Statistics* **15**, 305–327.

MacGibbon, B., Tomberlin, T.J. (1989). Small area estimates of proportions via empirical Bayes techniques. *Survey Methodology* **15**, 237–252.

Madow, W.G. (1948). On the limiting distributions of estimates based on samples from finite universe. *The Annals of Mathematical Statistics* **19**, 535–545.

Madow, W.G. (1949). On the theory of systematic sampling, II. *The Annals of Mathematical Statistics* **20**, 333–354.

Magee, L. (1998). Improving survey-weighted least squares regression. *Journal of the Royal Statistical Society, Series B* **60**, 115–126.

Mahalanobis, P.C. (1946). Recent experiments in statistical sampling in the Indian Statistical Institute. *Journal of the Royal Statistical Society* **109**, 325–378.

Majumdar, H., Sen, P.K. (1978). Invariance principles for jackknifing U-statistics for finite population sampling and some applications. *Communications in Statistics: Theory and Methods A* **7**, 1007–1025.

Mak, T.K., Kuk, A. (1993). A new method for estimating finite-population quantiles using auxiliary information. *The Canadian Journal of Statistics* **21**(1),29–38.

Malec, D., Davis, W.W., Cao, X. (1999). Model-based small area estimates of overweight prevalence using sample selection adjustment. *Statistics in Medicine* **18**, 3189–3200.

Malec, D., Sedransk, J. (1985). Bayesian inference for finite population parameters in multistage cluster sampling. *Journal of the American Statistical Association* **80**, 897–902.

Malec, D., Sedransk, J., Moriarity, C.L., LeClere, F.B. (1997). Small area inference for binary variables in the National Health Interview Survey. *Journal of the American Statistical Association* **92**, 815–826.

Malec, D., Sedransk, J., Tompkins, L. (1993). Bayesian predictive inference for small areas for binary variables in the National Health Interview Survey. In: Gatsonis, C., Hodges, J.S., Kass, R.E., Singpurwalla, N.D. (Eds.), *Case Studies in Bayesian Statistics*. Springer-Verlag, New York, pp. 377–389.

Manly, B., McDonald, L., Thomas, D. (2002). *Resource Selection by Animals*. Chapman & Hall, London.

Manski, C.F. (2001). Daniel McFadden and the analysis of discrete choice. *Scandinavian Journal of Economics* **103**, 217–229.

Manski, C.F., McFadden, D. (Eds.). (1981). *Structural Analysis of Discrete Data with Econometric Applications*. Wiley, New York.

Maravall, A. (1985). On structural time series models and the characterization of components. *Journal of Business and Economic Statistics* **3**, 350–355.

Marker, D.A. (2001). Producing small area estimates from national surveys: methods for minimizing use of indirect estimators. *Survey Methodology* **27**, 183–188.

Martuzzi, M., Elliott, P. (1996). Empirical Bayes estimation of small area prevalence of non-rare conditions. *Statistics in Medicine* **15**, 1867–1873.

Mayor, J.A. (2002). Optimal cluster selection probabilities to estimate the finite population distribution function under PPS cluster sampling. *Sociedad de Estadística e Investigación Operativa TEST* **11**(1), 73–88.

McCarthy, P.J. (1969). Pseudo-replication: half samples. *Review of International Statistical Institute* **37**, 239–264.

McCarthy, P.J., Snowden, C.B. (1985). The bootstrap and finite population sampling. In: Hyattsville, Md, *Vital and Health Statistics* (Series 2, No. 95). Public Health Service Publication 85-1369, U.S. Government Printing Office, Washington, DC.

McCullagh, P., Nelder, J.A. (1989). *Generalized Linear Models* (2nd ed.). Chapman and Hall, London.

McCullagh, P., Nelder, J.A. (1999). *Generalized Linear Models*. Chapman & Hall, London.

McCulloch, C.E., Searle, S.R. (2001). *Generalized, Linear, and Mixed Models*. John Wiley & Sons, New York.

McDowell, A., Engel, A., Massey, J.T., Maurer, K. (1981). Plan and operation of the second National Health and Nutrition Examination Survey, 1976–80. *Vital and Health Statistics,* Series 11, No. 15, National Center for Health Statistics, Washington, DC.

McLaren, C.H. (1999). Designing rotation patterns and filters for trend estimation in repeated surveys. Unpublished PhD thesis. University of Wollongong, Wollongong, Australia.

McLaren, C.H., Steel, D.G. (2000). The impact of different rotation patterns on the sampling variance of seasonally adjusted and trend estimates. *Survey Methodology* **26**, 163–172.

McLaren, C.H., Steel, D.G. (2001). Rotation patterns and trend estimation for repeated surveys using rotation group estimates. *Statistica Neerlandica* **55**, 221–238.

McLeish, D.L. (1984). Estimation for aggregate models: the aggregate Markov chain. *The Canadian Journal of Statistics* **12**, 265–282.

Meeden, G. (1992). Basu's contributions to the foundations of sample survey. In: Ghosh, M., Pathak, P.K. (Eds.), *IMS Lecture Notes Monograph Series*, vol. 17. pp. 178–186.

Meeden, G. (1995). Median estimation using auxiliary information. *Survey Methodology* **21**(1), 71–77.

Meeden, G., Ghosh, M. (1983). Chosing between experiments: applications to finite population sampling. *The Annals of Statistics* **11**, 296–305.

Meza, J., Chen, S., Lahiri, P. (2003). Estimation of lifetime alcohol abuse for Nebraska counties. Unpublished manuscript.

Midzuno, H. (1952). On the sampling system with probability proportionate to sum of sizes. *Annals of the Institute of Statistical Mathematics* **3**, 99–107.

Miettinen, O.S. (1985). The case-control study: valid selection of subjects. *American Journal of Epidemiology* **135**, 1042–1050.

Milbrodt, H. (1987). A note on Hájek's theory of rejective sampling. *Metrika* **34**, 275–281.

Millar, R.B. (1992). Estimating the size selectivity of fishing gear by conditioning on the total catch. *Journal of the American Statistical Association* **87**, 962–967.

Miller, M.E., Ten Have, T.R., Reboussin, B.A., Lohman, K.K., Rejeski, W.J. (2001). A marginal model for analyzing discrete outcomes from longitudinal surveys with outcomes subject to multiple-cause nonresponse. *Journal of the American Statistical Association* **96**, 844–857.

Miller, R.G. (1974a). An unbalanced jackknife. *The Annals of Statistics* **2,** 880–891.

Miller, R.G. (1974b). The jackknife – a review. *Biometrika* **61**(1), 1–15.

Miller, R.G. Jr., Sen, P.K. (1972). Weak convergence of U-statistics and von Mises's differentiable statistical functions. *The Annals of Mathematical Statistics* **43**, 31–41.

Modarres, R. (2002). Efficient nonparametric estimation of a distribution function. *Computational Statistics and Data Analysis* **39**, 75–95.

Mohamed, W.N., Diamond, I., Smith, P.W.F. (1998). The determinants of infant mortality in Malaysia: a graphical chain modelling approach. *Journal of the Royal Statistical Society* **A161**, 349–366.

Molina, E.A., Smith, T.M.F., Sugden, R.A. (2001). Modelling overdispersion for complex survey data. *International Statistical Review* **69**, 373–384.

Molina, I., Saei, A., Lombardia, M.J. (2007). Small area estimates of labour force participation under a multinomial logit mixed model. *Journal of the Royal Statistical Society Series (A)* **170**, 975–1000.

Montanari, G.E. (1987). Post-sampling efficient qr-prediction in large-scale surveys. *International Statistical Review* **55**, 191–202.

Montanari, G.E., Ranalli, M.G. (2002). Asymptotically efficient generalized regression estimators. *Journal of Official Statistics* **18**, 577–589.

Montanari, G.E., Ranalli, M.G. (2005). Nonparametric model calibration estimation in survey sampling. *Journal of the American Statistical Association* **100**(472), 1429–1442.

Moore, D.S., Spruill, M.C. (1975). Unified large sample theory of general chi-squared statistics for tests of fit. *The Annals of Statistics* **3**, 599–616.

Morris, C. (1982). Natural exponential families with quadratic variance functions. *The Annals of Statistics* **10**, 65–80.

Morris, C. (1983a). Natural exponential families with quadratic variance functions: statistical theory. *The Annals of Statistics* **11**, 515–529.

Morris, C. (1983b). Parametric empirical Bayes confidence intervals. In: Box, G.E.P., Leonard, T., Wu, C.F.J. (Eds.), *Scientific Inference, Data Analysis and Robustness*. Academic Press, New York, pp. 25–50.

Morris, C. (1983c). Parametric empirical Bayes inference: theory and applications. *Journal of the American Statistical Association* **78**, 47–54.

Mukhopadhyay, P. (1998). *Theory and Methods of Survey Sampling*. Prentice-Hall of India, New Delhi, India.

Mukhopadhay, P., Maiti, T. (2004). Two-stage nonparametric approach for small area estimation. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, Alexandria, VA, 4058–4065.

Murthy, M.M., Sethi, V.K. (1965). Self-weighting design at tabulation stage. *Sankhya, Series B* **27**, 201–210.

Nandi, H.K., Sen, P.K. (1963). On the properties of U-statistics when the observations are not independent. Part two: unbiased estimation of the parameters of a finite population. *Calcutta Statistical Association Bulletin* **12**, 125–148.

Nandram, B. (1999). An empirical Bayes prediction interval for the finite population mean of small area. *Statistica Sinica* **9**, 325–343.

Nandram, B., Sedransk, J. (1993). Bayesian predictive inference for a finite population proportion: two-stage cluster sampling. *Journal of the Royal Statistical Society, Series B* **55**, 399–408.

Narain, R.D. (1951). On sampling without replacement with varying probabilities. *Journal of the Indian Society of Agricultural Statistics* **3**, 169–175.

Nascimento Silva, P.L.D., Skinner, C.J. (1995). Estimating distribution functions with auxiliary information using poststratification. *Journal of Official Statistics* **11**(3), 277–294.

Nathan, G. (1999). *A Review of Sample Attrition and Representativeness in Three Longitudinal Surveys*. GSS Methodology Series No. 13. Office of National Statistics, London.

Nathan, G., Holt, D. (1980). The effect of survey design on regression analysis. *Journal of the Royal Statistical Society, Series B* **42**, 377–386.

National Center for Health Statistics. (1968). *Synthetic State Estimates of Disability*. P.H.S. Publications 1759, U.S. Government Printing Office, Washington, DC.

National Center for Health Statistics. (1976). NCHS growth charts, 1976. *Monthly Vital Statistics Report*, vol. 25, No. 3, Suppl. (HRA) 76-1120. Health Resources Administration, Rockville, MD.

Nelder, J.A., Wedderburn, R.W.M. (1972). Generalized linear models. *Journal of the Royal Statistical Society, Series A* **135**, 370–384.

Neuhaus, J., Scott, A.J., Wild, C.J. (2002). The analysis of retrospective family studies. *Biometrika* **89**, 23–37.

Neuhaus, J., Scott, A.J., Wild, C.J. (2006). Family-specific approaches to the analysis of retrospective family data. *Biometrics* **62**, 488–494.

Newcombe, R.G. (2001). Logit confidence interval and the inverse sinh transformation. *The American Statistician* **55**, 200–202.

Neyman, J. (1934). On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society* **97**(4), 558–625.

Ng, M.P., Donadio, M. (2006). Computing inclusion probabilities for order sampling. *Journal of Statistical Planning and Inference* **136**(11), 4026–4042.

Niyonsenga, T. (1994). Nonparametric estimation of response probabilities in sampling theory. *Survey Methodology* **20**, 177–184.

Niyonsenga, T. (1997). Response probability estimation. *Journal of Statistical Planning and Inference* **59**, 111–126.

Nordberg, L. (1989). Generalized linear modeling of sample survey data. *Journal of Official Statistics* **5**, 223–239.

Núñez-Antón, V., Zimmermann, D.L. (2000). Modeling nonstationary longitudinal data. *Biometrics* **56**, 699–705.

O'Hara Hines, R.J., Carter, E.M. (1993). Improved added variable and partial residual plots for the detection of influential observations in generalized linear models (with discussion). *Applied Statistics* **42**, 3–20.

Oh, H.L., Scheuren, F.J. (1983). Weighting adjustments for unit non-response. In: Madow, W.G., Olkin, I., Rubin, D.B. (Eds.), *Incomplete Data in Sample Surveys (Vol. 2): Theory and Bibliographies*. Academic Press, New York; London, pp. 143–184.

Ohlsson E. (1986). Asymptotic normality of the Rao-Hartley-Cochran estimator: an application of the martingale CLT. *Scandinavian Journal of Statistics*, *Theory and Applications* **13**, 17–28.

Ohlsson E. (1989a). Variance estimation in the Rao-Hartley-Cochran procedure. *Sankhya, Series B* **51**, 348–361.

Ohlsson E. (1989b). Asymptotic normality for two-stage sampling from a finite population. *Probability Theory and Related Fields* **81**, 341–352.

Ohlsson E. (1995). Sequential Poisson Sampling. Research Report, Institute of Actuarial Mathematics and Mathematical Statistics, Stockholm University, Stockholm, Sweden.

Opsomer, J.D., Breidt, F.J., Moisen, G.G., Kauermann, G. (2007). Model-assisted estimation of forest resources with generalized additive models. *Journal of the American Statistical Association* **102**, 400–416.

Opsomer, J.D., Claeskens, G., Ranalli, M.G., Kauermann, G., Breidt, F.J. (2008). Nonparametric small area estimation using penalised spline regression. *Journal of the Royal Statistical Society, Series B*, **70**, 265–286.

Opsomer, J.D., Miller, C.P. (2005). Selecting the amount of smoothing in nonparametric regression estimation for complex surveys. *Journal of Nonparametric Statistics* **17**, 593–611.

Orchard, T., Woodbury, M.A. (1972). A missing information principle: theory and application. *Proceedings of the 6th Berkeley Symposium on Mathematical Statistics and Probability* **1**, 697–715.

Owen, A.B. (1987). *Nonparametric Conditional Estimation*. Technical Report No. 265. Department of Statistics, Stanford University, Stanford, CA.

Owen, A.B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika* **75**, 237–249.

Owen, A.B. (2001). *Empirical Likelihood*. Chapman and Hall, New York.

Park, M., Fuller, W.A. (2005). Towards nonnegative regression weights for survey samples. *Survey Methodology* **31**, 85–93.

Park, Y.S., Kim, K.W., Choi, J.W. (2001). One-level rotation design balanced on time in monthly sample and in rotation group. *Journal of the American Statistical Association* **96**, 1483–1496.

Park, Y.S., Kim, K.W., Kim, N.Y. (2003). Three-way balanced multi-level rotation sampling designs. *Journal of the Korean Statistical Society* **32**, 245–259.

Patterson, H.D. (1950). Sampling on successive occasions with partial replacement of units. *Journal of the Royal Statistical Society, Series B* **12**, 241–255.

Petersen, C.G.J. (1896). The yearly immigration of young plaice into the Limfjord from the German sea. *Report of the Danish Biological Station* **6**, 1–48.

Petrucci, A., Pratesi, M., Salvati, N. (2005). Geographic information in small area estimation: small area models and spatially correlated random area effects. *Statistics in Transition* **7**, 609–623.

Pfeffermann, D. (1988). The effect of sampling design and response mechanism on multivariate regression-based predictors. *Journal of the American Statistical Association* **83**, 824–833.

Pfeffermann, D. (1991). Estimation and seasonal adjustment of population means using data from repeated surveys. *Journal of Business and Economic Statistics* **9**, 163–177 (with discussion).

Pfeffermann, D. (1993). The role of sampling weights when modeling survey data. *International Statistical Review* **61**, 317–337.

Pfeffermann, D. (1994). A general method for estimating the variances of X-11 seasonally adjusted estimators. *Journal of Time Series Analysis* **15**, 85–116.

Pfeffermann, D. (1996). The use of sampling weights for survey data analysis. *Statistical Methods in Medical Research* **5**, 239–261.

Pfeffermann, D. (2002). Small area estimation: new developments and directions. *International Statistical Review* **70**, 125–143.

Pfeffermann, D., Barnard, C.H. (1991). Some new estimators for small area means with application to the assessment of farmland values. *Journal of Business and Economic Statistics* **9**, 73–83.

Pfeffermann, D., Burck, L. (1990). Robust small area estimation combining time series and cross-sectional data. *Survey Methodology* **16**, 217–237.

Pfeffermann, D., Feder, M., Signorelli, D. (1998). Estimation of autocorrelations of survey errors with application to trend estimation in small areas. *Journal of Business & Economic Statistics* **16**, 339–348.

Pfeffermann, D., Glickman, H. (2004). Mean square error approximation in small area estimation by use of parametric and nonparametric bootstrap. *Proceedings of the Survey Research Methods Section,* American Statistical Association.

Pfeffermann, D., Holmes, D. (1985). Robustness considerations in the choice of a method of inference for regression analysis of survey data. *Journal of the Royal Statistical Society, Series A* **198**, 268–278.

Pfeffermann, D., Krieger, A.M., Rinott, Y. (1998a). Parametric distributions of complex survey data under informative probability sampling. *Statistica Sinica* **8**, 1087–1114.

Pfeffermann, D., Moura, F.A.S., Nascimento Silva, P.L.A. (2006). Multilevel modeling under informative sampling. *Biometrika* **93**, 943–959.

Pfeffermann, D., Nathan, G. (1985). Problems in model identification based on data from complex sample surveys. *Bulletin of the International Statistical Institute* **51**, 12.2.1–12.2.17.

Pfeffermann, D., Nathan, G. (2001). Imputation for wave nonresponse – existing methods and a time series approach. In: Groves, R., Dillman, D., Eltinge, J., Little, R. (Eds.), *Survey Nonresponse*, Chapter 28. Wiley, New-York, pp. 417–429.

Pfeffermann, D., Skinner, C.J., Holmes, D.J., Goldstein, H., Rasbash, J. (1998b). Weighting for unequal selection probabilities in multi-level models (with discussion). *Journal of the Royal Statistical Society, Series B* **60**, 23–76.

Pfeffermann, D., Skinner, C.J., Humphreys, K. (1998c). The estimation of gross flows in the presence of measurement errors using auxiliary variables. *Journal of the Royal Statistical Society, Series A* **161**, 12–32.

Pfeffermann, D., Sverchkov, M. (1999). Parametric and semi-parametric estimation of regression models fitted to survey data. *Sankhya, Series B* **61**(Pt. 1): 166–186.

Pfeffermann, D., Sverchkov, M. (2003). Fitting generalized linear models under informative probability sampling. In: Chambers, R.L., Skinner, C.J. (Eds.), *Analysis of Survey Data*. Wiley, New York, pp. 175–195.

Pfeffermann, D., Sverchkov, M. (2005). Small area estimation under informative sampling. *Statistics in Transition* **7**, 675–684.

Pfeffermann, D., Sverchkov, M. (2007). Small area estimation under informative probability sampling of areas and within the selected areas. *Journal of the American Statistical Association* **102**, 1427–1439.

Pfeffermann, D., Tiller, R. (2005). Bootstrap approximation to prediction MSE for state space models with estimated parameters. *Journal of Time Series Analysis* **26**, 893–916.

Pfeffermann, D., Tiller, R. (2006). Small-area estimation with state–space models subject to benchmark constraints. *Journal of the American Statistical Association* **101**, 1387–1397.

Pirkle, J.L., Schwartz, J., Landis, J.R., Harlan, W.R. (1985). The relationship between blood lead levels and blood pressure and its cardiovascular risk implications. *American Journal of Epidemiology* **121**, 246–258.

Poterba, J.M., Summers, L.H. (1986). Responding error and labor market dynamics. *Econometrica* **54**, 1319–1338.

Prasad, N.G.N., Rao, J.N.K. (1990). The estimation of mean squared errors of small area estimators. *Journal of the American Statistical Association* **85**, 163–171.

Prasad, N.G.N., Rao, J.N.K. (1999). On robust small area estimation using a simple random effects model. *Survey Methodology* **25**, 67–72.

Prášková, Z. (1982). Rate of convergence for simple estimate in rejective sampling. In: Grossmann, W., Pflug, G., Wertz, W. (Eds.), *Probability and Statistical Inference*. Reidel, Dordrecht, The Netherlands, pp. 307–317.

Prášková, Z. (1984). On the rate of convergence in Sampford-Durbin sampling from a finite population. *Statistics & Decision* **2**, 339–350.

Prášková, Z. (1988). On the convergence to the Poisson distribution in rejective sampling from a finite population. In: Grossmann, W., Mogyoródy, J., Vincze, I., Wertz, W. (Eds), Probability Theory and Mathematical Statistcs with Applications. Reidel, Dordrecht, The Netherlands, pp. 285–294.

Prášková, Z. (1995). On Hájek's conjecture in stratified sampling. *Kybernetika* **31**, 303–314.

Prášková, Z., Sen, P.K. (1998). The Hájek perspectives in finite population sampling. In: Hušková, M., Beran, R., Dupač, V. (Eds.), *Collected Works of Jaroslav Hájek - With Commentaries*. J. Wiley, New York pp. 37–43.

Pregibon, D. (1981). Logistic regression diagnostics. *The Annals of Statistics* **9**, 705–724.

Prentice, R.L. (1986). A case-cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika* **73**, 1–11.

Prentice, R.L., Pyke, R. (1979). Logistic disease incidence models and case-control studies. *Biometrika* **66**, 403–411.

Purcell, N.J., Kish, L. (1980). Postcensal estimates for local areas (or domains). *International Statistical Review* **48**, 3–18.

Qin, J., Lawless, J. (1994). Empirical likelihood and general estimating equations. *The Annals of Statistics* **22**, 300–325.

Qin, Y., Rao, J.N.K., Ren, Q. (2006). Confidence intervals for parameters of the response variable in a linear model with missing data. Technical Report, Laboratory for Research in Statistics and Probability, Carleton University, Ottawa, Canada.

Quenouille, M.H. (1949a). Approximate tests of correlation in time series. *Journal of the Royal Statistical Society, Series B: Statistical Methodology* **11**, 18–84.

Quenouille, M.H. (1949b). Problems in plane sampling. *The Annals of Mathematical Statistics* **20**, 335–375.

Quenouille, M.H. (1956). Notes on bias in estimation. *Biometrika* **43**, 353–360.

Rabe-Hesketh, S., Skrondal, A. (2006). Multilevel modeling of complex survey data. *Journal of the Royal Statistical Society A* **169**, 805–827.

Raftery, A.E. (1996). Hypothesis testing and model selection. In: Gilks, W.R., Spigelhalter, D., Richardson, S. (Eds.), *Markov Chair Monte Carlo is Practice*. Chrpman and Hall, London, pp. 163–188.

Raghunathan, T.E. (1993). A quasi-empirical Bayes method for small area estimation. *Journal of the American Statistical Association* **88**, 1444–1448.

Ramakrishnan, M.K. (1973). An alternative proof of the admissibility of the Horvitz-Thompson estimator. *The Annals of Statistics* **1**, 577–579.

Rao, C.R. (1973). *Linear Statistical Inference and its Applications*. Wiley, New York.

Rao Jammalamadaka, S., Michaletzky, G., Todorovic, P. (1991). Large sample distribution of the sample total in a generalized rejective sampling scheme. *Statistics and Probability Letters* **11**, 463–468.

Rao, J.N.K. (1971). Some thoughts on the foundations of survey sampling. *Journal of Indian Society of Agricultural Statistics* **23**, 69–82.

Rao, J.N.K. (1973). On double sampling for stratification and analytical surveys. *Biometrika* **60**, 125–133.

Rao, J.N.K. (1985). Conditional inference in survey sampling. *Survey Methodology* **11**, 15–31.

Rao, J.N.K. (1994). Estimating totals and distribution functions using auxiliary information at the estimation stage. *Journal of Official Statistics* **10**, 153–165.

Rao, J.N.K. (1996). On variance estimation with imputed survey data. *Journal of American Statistical Association* **91**, 499–506.

Rao, J.N.K. (1997). Developments in sample survey theory: an appraisal. *The Canadian Journal of Statistics* **25**, 1–21.

Rao, J.N.K. (1999). Some recent advances in model-based small area estimation. *Survey Methodology* **25**, 175–186.

Rao, J.N.K. (2003a). *Small Area Estimation*. John Wiley & Sons, Hoboken, New Jersey.

Rao, J.N.K. (2003b). Some new developments in small area estimation. *Journal of the Iranian Statistical Society* **2**, 145–169.

Rao, J.N.K. (2005a). Interplay between sample survey theory and practice: an appraisal. *Survey Methodology* **31**, 117–138.

Rao, J.N.K. (2005b). Inferential issues in small area estimation: some new developments. *Statistics in Transition* **7**, 513–526.

Rao, J.N.K. (2006). Bootstrap methods for analyzing complex sample survey data. To appear in *Proceedings of Statistics Canada Symposium 2006: Methodological Issues in Measuring Population Health*. Available at: http://www.statcan.gc.ca/pub/11-522-x/11-522-x2006001-eng.htm.

Rao, J.N.K., Ghangurde, P.D. (1972). Bayesian optimization in sampling finite populations. *Journal of the American Statistical Association* **67**, 439–443.

Rao, J.N.K., Graham, J.E. (1964). Rotation designs for sampling on repeated occasions. *Journal of the American Statistical Association* **59**, 492–509.

Rao, J.N.K., Hartley, H.O., Cochran, W.G. (1962). On a simple procedure of unequal probability sampling without replacement. *Journal of the Royal Statistical Society, Series B: Statistical Methodology* **24**(2), 482–491.

Rao, J.N.K., Kovar, J.G., Mantel, H.J. (1990). On estimating distribution functions and quantiles from survey data using auxiliary information. *Biometrika* **77**(2), 365–375.

Rao, J.N.K., Scott, A.J. (1984). On chi-squared tests for multiway tables with cell proportions estimated from survey data. *The Annals of Statistics* **12**, 46–60.

Rao, J.N.K., Scott, A.J. (1992). A simple method for the analysis of clustered binary data. *Biometrics* **48**, 577–585.

Rao, J.N.K., Scott, A.J. (1999). A simple method for analyzing over-dispersion in clustered Poisson data. S*tatistics in Medicine* **18**, 1373–1385.

Rao, J.N.K., Scott, A.J., Skinner, C.J. (1998). Quasi-score tests with survey data. *Statistica Sinica* **8**, 1059–1070.

Rao, J.N.K., Shao, J. (1992). Jackknife variance estimation with survey data under hot deck imputation. *Biometrika* **79**(4), 811–822.

Rao, J.N.K., Shao, J. (1996). On balanced half-sample variance estimation in stratified random sampling. *Journal of American Statistical Association* **91**, 343–348.

Rao, J.N.K., Shao, J. (1999). Modified balanced repeated replication for complex survey data. *Biometrika* **86**(2), 403–415.

Rao, J.N.K., Singh, A.C. (1997). A ridge-shrinkage method for range-restricted weight calibration in survey sampling. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, Washington, DC, 57–65.

Rao, J.N.K., Sitter, R.R. (1995). Variance estimation under two-phase sampling with application to imputation for missing data. *Biometrika* **82**(2), 453–460.

Rao, J.N.K., Tausi, M. (2004). Estimating function jackknife variance estimators under stratified multistage sampling. *Communications in Statistics. Theory and Methods* **33**, 2087–2095.

Rao, J.N.K., Thomas, D.R. (1989). Chi-squared tests for contingency tables. In: Skinner, C.J., Holt, D., Smith, T.M.F. (Eds.), *Analysis of Complex Surveys.* Wiley, Chichester, UK, pp. 89–114.

Rao, J.N.K., Thomas, D.R. (2003). Analysis of categorical response data from complex surveys: an appraisal and update. In: Chambers, R.L., Skinner, C.J. (Eds.), *Analysis of Survey Data*. John Wiley, New York, pp. 85–108.

Rao, J.N.K., Vijayan, K. (1977). On estimating the variance in sampling with probability proportional to aggregate size. *Journal of the American Statistical Association* **72**, 579–584.
Rao, J.N.K., Wu, C. (2005). Empirical likelihood approach to calibration using survey data. *Proceedings of 55th Session of International Statistics Institute*, Sydney, Australia.
Rao, J.N.K., Wu, C.F.J. (1985). Inference from stratified samples: second-order analysis of three methods for nonlinear statistics, *Journal of American Statistical Association* **80**, 620–630.
Rao, J.N.K., Wu, C.F.J. (1988). Resampling inference with complex survey data. *Journal of the American Statistical Association* **83**, 231–241.
Rao, J.N.K., Wu, C.F.J., Yue, K. (1992). Some recent work on resampling methods for complex surveys. *Survey Methodology* **18**, 209–217.
Rao, J.N.K., Yu, M. (1992). Small area estimation by combining time series and cross-sectional data. *Proceedings of the Survey Research Section*, American Statistical Association, 1–9.
Rao, J.N.K., Yu, M. (1994). Small area estimation by combining time series and cross-sectional data. *The Canadian Journal of Statistics* **22**, 511–528.
Rao, J.N.K., Yung, W., Hidiroglou, M.A. (2002). Estimating equations for the analysis of survey data using poststratification information. *Sankhya* **64**, 364–378.
Rao, K.C., Robson, D.S. (1974). A chi-square statistic for goodness-of-fit tests within the exponential family. *Communications in Statistics* **3**, 1139–1153.
Ripley, B.D. (1996). *Pattern Recognition and Neural Networks*. Cambridge University Press, Cambridge, UK.
Rivest, L.P., Belmonte, E. (1999). The conditional mean square errors of small area estimators in survey sampling. Technical Report, Laval University, Quebec, Canada.
Rivest, L.-P., Belmonte, E. (2000). A conditional mean squared error of small area estimators. *Survey Methodology* **26**, 67–78.
Rivest, L.-P., Vandal, N. (2003). Mean squared error estimation for small areas when the small area variances are estimated. *Proceedings of the International Conference on Recent Advances in Survey Sampling*, Laboratory for Research in Statistics and Probability, Carleton University, Ottawa, ON, Canada.
Roberts, G.A., Ren, Q., Rao, J.N.K. (2009). Using marginal mean models with data from longitudinal surveys with a complex design: some advances in methods. In: Lynn, P. (Ed.), *Methodology of Longitudinal Surveys*. Wiley: New York.
Roberts, G.R., Rao, J.N.K., Kumar, S. (1987). Logistic regression analysis of sample survey data. *Biometrika* **74**, 1–12.
Roberts, G.R., Ren, Q., Rao, J.N.K. (2008). Marginal models for longitudinal surveys. In: Lynn, P. (Ed.), *Methodology of Longitudinal Surveys*. Wiley, New York.
Robins, J.M., Rotnitzky, A., Zhao, L.P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association* **90**, 106–121.
Rosén, B. (1970). On the coupon collector's waiting time. *The Annals of Mathematical Statistics* **41**, 1952–1969.
Rosén, B. (1972a). Asymptotic theory for successive sampling with varying probabilities without replacement, I. *The Annals of Mathematical Statistics* **43**, 373–397.
Rosén, B. (1972b). Asymptotic theory for successive sampling with varying probabilities without replacement, II. *The Annals of Mathematical Statistics* **43**, 748–776.
Rosén, B. (1974). Asymptotic theory for Des Raj estimators. In: Hájek, J. (Ed.), *Proceedings of the Prague Conference on Asymptotic Statistics*, vol. I, Charles University, Prague, Czech Republic, pp. 313–330.
Rosén, B. (1997a). Asymptotic theory for order sampling. *Journal of Statistical Planning and Inference* **62**, 135–158.
Rosén, B. (1997b). On sampling with probability proportional to size. *Journal of Statistical Planning and Inference* **62**, 159–191.
Rosén, B. (2000). On inclusion probabilities for order πps sampling. *Journal of Statistical Planning and Inference* **90**, 117–143.
Rothman, K.J., Greenland, S. (1998). *Modern Epidemiology* (2nd ed.). Lippincott-Raven, Philadelphia, PA.
Rotnitzky, A., Robins, J.M., Scharfstein, D.O. (1998). Semiparametric regression for repeated outcomes with nonignorable nonresponse. *Journal of the American Statistical Association* **93**, 1321–1339.
Royall, R.M. (1968). An old approach to finite population sampling theory. *Journal of the American Statistical Association* **63**, 1269–1279.
Royall, R.M. (1970). On finite population sampling theory under certain linear regression models. *Biometrika* **57**, 377–387.

Royall, R.M. (1970a). Finite population sampling - On labels in estimation. *The Annals of Mathematical Statistics* **41**, 1774–1779.

Royall, R.M. (1971). Linear regression models in finite population sampling theory. In: Godambe, V.P., Sprott, D.A. (Eds.), *Foundations of Statistical Inference*. Holt, Rinehart and Winston, Toronto, ON, Canada, pp. 259–274.

Royall, R.M. (1976a). The linear least-squares prediction approach to two-stage sampling. *Journal of the American Statistical Association* **71**, 657–664.

Royall, R.M. (1976b). Current advances in sampling theory: implications for human observational studies. *American Journal of Epidemiology* **104**, 463–473.

Royall, R.M. (1986). The prediction approach to robust variance estimation in two-stage cluster sampling. *Journal of the American Statistical Association* **81**, 119–123.

Royall, R.M. (1992). Robustness and optimal design under prediction models for finite populations. *Survey Methodology* **18**, 179–185.

Royall, R.M. (1994). Discussion of "sample surveys 1975–1990; an age of reconciliation?" by T.M.F. Smith. *International Statistical Review* **62**, 19–21.

Royall, R.M., Cumberland, W.G. (1978). Variance estimation in finite population sampling. *Journal of the American Statistical Association* **73**, 351–358.

Royall, R.M., Cumberland, W.G. (1981). An empirical study of the ratio estimator and estimators of its variance. *Journal of the American Statistical Association* **76**, 66–77.

Rubin, D.B. (1976). Inference and missing data. *Biometrika* **63**, 605–614.

Rubin, D.B. (1978). Multiple imputation in sample surveys – a phenomenological Bayesian approach to nonresponse. *Proceedings of Survey Research Methods Section,* American Statistical Association, 20–34.

Rubin, D.B. (1985). The use of propensity scores in applied Bayesian inference. In: Bernardo, J.M., Degroot, M.H., Lindley, D.V., Smith, A.F.M. (Eds.), *Bayesian Statistics, Volume 2*. Elsevier Science Publishers B.V., Amsterdam: North-Holland, pp. 463–472.

Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. John Wiley, New York.

Rubin, D.B. (1996). Multiple imputation after 18+ years (with discussion). *Journal of the American Statistical Association* **91**, 473–489.

Rubin-Bleuer, S., Schiopu Kratina, I. (2005). On the two-phase framework for joint model and design-based inference. *The Annals of Statistics* **33**, 2789–2810.

Rubinstein, A. (1991). Comments on the interpretation of game theory. *Econometrica* **59**, 909–924.

Rueda, M., Martinez, S., Martinez, H., Arcos, A. (2007a). Estimation of the distribution function with calibration methods. *Journal of Statistical Planning and Inference* **137**(2), 435–448.

Rueda, M., Martinez, S., Sánchez, I. (2007b). Estimation of the Distribution Function using Nonparametric Regression, Technical Report, University of Granada.

Ruppert, D., Sheather, S.J., Wand, M.P. (1995). An effective bandwidth selector for local least squares regression. *Journal of the American Statistical Association* **90**, 1257–1270.

Ruppert, D., Wand, M.P., Carroll, R.J. (2003). *Semiparametric Regression*. Cambridge University Press, Cambridge, UK.

Rust, K.F. (1985). Variance estimation for complex estimators in sample surveys. *Journal of Official Statistics* **1**, 381–397.

Rust, K.F., Rao, J.N.K. (1996). Variance estimation for complex surveys using replication techniques. *Statistical Methods in Medical Research* **5**, 283–310.

Saei, A., Chambers, R. (2004). Small area estimation under linear and generalized linear mixed models with time and area effects. EURAREA Consortium 2004, *Project Reference Volume*. Available at: www.statistics.gov.uk/eurarea.

Saigo, H., Shao, J., Sitter, R.R. (2001). A repeated half-sample bootstrap and balanced repeated replications for randomly imputed data. *Survey Methodology* **27**, 189–196.

Sampford, M.R. (1967). On sampling without replacement with unequal probabilities of selection. *Biometrika* **54**, 499–513.

Samuel, E. (1968). Sequential maximum likelihood estimation of the size of a population. *The Annals of Mathematical Statistics* **39**, 1057–1068.

Sánchez, B.N., Budtz-Jorgensen, E., Ryan, L.M., Hu, H. (2005). Structural equation models: A review with applications to environmental epidemiology. *Journal of the American Statistical Association* **100**, 1443–1455.

Sanderson, M., Placek, P.J., Keppel, K.G. (1991). The 1988 national maternal and infant health survey: design, content, and data availability. *Birth* **18**, 26–32.

Särndal, C.-E. (1980). On π-inverse weighting versus best linear unbiased weighting in probability sampling. *Biometrika* **67**, 639–650.

Särndal, C.-E. (1981). Frameworks for inference in survey sampling with applications to small area estimation and adjustment for nonresponse. *Bulletin of the International Statistical Institute* **49**, 494–513.

Särndal, C.-E. (1984). Design-consistent versus model-dependent estimation for small domains. *Journal of the American Statistical Association* **79**, 624–631.

Särndal, C.-E. (1996). Efficient estimators with simple variance in unequal probability sampling. *Journal of the American Statistical Association* **91**, 1289–1300.

Särndal, C.-E. (2001). Design-based methodologies for domain estimation. In: Lehtonen, R., Djerf, K. (Eds.), *Proceedings of the Symposium on Advances in Domain Estimation*. Reviews 2001/5. Statistics Finland.

Särndal, C.-E. (2007). The calibration approach in survey theory and practice. *Survey Methodology* **33**, 99–119.

Särndal, C.-E., Hidiroglou, M.A. (1989). Small domain estimation: a conditional analysis. *Journal of the American Statistical Association* **84**, 266–275.

Särndal, C.-E., Lundström, S. (2005). *Estimation in Surveys with Nonresponse*. John Wiley & Sons, Chichester, UK.

Särndal, C.-E., Swensson, B. (1987). A general view of estimation for two phases of selection with applications to two-phase sampling and nonresponse. *International Statistical Review* **55**, 279–294.

Särndal, C.-E., Swensson, B., Wretman, J.H. (1989). The weighted residual technique for estimating the variance of the general regression estimator of the finite population total. *Biometrika* **76**(3), 527–537.

Särndal, C.-E., Swensson, B., Wretman, J.H. (1992). *Model Assisted Survey Sampling*. Springer-Verlag, New York.

Särndal, C.-E., Wright, R. (1984). Cosmetic form of estimators in survey sampling. *Scandinavian Journal of Statistics* **11**, 146–156.

SAS. (1990). *SAS/GRAPH Software: Reference, Version 6* (1st ed., vol. 1). SAS Institute Inc, Cary, NC.

Schaible, W.L. (Ed.). (1996). *Indirect Estimation in U.S. Federal Programs.* Springer, New York.

Schaible, W.L., Brock, D.B., Schnack, G.A. (1977). An empirical comparison of the simple inflation, synthetic and composite estimators for small area statistics. *Proceedings of the Social Statistics Section*, American Statistical Association, 1017–1021.

Schnabel, Z.E. (1938). The estimation of the total fish population of a lake. *The American Mathematical Monthly* **45**, 348–352.

Scott, A.J. (1975). On admissibility and uniform admissibility in finite population sampling. *The Annals of Statistics* **3**, 489–491.

Scott, A.J. (1977). On the problem of randomization in survey sampling. *Sankhya Series C* **39**, 1–9.

Scott, A.J. (2006). Population-based case control studies. *Survey Methodology* **32**, 123–132.

Scott, A.J., Brewer, K.R.W., Ho, E.W.H. (1978). Finite population sampling and robust estimation. *Journal of the American Statistical Association* **73**, 359–361.

Scott, A.J., Smith, T.M.F. (1969). Estimation in multistage surveys. *Journal of the American Statistical Association* **64**, 830–840.

Scott, A.J., Smith, T.M.F. (1974). Analysis of repeated surveys using time series methods. *Journal of the American Statistical Association* **69**, 674–678.

Scott, A.J., Smith, T.M.F. (1975). Minimax designs for sample surveys. *Biometrika* **62**, 353–357.

Scott, A.J., Smith, T.M.F., Jones, R.G. (1977). The application of time series methods to the analysis of repeated surveys. *International Statistical Review* **45**, 13–28.

Scott, A.J., Wild, C.J. (1986). Fitting logistic models under case-control or choice-based sampling. *Journal of the Royal Statistical Society, Series B* **48**, 170–182.

Scott, A.J., Wild, C.J. (1991). Fitting logistic models in stratified case-control studies. *Biometrics* **47**, 497–510.

Scott, A.J., Wild, C.J. (1997). Fitting regression models to case-control data by maximum likelihood. *Biometrika* **84**, 57–71.

Scott, A.J., Wild, C.J. (2001a). The analysis of clustered case-control studies. *Applied Statistics* **50**, 57–71.

Scott, A.J., Wild, C.J. (2001b). Fitting regression models to case-control data by maximum likelihood. *Journal of Statistical Planning and Inference* **96**, 3–27.

Scott, A.J., Wild, C.J. (2002). On the robustness of weighted methods for fitting model to case-control data by maximum likelihood. *Journal of the Royal Statistical Society, Series B* **64**, 207–220.

Scott, A.J., Wild, C.J. (2007). Maximum likelihood methods for stratified cluster sampling with informative stratification. *Journal of Applied Mathematics and Decision Sciences* **10**, 148–159.

Scott, A.J., Wu, C. (1981). On the asymptotic distribution of ratio and regression estimators. *Journal of the American Statistical Association* **76**, 98–102.

Scott, S., Sverchkov, M., Pfeffermann, D. (2004). Variance measures for seasonally adjusted employment and employment change. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, Alexandria, VA, 1328–1335.

Scott, S., Sverchkov, M., Pfeffermann, D. (2005). Variance measures for X-11 seasonal adjustment: a summing up of empirical work. *Proceedings of the Section on Survey Research Methods*, American Statistical Association Alexandria, VA, 3534–3545.

Seaman, S.R., Richardson, S. (2004). Equivalence of prospective and retrospective models in the Bayesian analysis of case-control studies. *Biometrika* **91**, 15–25.

Seber, G.A.F. (1973). The Estimation of Animal Abundance and Related Parameters. Charles Griffin & Company Ltd., London.

Sedransk, N., Sedransk, J. (1979). Distinguishing among distributions using data from complex sample designs. *Journal of the American Statistical Association* **74**(368), 754–760.

Sen, A.R. (1953). On the estimate of the variance in sampling with varying probabilities. *Journal of the Indian Society of Agricultural Statistics* **5**, 119–127.

Sen, A.R., Sen, P.K. (1981). Schnabel type estimators for closed populations with multiple markings. *Sankhya B* **43**, 68–80.

Sen, P.K. (1970). The Hájek-Rényi inequality for sampling from a finite population. *Sankhya A* **32**, 181–188.

Sen, P.K. (1972). Finite population sampling and weak convergence to a Brownian bridge. *Sankhya A* **34**, 85–90.

Sen, P.K. (1977). Some invariance principles relating to jackknifing and their role in sequential analysis. *The Annals of Statistics* **5**, 315–329.

Sen, P.K. (1979). Invariance principles for the coupon collector's problem: a martingale approach. *The Annals of Statistics* **7**, 372–380.

Sen, P.K. (1980). Limit theorems for an extended coupon collector's problem and for successive sub-sampling with varying probabilities. *Calcutta Statistical Association Bulletin* **29**, 113–132.

Sen, P.K. (1982a). On the asymptotic normality in sequential sampling tagging. *Sankhya A* **44**, 352–363.

Sen, P.K. (1982b). A renewal theorem for an urn model. *The Annals of Probability* **10**, 838–843.

Sen, P.K. (1987). Sequential estimation of the size of a finite population. *REBRAPE* **1**, 113–137.

Sen, P.K. (1988). Asymptotics in finite population sampling. In: Krishnaiah, P.R., Rao, C.R., *Handbook of Statistics*, vol. 6. Elsevier Science B.V., Amsterdam, pp. 291–331.

Sen, P.K. (1995). The Hájek asymptotics for finite population sampling and its ramifications. *Kybernetika* **31**, 251–268.

Sen, P.K., Singer, J.M. (1993). *Large Sample Methods in Statistics. An Introduction with Applications*. Chapmann & Hall, London.

Serfling, R.J. (1980). *Approximation Theorems of Mathematical Statistics*. John Wiley and Sons, New York.

Shah, B.V., Barnwell, B.G., Bieler, G.S. (1995). *SUDAAN User's Manual*. Research Triangle Institute, Research Triangle Park, NC.

Shah, B.V., Vaish, A.K. (2006). Confidence intervals from complex survey data. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 3993–3396.

Shao J. (1994). L-statistics in complex survey problems. *The Annals of Statistics* **22**, 946–967.

Shao, J. (1996). Resampling methods in sample surveys. *Statistics* **27**, 203–237 (discussion 237–254).

Shao, J. (2003). Impact of the bootstrap on sample surveys. *Statistical Science* **18**, 191–198.

Shao, J., Chen, Y., Chen, Y. (1998). Balanced repeated replication for stratified multistage survey data under imputation. *Journal of the American Statistical Association* **93**, 819–831.

Shao, J., Sitter, R.R. (1996). Bootstrap for imputed survey data. *Journal of the American Statistical Association* **91**, 1278–1288.

Shao, J., Tu, D. (1995). *The Jackknife and Bootstrap*. Springer, New York.

Shao, J., Wu, C.F.J. (1989). A general theory for jackknife variance estimation. *The Annals of Statistics* **17**, 1176–1197.

Shen, W., Louis, T.A. (1998). Triple-goal estimates in two-stage hierarchical models. *Journal of the Royal Statistical Society, Series B* **60**, 455–471.

Shiskin, J., Young, A.H., Musgrave, J.C. (1967). *The X11 Variant of the Census Method II Seasonal Adjustment Program.* Technical Paper 15, Bureau of the Census, US Department of Commerce, Washington, DC.

Shuster, E.F. (1973). Median on the goodness-of-fit problem for continuous symmetric distributions. **68**, 713–715; Corrigenda (1974). *Journal of the American Statistical Association* **69**, 288.

Silva, D.B.N., Smith, T.M.F. (2001). Modelling compositional time series from repeated surveys. *Survey Methodology* **27**, 205–215.

Simonoff, J.S. (1996). *Smoothing Methods in Statistics*. Springer-Verlag, New York.

Singer, J.D., Willet, J.B. (2003). *Applied Longitudinal Data Analysis: Modelling Change and Event Occurrence*. Wiley, New York.

Singh, A.C. (1985). On optimal asymptotic tests for analysis of categorical data from sample surveys. *Methodology Branch working paper SSMD 86-002*, Statistics Canada, Ottawa, ON, Canada.

Singh, A.C. (1987). On the optimality and a generalization of Rao-Robson's statistic. *Communications in Statistics, Theory & Methods* **16**(11), 3255–3273.

Singh, A.C. (2006). Some problems and proposed solutions in developing a small area estimation product for clients. *Proceedings of the Survey Research Section*, American Statistical Association, Washington, DC, 3673–3683.

Singh, A.C., Folsom, R.E. Jr., Vaish, A.K. (2005). Small area modeling for survey data with smoothed error covariance structure via generalized design effects. *Federal Committee on Statistical Methods Conference Proceedings*, Washington, DC. Available at: www.fcsm.gov.

Singh, A.C., Kennedy, B., Wu, S. (2001). Regression composite estimation for the Canadian labour force survey with a rotating panel design. *Survey Methodology* **27**, 33–44.

Singh, A.C., Mantel, H.J., Thomas, B.W. (1994). Time series EBLUPs for small areas using survey data. *Survey Methodology* **20**, 33–43.

Singh, A.C., Mohl, C.A. (1996). Understanding calibration estimators in survey sampling. *Survey Methodology* **22**, 107–115.

Singh, A.C., Rao, J.N.K. (1995). On the adjustment of gross flow estimates for classification error with application to data from the Canadian Labour Force Survey. *Journal of the American Statistical Association* **90**, 478–488.

Singh, A.C., Rao, R.P. (1997). Optimal instrumental variable estimation for linear models with stochastic regressors using estimating functions. In: Basawa, I.V., Godambe, V.P., Taylor, R.L. (Eds.), *Selected Proceedings of the Symposium on Estimating Functions*, IMS Lecture Notes-Monograph Series, Institute of Mathematical Statistics, California: Hayward, vol. 32. pp. 177–192.

Singh, A.C., Roberts, G.R. (1992). Analysis of cross-classified categorical time series of counts. *International Statistical Review* **60**, 321–335.

Singh, A.C., Stukel, D.M., Pfeffermann, D. (1998). Bayesian versus frequentist measures of error in small area estimation. *Journal of the Royal Statistical Society, Series B* **60**, 377–396.

Singh, A.C., Sutradhar, B.C. (1989). Testing proportions for Markov dependent Bernoulli trials. *Biometrika* **76**(4), 809–813.

Singh, A.C., Wu, S. (1998). Hierarchical covariance modeling for nonlinear regression with random parameters. *Proceedings of the Section on Survey Research Methods,* American Statistical Association, Washington, DC, 64–73.

Singh, M.P., Gambino, J., Mantel, H.J. (1994). Issues and strategies for small area data. *Survey Methodology* **20**, 3–14.

Singh, S., Joarder, A.H., Tracy, D.S. (2001). Median estimation using double sampling. **43**(1), 33–46.

Sinha, B.K., Sen, P.K. (1989). On averaging over distinct units in sampling with replacement. *Sankhya B* **51**, 65–83.

Sitter, R.R. (1992a). A resampling procedure for complex survey data. *Journal of the American Statistical Association* **87**, 755–765.

Sitter, R.R. (1992b). Comparing three bootstrap methods for survey data. *The Canadian Journal of Statistics* **20**(2), 135–154.

Sitter, R.R. (1997). Variance estimation for the regression estimator in two-phase sampling. *Journal of the American Statistical Association* **92**, 780–787.

Sitter, R.R., Wu, C. (2001). A note on Woodruff confidence intervals for quantiles. *Statistics and Probability Letters* **52**, 353–358.

Sitter, R.R., Wu, C. (2002). Efficient estimation of quadratic finite population functions in the presence of auxiliary information. *Journal of the American Statistical Association* **97**, 535–543.

Skinner, C.J. (1989a). Domain means, regression and multivariate analysis. In: Skinner, C.J., Holt, D., Smith, T.M.F. (Eds.), *Analysis of Complex Surve*ys. Wiley, New York, pp. 59–88.

Skinner, C.J. (1989b). Introduction to part A. In: Skinner, C.J., Holt, D., Smith, T.M.F. (Eds.), *Analysis of Complex Surveys*, Chapter 2. Wiley, Chichester, UK, pp. 23–58.

Skinner, C.J. (1994). Sample models and weights. In: *Proceedings of the Section on Survey Research Methods*. American Statistical Association, pp. 133–142.

Skinner, C.J., Holmes, D.J (2003). Random effects models for longitudinal data. In: Skinner, C., Chambers, R. (Eds.), *Analysis of Survey Data*. Wiley, New York, pp. 205–219.

Skinner, C.J., Holt, D., Smith, T.M.F. (Eds.) (1989). *Analysis of Complex Surveys*. Wiley, New York.

Smith, D.D. (2001). *Minimum Hellinger Distance Estimation for the Exponential Distribution and Hierarchical Bayesian Approaches in Small Area Estimation*. Unpublished Ph.D. dissertation. Department of Statistics, University of Georgia, Athens, GA.

Smith, J.P., Thomas, D. (2003). Remembrances of things past: test-retest reliability of retrospective migration histories. *Journal of the Royal Statistical Society A* **166**, 23–49.

Smith, K., Joshi, H. (2002). The Millennium Cohort Study. *Population Trends* **107**, 30–35.

Smith, T.M.F. (1976). The foundations of survey sampling: a review. *Journal of the Royal Statistical Society A* **139**, 183–204.

Smith, T.M.F. (1978). Principles and problems in the analysis of repeated surveys. In: Krishnan Namboodiri, N. (Ed.), *Survey Sampling and Measurement*. Academic Press, New York; London, pp. 201–216.

Smith, T.M.F. (1984). Present position and potential developments: some personal views, sample surveys. *Journal of the Royal Statistical Society A* **147**, 208–221.

Smith, T.M.F. (1988). To weight or not to weight, that is the question. In: Bernardo, J.M., Degroot, M.H., Lindley, D.V., Smith, A.F.M. (Eds.), *Bayesian Statistics, Volume 3*. Oxford: Oxford University Press, pp. 437–451.

Smith, T.M.F. (1994). Sample surveys 1975–1990; an age of reconciliation?" (with discussion). *International Statistical Review* **62**, 3–34.

Smith, T.M.F. (1997). Social surveys and social science. *The Canadian Journal of Statistics* **25**, 23–44.

Smith, T.M.F., Njenga, E. (1992). Robust model-based methods for analytic surveys. *Survey Methodology* **18**, 187–208.

Solon, G. (1989). The value of panel data in economic research. In: Kasprzyk, D., Duncan, G., Kalton, G., Singh, M.P. (Eds.), *Panel Surveys*. Wiley, New York, pp. 486–496.

Stapleton, L. (2002). The incorporation of sample weights into multilevel structural equation models. *Structural Equation Modeling* **9**, 475–502.

Steel, D. (1997). Producing monthly estimates of unemployment and employment according to the international labour office definition (Disc: p33-46). *Journal of the Royal Statistical Society, Series A* **160**, 5–33.

Steel, D.G. (1996). *Options for Producing Monthly Estimates of Unemployment According to the ILO Definition*. Central Statistical Office, London.

Steel, D.G. (2004). Sampling in time. In: Kimberly Kempf-Leonard (Ed.), San Diego, CA, *Encyclopaedia of Social Measurement*. Academic Press, pp. 823–828.

Steel, D.G., McLaren, C.H. (2000a). Designing for trend estimation in repeated business surveys. *Proceedings of the Second International Conference on Establishment Surveys*, Invited Papers, American Statistical Association, Alexandria, VA, pp. 799–808.

Steel, D.G., McLaren, C.H. (2000b). The effect of different rotation patterns on the revisions of trend estimates. *Journal of Official Statistics* **16**, 61–76.

Steel, D.G., McLaren, C.H. (2002). In search of a good rotation pattern. In: Gulati, G., Lin, Y.-X., Mishra, S., Rayner, J. (Eds.), *Advances in Statistics, Combinatorics and Related Areas*. World Scientific, New Jersey, 799–808.

Stefanski, L.A., Bay, J.M. (1996). Simulation extrapolation deconvolution of finite population distribution function estimators. *Biometrika* **83**(2), 407–417.

Stenger, H. (1979). A minimax approach to randomization and estimation in survey sampling. *The Annals of Statistics* **7**, 395–399.

Stenger, H. (2002). Regression analysis and random sampling. *Journal of Statistical Planning and Inference* **102**, 169–178.

Stephan, F.F. (1942). An iterative method of adjusting sample frequency tables when expected marginal totals are known. *The Annals of Mathematical Statistics* **13**, 166–178.

Stevens, R.G., Jones, D.Y., Micozzi, M.S., Taylor, P.R. (1988). Body iron stores and the risk of cancer. *New England Journal of Medicine* **319**, 1047–1052.

Stone, C.J. (1977). Consistent nonparametric regression (with discussion). *The Annals of Statistics* **5**, 595–645.

Stone, C.J., Koo, C. (1985). Additive splines in statistics. *Proceedings of the Statistical Computing Section*, American Statistical Association, 45–48.

Stroud, T.W.F. (1991). Hierarchical Bayes predictive means and variances with application to sample survey inference. *Communications in Statistics, Part A-Theory and Methods* **20**, 13–36.

Stroud, T.W.F. (1994). Bayesian inference from categorical survey data. *The Canadian Journal of Statistics* **22**, 33–45.

Stukel, D.M., Hidiroglou, M.A., Särndal, C.-E. (1996). Variance estimation for calibration estimators: a comparison of jackknifing versus Taylor linearization. *Survey Methodology* **22**, 117–125.

Sugden, R.A., Smith, T.M.F. (1984). Ignorable and informative designs in survey sampling inference. *Biometrika* **71**, 495–506.

Sutradhar, B.C., Rao, R.P. (2003). On quasi likelihood inference in generalized linear mixed models with two components of dispersion. *The Canadian Journal of Statistics* **31**, 415–435.

Sverchkov, M., Pfeffermann, D. (2004). Prediction of finite population totals based on the sample distribution. *Survey Methodology* **30**(1), 79–92.

Tam, S.M. (1984). On covariances from overlapping samples. *The American Statistician* **38**, 288–289.

Tang, E.-T. (2008). On the estimation of mean squared prediction error in small area estimation and related topics. Ph.D. dissertation, Department of Statistics, University of California, Davis, CA.

Tate, A.R., Calderwood, L., Dezateux, C., Joshi, H. (2006). Mother's consent to linkage of survey data with her child's birth records in a multi-ethnic national cohort study. *International Journal of Epidemiology* **35**, 294–298.

Thomas, D.R., Rao, J.N.K. (1987). Small sample comparison of test and power for simple goodness-of-fit statistics under cluster sampling. *Journal of the American Statistical Association* **82**, 630–636.

Thomas, D.R., Singh, A.C., Roberts, G.R. (1996). Tests of independence on two way tables under cluster sampling: an evaluation. *International Statistical Review* **64**, 295–311.

Thompson, M.E. (1997). *Theory of Sample Surveys*. Chapman and Hall, London.

Thompson, S.K., Seber, G.A.F. (1996). *Adaptive Sampling*. John Wiley & Sons, New York.

Tiller, R.B. (1992). Time series modeling of sample survey data from the US current population survey. *Journal of Official Statistics* **8**, 149–166.

Tillé, Y. (1999). Estimation in surveys using conditional inclusion probabilities: complex designs. *Survey Methodology* **25**, 57–66.

Tillé, Y. (2006). *Sampling Algorithms*. Springer Science+Business Media, Inc., New York.

Torabi, M. (2006). *Some Contributions to Small Area Estimation*. Unpublished Ph.D. dissertation. School of Mathematics and Statistics, Carleton University, Ottawa, ON, Canada.

Torabi, M., Rao, J.N.K. (2008). Small area estimation under a two-level model. *Survey Methodology* **34**, 11–17.

Trabelsi, A., Hillmer, S.C. (1990). Bench-marking time series with reliable bench-marks. *Applied Statistics* **39**, 367–379.

Truett, J., Cornfield, J., Kannel, W. (1967). A multivariate analysis of coronary heart disease in Framingham. *Journal of Chronic Disease* **20**, 511–524.

Tukey, J. (1958). Bias and confidence in not quite large samples. *The Annals of Mathematical Statistics* **29**, 614.

Tzavidis, N., Salvati, N., Pratesi, M., Chambers, R. (2007). M-quantile models with application to poverty mapping. *Statistical Methods and Applications* (SpringerLink Online publication, October 2007).

University of Essex. (2006). *British Household Panel Survey; Waves 1–14, 1991–2005*, 2nd ed. UK Data Archive. SN: 5151. Institute for Social and Economic Research, Colchester, Essex.

Valliant, R. (1985). Nonlinear prediction theory and the estimation of proportions in a finite population. *Journal of the American Statistical Association* **80**, 631–641.

Valliant, R. (2004). The effect of multiple weighting steps on variance estimation. *Journal of Official Statistics* **20**(1), 1–18.

Valliant, R., Dorfman, A.H., Royall, R.M. (2000). *Finite Population Sampling and Inference: A Prediction Approach*. John Wiley and Sons, New York.

Vidakovič, B. (1999). *Statistical Modeling by Wavelets*. John Wiley & Sons, Inc, New York.

Víšek, J.Á. (1979). Asymptotic distribution of simple estimate for rejective, Sampford and successive sampling. In: Jurečková, J. (Ed.), *Contribution of Statistics, Jaroslav Hájek Memorial Volume*. Academia, Prague, Czech Republic, pp. 71–78.

Waksberg, J. (1998). Random digit dialing sampling for case-control studies. In: Armitage, P. Colton, T. (Eds.), *Encyclopedia of Biostatistics.* Wiley, New York, pp. 3678–3682.

Wallis, K.F. (1974). Seasonal adjustment and relations between variables. *Journal of the American Statistical Association* **69**, 18–31.

Wand, M.P. (2003). Smoothing and mixed models. *Computational Statistics* **18**, 223–249.

Wand, M.P., Coull, B.A., French, J.L., Ganguli, B., Kammann, E.E., Staudenmayer, J., Zanobetti, A. (2005). *SemiPar 1.0. R package*. Available at: http://cran.r-project.org.

Wand, M.P., Jones, M.C. (1995). *Kernel Smoothing*. Chapman and Hall, London.

Wang, D., Chen, S.X. (2009). Empirical likelihood for estimating equations with missing values. *The Annals of Statistics* **37**, 490–517.

Wang, S., Dorfman, A.H. (1996). A new estimator of the finite population distribution function. *Biometrika* **83**, 639–652.

Wang, J., Fuller, W.A. (2003). The mean-squared error of small area predictors constructed with estimated area variances. *Journal of the American Statistical Association* **98**, 716–723.

Wang, Q., Rao, J.N.K. (2002a). Empirical likelihood-based inference in linear models with missing data. *Scandinavian Journal of Statistics* **29**, 563–576.

Wang, Q., Rao, J.N.K. (2002b). Empirical likelihood-based inference under imputation for missing response data. *The Annals of Statistics* **30**, 896–924.

Wang, X., Zuckerman, B., Coffman, G.A., Corwin, M.J. (1995). Familial aggregation of low birth weight among whites and blacks in the United States. *New England Journal of Medicine* **333**, 1744–1749.

Wedderburn, R.W.M. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika* **61**, 439–447.

Welsh, A.H., Ronchetti, E. (1998). Bias-calibrated estimation from sample surveys containing outliers. *Journal of the Royal Statistical Society, Series B* **60**(2), 413–428.

Wermuth, N., Lauritzen, S.L. (1990). On substantive research hypotheses, conditional independence graphs and graphical chain models. *Journal of the Royal Statistical Society B* **52**, 21–50.

Wheeless, S.C., Shah, B.V. (1988). Results of a simulation for comparing two methods for estimating quantiles and their variances for data from a survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 722–727.

White, J.E. (1982). A two stage design for the study of the relationship between a rare exposure and a rare disease. *American Journal of Epidemiology* **115**, 119–128.

Wolter, K. (1985). *Introduction to Variance Estimation*. Springer-Verlag, New York.

Wolter, K. (2007). *Introduction to Variance Estimation* (2nd ed.). Springer Verlag, New York.

Wolter, K.M. (1979). Composite estimation in finite populations. *Journal of the American Statistical Association* **74**, 604–613.

Wolter, K.M., Monsour, N.J. (1981). On the problem of variance estimation for a deseasonalised series. In: Krewski, D., Platek, R., Rao, J.N.K. (Eds.), *Current Topics in Survey Sampling*. Academic Press, New York, pp. 367–403.

Woodruff, R.S. (1952). Confidence intervals for medians and other position measures. *Journal of the American Statistical Association* **47**, 635–646.

Wrensch, M., Lee, M., Miike, R., Newman, B., Barger, G., Davis, R., Wiencke, J., Neuhaus, J. (1997). Familial and personal medical history of cancer and nervous system conditions among adults with glioma and controls. *American Journal of Epidemiology* **145**, 581–93.

Wu, C. (2003). Optimal calibration estimators in survey sampling. *Biometrika* **90**, 937–951.

Wu, C. (2004a). Combining information from multiple surveys through the empirical likelihood method. *The Canadian Journal of Statistics* **32**, 15–26.

Wu, C. (2004b). Some algorithmic aspects of the empirical likelihood method in survey sampling. *Statistica Sinica* **14**, 1057–1067.

Wu, C. (2005). Algorithms and R codes for the pseudo empirical likelihood method in survey sampling. *Survey Methodology* **31**, 239–243.

Wu, C.F.J. (1991). Balanced repeated replications based on mixed orthogonal arrays. *Biometrika* **78**, 181–188.

Wu, C.F.J., Deng, L.Y. (1983). Estimation of variance of the ratio estimator: an empirical study. In: Box, G.E.P., Wu, C.F.J. (Eds.), *Scientific Inference, Data Analysis and Robustness*. Academic Press, New York, pp. 245–277.

Wu, C., Rao, J.N.K. (2006). Pseudo empirical likelihood ratio confidence intervals for complex surveys. *The Canadian Journal of Statistics* **34**, 359–375.

Wu, C., Sitter, R.R. (2001a). A model-calibration approach to using complete auxiliary information from survey data. *Journal of the American Statistical Association* **96**(453), 185–193.

Wu, C., Sitter, R.R. (2001b). Variance estimation for the finite population distribution function with complete auxiliary information. *The Canadian Journal of Statistics* **29**(2), 289–307.

Wu, Y.Y., Fuller, W.A. (2005). Preliminary testing procedures for regression with survey samples. In: *Proceedings of the Section on Survey Research Methods*. American Statistical Association, pp. 3683–3688.

Wu, Y.Y., Fuller, W.A. (2006). Estimation of regression coefficients with unequal probability samples. In: *Proceedings of the Section on Survey Research Methods.* American Statistical Association, pp. 3892–3899.

Yansaneh, I.S., Fuller, W.A. (1998). Optimal recursive estimation for repeated surveys. *Survey Methodology* **24**, 31–40.

Yates, F. (1949). *Sampling Methods for Censuses and Surveys* (1st ed.). Charles Griffin, London.

Yates, F. (1953). *Sampling Methods for Censuses and Surveys*. Charles Griffin, London.

Yates, F. (1960). *Sampling Methods for Censuses and Surveys*. Charles Griffin, London.

Yates, F., Grundy, P.M. (1953). Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society, Series B* **15**, 253–261.

Ybarra, L.M.R., Lohr, S.L. (2008). Small area estimation when auxiliary information is measured with error. To appear in *Biometrika* **95**.

Young, A.H. (1968). Linear approximations to the census and BLS seasonal adjustment methods. *Journal of the American Statistical Association* **63**, 445–471.

You, Y. (1999). *Hierarchical Bayes and Related Methods for Model-Based Small Area Estimation*. Unpublished Ph.D. dissertation. School of Mathematics and Statistics, Carleton University, Ottawa, ON, Canada.

You, Y. (2008). An integrated modeling approach to unemployment rate estimates for subprovincial areas of Canada. *Survey Methodology* **34**, 19–27.

You, Y., Rao, J.N.K. (2002a). A pseudo-empirical best linear unbiased prediction approach to small area estimation using survey weights. *The Canadian Journal of Statistics* **30**, 431–439.

You, Y., Rao, J.N.K. (2002b). Small area estimation using unmatched sampling and linking models. *The Canadian Journal of Statistics* **30**, 3–15.

You, Y., Rao, J.N.K. (2003). Pseudo-hierarchical Bayes small area estimation combining unit level models and survey weights. *Journal of Statistical Planning and Inference* **111**, 197–208.

You, Y., Rao, J.N.K., Gambino, J. (2003). Model-based unemployment rate estimation for the Canadian Labour Force Survey: a hierarchical Bayes approach. *Survey Methodology* **29**, 25–32.

Yuan, K.H., Jennrich, R.I. (1998). Asymptotics of estimating equations under natural conditions. *Journal of Multivariate Analysis* **65**, 245–260.

Yung, W. (1996). Contributions to poststratification in stratified multi-stage samples. Unpublished Ph.D. thesis, Carlton University, Ottawa, Canada.

Yung, W., Rao, J.N.K. (1996). Jackknife linearization variance estimators under stratified multi-stage sampling. *Survey Methodology* **22**, 23–31.

Zeger, S.L., Karim, M.R. (1991). Generalized linear models with random effects: a Gibbs sampling approach. *Journal of the American Statistical Association* **86**, 79–86.

Zeger, S.L., Liang, K.Y., Albert, P.S. (1988). Models for longitudinal data: a generalized estimating equation approach. *Biometrics* **44**, 1049–1060.

Zhao, L., Chen, X. (1990). Normal approximation for finite population U-statistics. *Acta Mathematicae Applicatae Sinica* **6**, 263–272.

Zheng, H., Little, R.J.A. (2003). Penalized spline model-based estimation of finite population total from probability-proportional-to-size samples. *Journal of Official Statistics* **19**, 99–117.

Zheng, H., Little, R.J.A. (2004). Penalized spline nonparametric mixed models for inference about a finite population mean from two-stage samples. *Survey Methodology* **30**, 209–218.

Zhong, C.X.B., Rao, J.N.K. (1996). Empirical likelihood inference under stratified random sampling using auxiliary information. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, pp. 798–803.

Zhong, C.X.B., Rao, J.N.K. (2000). Empirical likelihood inference under stratified sampling using auxiliary population information. *Biometrika* **87**, 929–938.

Zhu, L., Carlin, B., Gelfand, A. (2003). Hierarchical regression with misaligned spatial data: relating ambient ozone and pediatric asthma ER visits in Atlanta. *Environmetrics* **14**, 537–557.

Some of the key words in this Index appear also in the Index of Volume 29A.

Subject Index: Index of Vol. 29B

A

adaptive design, 550
added variable plot, 400, 401
additive model, 107, 305, 309
adequate summary, 427, 463
administrative data, 220, 231, 316, 448, *see also* Index of Vol. 29A
admissible estimator, 167, 545
analytic inference (study), 34, 114–115, 325, 425, 426, 455, 467
ANOVA estimators, 259–261
antedependence models, 321, 322
– unstructured, 215, 322
anticipated variance, 62–63, 69, 112, 424, 546
area-level models, 213, 255, 273
assisting model, 221, 224, 232, 244–246
asymptotic, 58–64
– framework, 90, 91, 192, 470
– normality, 9, 90, 339, 349, 358, 364, 428, 490, 498, 506, 508, 509, 513, 515, 517, 518, 521
autocorrelations, 279, 280, 301–303, 307, 308, 311
autoregressive time series models, 319
auxiliary information (data), 81, 87, 108, 112, 164, 186, 190, 192, 195, 202, 204, 212, 216, 219–221, 226, 233, 240, 247, 254, 323, 376–389, *see also* Index of Vol. 29A
auxiliary variable, 4, 11, 18, 55, 132, 196, 206, 216, 254, 279, 301, 325, *see also* Index of Vol. 29A

B

backfitting, 107
Bahadur representation, 91, 394
balanced half samples (BHS), 128–130, 132, 149
– grouped, 129
balanced repeated replication (BRR), 51, 128, 136, 137, 149, 150, 394, 415, 497, 510, *see also* Index of Vol. 29A
– Fay's method, 129
– repeatedly grouped, 129
– variance estimation for quantiles, 393–394
balanced sampling, weighted, 18–20
bandwidth, 104, 106, 111, 113, 217, 381, 382, 384, 385, 412–415
basic structural model, 214, 305, 321
Bayes
– empirical Bayes, 173, 179, 213, 252, 267–269, 276–278, 285
– estimator, 157, 159, 164–174, 253, 282, 481, 538, 539
– hierarchical Bayes, 8, 118, 154, 171, 179, 252, 270–273, 280–283, 522
– linear estimator, 154, 161–164, 177
– predictor, 176, 213, 267–269, 272, 273, 286
beef population, 380–382
benchmark constraint, 196–198, 200, 201, 203, 205, 206
benchmark variable, 55, 60, 63–65, 68, 81, 230
Bernoulli, 268, 281, 359, 538
Bernoulli model, 31
best linear unbiased predictor (BLUP), 8, 12, 18, 23, 118, 138, 162, 168, 220, 252, 256–257, *see also* Index of Vol. 29A
best predictor, 123, 139–141
bias-robust, 18
binary response, 245, 283, 447
birth cohort studies, 316, *see also* Index of Vol. 29A
BLUP, 8, 12, 18, 23, 118, 138, 162, 168, 220, 252, 256–257, *see also* Index of Vol. 29A
bonus sum after n coupons, 506
bootstrap, 96, 119, 124–125, 130–132, 136, 151, 137, 143, 150, 374, 386, 393, 498, 499, *see also* Index of Vol. 29A
– Bernoulli, 131
– estimates of bias and variance, 393
– mirror-match, 131, 132, 499
– parametric, 143, 145–148, 151
– rescaling, 130–131, 499
– sample, 96, 124, 125, 130, 131, 136, 145, 498, 499
– weight, 96, 124, 125, 130
– without-replacement, 130–131, 498
borrow strength, 224

boundary condition of an estimator of *cdf*, 373
Brewer-selection probabilities, 63
Bubble plots, 362, 398–402

C

calibration, 5, 56, 57, 64, 66–69, 73–80, 112, 113, 196, 220, 226, 230, 233, 237, 248, *see also* Index of Vol. 29A
– bias, 72
– of *cdf*, 386–389
– design balanced, 58, 67, 68
– equation, 55, 57, 58, 61, 65, 74, 196, 230, 231, 237, 248
– estimator, 5, 56, 58, 62, 65–70, 76, 81, 225, 230–231, 386–389
– internal bias, 247
– linear estimator, 57, 65, 66, 69, 70, 72, 231
– model bias, 72–73
– model calibration, 113, 220, 248
– nonlinear estimator, 73–74
– randomization-optimal estimator, 58, 64, 66–69
– truncated linear calibration, 74–75
– variance estimation, 9, 20–22, 26–28, 69–73, 311–313, 392–393, 442–443
– weight, 56–58, 63, 65, 66, 68, 74–76, 81, 230, 387
calibration groups, 230, 237
capture-mark-release-recapture (capture-recapture), 428, 490, 500, *see also* Index of Vol. 29A
case, 434, 438, 441, 442
case-augmented, 443–446
case-control bias, 446
case–control sampling, 431–434, 436, 445, 447–451
case-control studies, 34, 317, 330, 431–453,
– family studies, 451–453
– matched, 432
– population-based, 50, 431–453
– stratified, 431, 447, 449, 452
– two-phase, 447–451
– unmatched, 432
case-enriched, 443–445
case-supplemented, 443–445
causal diagram, 321
cdf, 371–395
census estimating equation, 42, 84, 89, 101, 350, 435, 445
central limit theorem, 340, 424, 491, 494, *see also* Index of Vol. 29A
Chambers–Dunstan (CD) estimator, 376–377, 380, 391, 395
Chao sampling, 514
choice-based, 432
Cholesky residuals, 215, 331, 341, 349, 358, 364, 366, 367
clustered populations, 23–29
cluster sampling, 169, 238, 395, 416, 489, 553
coefficient of variation (CV), 80, 121, 166, 167, 227, 232, 355–357, 436, *see also* Index of Vol. 29A
combined longitudinal and cross-sectional, 368
complex sampling, 9, 129, 215, 367, 427, 435, 438, 449
composite estimation, 235–236, 253–254, 297–305, 308
conditional independence graph, 321
conditional inference, 194, 224, 373
conditional mean, 217, 409–412
conditional percentile, 411, 413–415
confidence interval, 92, 108, 192, 224, 243, 394, 497, *see also* Index of Vol. 29A
consistency, 464, *see also* Index of Vol. 29A
constrained maximization, 194, 200
control, 223, 355, 357, 426, 431, 434
control variable, 55
convex hull, 192, 202, 205
convex loss, 535, 554
correlated binomial, 329, 345
correlation models, 212, 292, 301–304
– unstructured covariance models, 322
coupon collector's problem, 505, 520
covariance matrix, 14, 18, 23, 30–31, 139, 308, 323, 331, 339, 345, 347, 352–353, 363, 436
covariance smoothing, 357–358
coverage
– adjustment, 78
– probability, 148, 193, 204, 284, 285
cross-sectional, 36, 214, 252, 264, 267, 278, 280, 291, 322, 326, 368
– estimation, 212, 316
cross-validation, 113, 114, 485
cumulative distribution function (*cdf*), 85, 191, 371–395, 413–414
– plug-in estimate of variance of *cdf* estimator, 393
– poststratified estimator of, 385, 389
– residual corrected estimator of, 382

D

defining parameters, 84–85
descriptive inference (study), 34, 107–114, 431, 455, 467
design, *see also* Index of Vol. 29A
– balanced, 58
– *p*-unbiased estimators, 293, 541, 546–550
design-based (*see also* randomization-based)
– estimator, 4, 44, 46–47, 194, 216, 219, 220, 372–373, 482

– inference, 3, 34, 153, 221–222, 373, 546, *see also* Index of Vol. 29A
– perspective, 221, 372–373
design domain, *see* planned domain
design effect, 202–204, 216, 332, 354, 359, 435, 440, 450
design-expectation, 18, 20
design information, 45, 359, 427, 450, 460, 463, 476
design-population distribution, 41, 44, 459, 464
design variables, 6, 39, 54, 325, 440, 456, 460, 461–463, 469, 487
design weight, 112, 115, 150, 195, 198, 201, 202, 222, 238, 244, 247, 249, 350, 435, 439, 452
difference estimator, 60, 66, 70, 109, 111, 216, 228, 244, 378, 391, 441, 447
direct estimator, 223–224, 226–232
distance measure, 196, 230
distribution function, 192, 204, 216, 242–244, 371–395, 510
– empirical, 191, 244, 332, 362, 371, 375
– estimation, 373–390, 395
– finite population, 109, 114, 371–375
– inverting estimates, 390
domain, 219, 222–225, 232, 239, 333, 347, 362, *see also* Index of Vol. 29A
– estimation, 174–179, 219, 226–232
– level modeling, 330
double-expansion estimator (DEE), 138, *see also* Index of Vol. 29A

E

EBLUP, 118, 119, 141–143, 146, 220, 225, 244, 249, 253, 254, 257–263, 265, 266, 277, 278, 284, 286, 287, 484
Edgeworth expansion, 130, 131, 499
effective sample size, 435, 439
efficiency, 57, 111, 197, 214, 223, 374, 427, 432, 436, 437, 448, 450, 552
elementary estimates, 292, 298–301, 305, 306, 308
empirical Bayes, 173, 179, 213, 252, 267–269, 276–278, 285
empirical best linear unbiased predictor (EBLUP), 118, 119, 141–143, 146, 220, 225, 244, 249, 252, 253, 254, 257–263, 265, 266, 277, 278, 284, 286, 287, 484, *see also* Index of Vol. 29A
empirical distribution function, 191, 244, 332, 362, 371, 375
empirical likelihood, 9, 95, 207
– pseudo empirical likelihood, 9, 74, 194–205, 387
– ratio, 95, 101, 202–205
entropy, *see also* Index of Vol. 29A
– entropy variance, 357
– large entropy, 510–518
error variance, 15–23, 26–28, 30–31, 228, 238, 284
estimating equations (EE), 49, 318, 319, 465–467, 472, 480, 486, *see also* Index of Vol. 29A
– probability weighted, 9, 43, 466, 467
estimation, *see also* Index of Vol. 29A
– composite, 235–236, 253–254, 297–301, 303, 308
– error, 14, 21, 26, 28, 287
– linear, 9, 14, 17, 29, 42, 189, 513, 553
– of totals, 249, 379
– variance, 9, 20–22, 26–28, 50–52, 69–73, 122, 126–127, 134–137, 216, 222, 226, 248, 249, 311–313, 392–393, 415, 442–443, 464, 467, 473, 479
– weights, 17–18
Estrella's R^2-type measures, 331, 343, 349
event history analysis, 50–51, 322
exchangeability, 164, 391, 548, 549
expansion estimator, 16, 17, 23, 55, 56, 59, 67, 226, *see also* Index of Vol. 29A
– reweighted expansion estimator (REE), 138
expected information, 331, 339, 344, 348, 363
experimental studies, 317
exponential family, 161, 164, 270
exponential sampling, 515, 518
extended domain variable, 223, 234, 240

F

Fay's Jackknifed χ^2 test, 357
finite population
– distribution function, 109, 114, 372
– mean, 84, 139, 159, 164, 174, 254, 256, 267
– quantities, 3, 5, 9, 35, 85
– sampling, 37, 140, 154, 156, 489–522, 539–541
– totals, model-based prediction of, 11–31, 476–478
frame, 11, 15, 25, *see also* Index of Vol. 29A
frequency matching, 432

G

general exponential model (GEM), 58, 78–79
generalized additive model, 107, 113
generalized design effects (g-deffs), 332, 354
generalized difference estimator, 109, 111, 541
generalized estimating equations (GEEs), 318–319, 346
generalized inverse 22–23, 25
generalized least squares (GLS) estimator, 228, 257, 473, 484
– iterative, 49, 319, 324
generalized linear mixed model (GLMM), 244, 252, 269, 282–283
generalized linear models (GLMs), 14–17, 179–186, 318
generalized raking, 75, 79, 80

generalized regression estimator (GREG), 60–64, 132–133, 220, 229–230, 233–239, *see also* Index of Vol. 29A
general prediction theorem, 15
g-inverse, 22, 23, 25, 337, 359
Godambe–Joshi variance bound, 20, 112
graphical chain modeling (GCM), 214, 321
GREG (generalized regression estimator), 60–64, 132–133, 220, 229–230, 233–239, *see also* Index of Vol. 29A
gross flows, 214, 291, 303, 368
grouping for unit level models, 362
group level residuals, 362
$\mathcal{G}$-unbiased predictor, 551
g-weight, 230, 233, 241, 245, 249, *see also* Index of Vol. 29A

H

Hadamard matrix, 128–129
Hájek approximation, 227, 514
Hájek estimator, 4, 86, 195, 196, 216, 227, 231, 232, 375, 389, 391, 467, 478, *see also* Index of Vol. 29A
Hansen–Hurwitz variance estimator, 123, 227, 503
hat matrix, 29, 342, 364
Heaviside function, 371
hierarchical Bayes, 8, 118, 154, 171, 179, 252, 270–273, 280–283, 522
hinge estimates of quantiles, 390, 392
Horvitz–Thompson estimator (HT), 4, 20, 23, 25, 108, 111, 117, 166, 186, 221, 226–227, 330, 375, 428, 464, 498, 504–506, 513–516, 526, 527, 544, 553, *see also* Index of Vol. 29A
Hosmer–Lemeshow statistic, 332, 365, 367
household panel survey, 316

I

ignorability, 6, 8, 40–41, 427, 460, 461, 463, 484–486
imputation, 94, 99–101, 131, 134–137, 207, 301, 323, 407–409, 476, *see also* Index of Vol. 29A
– nearest neighbor, 325
– regression, 100, 207, 325
– weighted hot-deck 135, 136
inclusion probabilities, 7, 20, 86, 194, 222, 226, 246, 249, 326, 373, 456, 457, 461, 463, 465, 469, 470, 479, 503, 511, 517, 555, *see also* Index of Vol. 29A
– first-order, 194, 198, 326, 350, 503
– second-order, 226, 227, 330, 345, 350, 357, 367, 373, 374, 378, 503
– target, 517, 518
indicator variable, 68, 184, 201, 279, 336, 373, 384
indirect estimator, 212, 223–224, 233–244
inferential estimation, 215, 331, 344–345, 367
inferential testing, 215, 331, 343–344, 349–350, 358–361, 367
influential points, 217, 342, 358, 364, 401, 419
information unbiasedness, 347, 348, 354
informative response, 455, 456
informative sampling, 6, 40–41, 215, 221, 247, 252, 325, 421, 427, 455–487, 524
instability of covariance matrix, 215
instrumental variable, 8, 65–66, 238
instrument vector, 230, 231, 237
integrating data, 34, 52–53
intervention, 214, 318
inverse testing, 94–96
irregular component, 309–312
isotonic regression, 374
iterative proportional fitting, 57, 74, 79, 199, 336, *see also* Index of Vol. 29A

J

jackknife, 5, 9, 22, 29, 48, 52, 81, 123, 125–127, 132–140, 143–146, 149–151, 249, 269, 356, 415, 495, 498, *see also* Index of Vol. 29A
– delete-d jackknife, 125, 132
– delete-one-cluster jackknife, 127–128
– delete-one jackknife, 125
– estimates (of variance of *cdf* estimator), 393
– variance estimation, 22, 29, 52, 127–133, 135, 138, 139, 149, 227, 323, 415, 497, 498
jittering, 217, 398, 402, 404–406

K

Kalman filter, 214, 305–308, 311, 324
kernel-based estimation, 104, 384, 480
kernel density estimation, 104, 105, 114
kernel smoothing, 212, 384, 398, 409, 411–417
Kullback–Leibler information, 486
Kuo estimator, 384

L

Lagrange multiplier, 191, 196, 200–201, 205, 555
latent variables, 321, 325, 456, *see also* Index of Vol. 29A
length-biased sampling, 86
leverage, 21
likelihood, *see also* Index of Vol. 29A
– equations, 98, 437, 466, 467, 472
– full likelihood, 471, 474–475
– inferences, 189, 212, 459, 460, 468
– maximum likelihood estimator (MLE), 30, 42, 43, 49, 83, 115, 119, 173, 229, 258, 319, 329, 387, 426, 436, 445, 448, 462, 500
– principle, 154–157, 186

– probability-weighted likelihood, 481, 482
– ratio test statistic, 101, 119, 215, 343, 441
– sample, 472, 474, 475
Linear estimator 14, 17, 30–31
linear model, 5, 8, 14–17, 29, 31, 33, 113, 141, 179, 197, 340, 388
linear regression estimator, 16, 17, 137, 194
local polynomial regression, 104, 105, 107, 110, 413
local regression smoother, 411–413
logistic (logit) model, 33, 139, 244, 245, 267, 281, 317, 329, 330, 335, 337, 349, 362, 388, 399, 400–401, 426, 433, 435, 440, 444, 460, 466, 482, 486
log-linear models, 329, 335, 336, 337, 354, *see also* Index of Vol. 29A
longitudinal categorical data, 368
longitudinal surveys, 211, 214, 289, 292, 315–327, 368, *see also* Index of Vol. 29A
loss function, 375, 428, 526, 531, 533, 535

M

marginal totals, 198, 199, 201, 202, 300
Markov transition models, 319
martingale properties, 494
maximum likelihood estimator (MLE), 30, 42, 43, 49, 83, 115, 119, 173, 229, 258, 319, 329, 387, 426, 436, 445, 448, 462, 500
mean-of-ratios estimator, 16–17, 67
mean squared error (MSE), 56, 58, 62, 68, 79, 81, 90, 142, 143, 149, 153, 222, 252, 313, 442, 476, 478–479, *see also* Index of Vol. 29A
– relative, 57, 58
measurement model, 321
median, 85, 217, 242, 262, 265, 267, 274, 390, 392, 414
missing at random (MAR), 207, 323, 425, 448, 459, 463, *see also* Index of Vol. 29A
missing completely at random (MCAR), 323, *see also* Index of Vol. 29A
missing data, 99, 215, 292, 322–324, 368, 407–409, 446, 448
mixed categorical models, 367
mixed models, 118, 123, 141, 148, 212, 244, 245, 247, 249, 254, 270
– linear, 118, 252, 253, 320
model-assisted, 12, 15, 26, 109, 111, 113, 121, 153, 211, 220, 223, 225, 246, 373
model-based, 5, 8, 11, 12, 17, 21, 33, 44, 45–46, 48, 61, 109, 121, 142, 153, 213, 251, 253, 255, 310, 323, 373, 424, 476, 482
– estimators, 7, 41, 42, 45–46, 91, 109, 220, 257, 372, 380, 468
– inference, 3, 7, 33, 115, 153, 284, 424, 429, 467, 546
– model-unbiased, 25, 69
– perspective, 149, 372–373, 382
model-design-based inference, 3, 39–41, 44, 45, 49, 50, 54, 153, 211, 221–223, 224, 546
model diagnostics, 212, 215, 284, 331, 332, 340–343, 349, 357–358, 363–364, 366, 380
model-free calibration, 220, 230
model group, 236–237, 239
model misspecification, 12, 41, 108, 115, 187, 212, 216, 373, 380, 462
model selection, 215, 284, 331, 333–340, 346–349, 351–357, 363, 366, 380
monotonicity of estimator of *cdf*, 373
multilevel models, 24, 49, 214, 290, 319–321, 466, 474, 480–481
multistage, 433, 502, 518
– sampling, 11, 15, 59, 126–127, 130, 135, 164, 168–174, 292, 428, 435, 452, 518
multivariate, 91, 97, 113, 173, 262, 273
– counting processes, 215, 322

N

naive estimator (of *cdf*), 261, 269, 376
nearest neighbor imputation, 325, *see also* Index of Vol. 29A
nested hypotheses, 215, 331, 339, 348–349, 353, 354
neural network, 106, 113, 248
Neyman's score function, 215, 331, 338
Neyman's score statistic, 331, 365
noninformative design, 547, 550
nonlinear models, 5, 29–31, 233, 244, 329
nonparametric estimation, 479–480
nonparametric regression, 9
nonparametric regression estimation, 9, 115, 381, 383–386, 395
– of *cdf*, 276–277, 381, 383–386
– of variances, 381
nonresponse, 3, 76–78, 79, 99, 115–118, 134, 215, 217, 322–325, 398, 407, 456–463, 469, 475, *see also* Index of Vol. 29A
– bias, 116, 317
– informative, 323, 327, 456, 476
– missing at random (MAR), 207, 323, 425, 448, 459, 463
– missing completely at random (MCAR), 323
– not missing at random (NMAR), 323, 425, 463, *see also* Index of Vol. 29A
– unit, 57, 58, 76–78, 80, 81, 99, 116, 135, 217, 231, 368, 407, 455
nonresponse mechanism, 38
normalized weights, 54, 195
not missing at random (NMAR), 323, 425, 463

nuisance parameter, 84, 97–99, 215, 331, 334, 338, 359, 364
– adjusted score, 215, 331, 338

O

observational studies, 214, 317, 446
observed at random, 459
observed information, 224, 331, 335, 344, 363, 452
optimal estimating function, 88, 243, 270, 329
optimal estimator, 194, 227, 238, 306, 348, 547
optimal predictor, 24, 213
ordered categorical data, 368
order sampling, 335–337, 515–517, *see also* Index of Vol. 29A
outliers, 221, 381, 419, 535, *see also* Index of Vol. 29A
overdispersion, 283, 329, 345, 347

P

panel estimates, 299, 302, 303, 306, 308
panel surveys, *see also* Index of Vol. 29A
– household panel survey, 315, 316
– repeated, 315
parameter space, 90, 146, 155, 192, 207, 428, 524, 526, 528, 532, 541, 548
Pareto sampling, 517–518
partial residual plot, 415, 416
path analysis, 214, 321, *see also* Index of Vol. 29A
Pearson's Chi-square test, 215, 339
penalized spline regression, 105–107, 112, 118
penalty, 105, 106, 111, 113, 117
percentile, 53, 217, 241–244, 284, 363, 407, 409–412
– as group boundaries, 363
planned domain, 222, 225, 231–232, 234
Poisson sampling, 59, 63, 67, 69–70, 76, 226, 249, 334, 336, 359, 360, 458, 474, 505, 510–511, *see also* Index of Vol. 29A
– conditional, 511, 512
Polya posterior, 392
polynomial model, 19
polytomous response, 245, 335, 362
pooling approach to data integration, 52–53
population, 3, 8, 11, 15, 23–29, 50, 68, 84, 121, 130, 140, 156, 162, 164, 190, 201, 216, 221, 228, 292, 379, 392, 431, 456, 476, 489, 523, 524, *see also* Index of Vol. 29A
– target, 34–40, 50–52, 131, 206, 425, 431, 433, 434, 460, 482
population-averaged approach, 318–319
population distribution (model), 455–487
population mean, 4, 8, 16, 19, 38, 68, 84, 109, 133, 139, 154, 159, 162, 164–174, 196, 197, 201, 205, 254, 297, 429, 462, 464, 465, 475, 490, 533, 546
population of inference, 34
population size, estimation of, 428, 499–502
posterior distribution, 121, 156, 165, 172, 180, 253, 282
posterior linearity, 161–163, 171, 177
poststratified estimator, 4, 26, 385, 389
poststratum, 25, 57, 277, 385
prediction, 8, 11–31, 55, 76, 79, 138–140, 157, 372, 455, 460, 468, 475–479, 482, 550–552, *see also* Index of Vol. 29A
– augmented regression, 324
prediction form, 66
prediction model, 55, 65, 76, 78, 80–82
prediction-unbiased, 15, 17, 18, 25, 29
– estimator of total, 25
prediction variance, 15, 122, 123, 138–140, 149
predictive distribution, 158, 169, 178, 271, 286
predictor, 8, 12, 22, 24, 28, 56, 79, 123, 140, 176, 213, 267, 272, 286, 342, 424, 476–478, 483, 551
– linear, 551, 552
primary domain, *see* planned domain
prior distribution, 121, 153, 156, 161, 162, 178, 213, 252, 271, 276, 280
probability inequality, 490
probability proportional to size (pps) sampling, 20, 67, 131, 191, 193, 195, 226, 231, 246, 357, 429, 469, 499, 504, 553
probability sampling, 25, 33, 59, 81, 121, 189, 194, 227, 330, 372, 390, 426, 455, 456, 486
– varying, 489, 490, 502–510, 518–521
probability weighted estimating equations, 9, 43, 466, 467
probability weighted likelihood, 481, 482
probability weighting, 99, 427, 464–468, 487
profile estimating function, 98
projection form, 65, 66
proportionality condition, 332, 353, 354, 357
prospective measurement, 214, 316, 317
prospective sampling, 432
pseudo empirical likelihood, 9, 74, 194–205, 387
pseudo errors, 302
pseudo maximum likelihood
– approach (method), 229, 326, 426, 351
– estimator, 43, 326, 351, 426, 466, 480, 481
– multilevel, 49
– weights, 439, 441–443, 451
pseudo population, 131, 478, 479
pseudoreplication, 128
pseudovalues, 126, 495

Q

quadratic loss, 272, 429, 465, 527–529, 533–536, 538–541, 546, 548–549, 553
quadratic score statistic, 215, 331

quadratic variance function, 161, 270
qualitative auxiliaries, 17, 22–23
quantile, 20, 85, 91, 92, 96, 132, 136, 192, 196, 203, 204, 207, 211, 212, 216, 219, 342, 366, 371–395, 413, 414, 417, 510
– position estimator, 391, 392
quasi-likelihood (*ql)* estimation, 186, 212, 216, 329, 345–350, 366, 367
– weighted, 216, 350–361, 366
quasi-randomization, 58, 76–80, 81, 117
quasi-score functions, 215, 346–348
– weighted, 215, 331, 345–348, 350–352, 359, 362, 366

R

raking, 57, 198–202, 206, *see also* Index of Vol. 29A
random effects, 34, 39, 49–50, 119, 141–143, 150–151, 180, 212, 214, 245, 246, 249, 255, 257, 258, 264, 265, 279, 283, 286, 318–320, 367, 451, 466, 480–484
random group, 51, 122, 129, 320
randomization, 12, 38–40, 43, 189, 212, 350, 351, 353, 423–426, 458–459, 464, 465, 467, 468, 472, 473, 477, 478, 485, 487, 499, 529
– approach, 121, 212, 219
– consistency, 58–59, 61, 65, 76, 464, 465, 467
randomization-based, 3, 4, 55, 58, 61, 62, 70, 72, 73, 81, 211, 424 (*see also* design-based)
range-restricted weights, 74, 82, 197
range restrictions, 82, 329
Rao–Kovar–Mantel (RKM) estimator, 378–381, 383, 385, 387
Rao–Scott's corrected χ^2 tests, 331, 355, 357
ratio, 241–242
ratio estimator, 4, 12, 16–19, 66, 137, 138, 166, 168, 186, 224, 242, 254, 300, 391, 497, 526, 527, *see also* Index of Vol. 29A
regression, *see also* Index of Vol. 29A
– estimator, 4, 5, 8, 9, 16, 17, 89, 100, 109, 111, 112, 137, 228, 229, 235, 491
– imputation, 100, 207, 323, 325
– isotonic, 374
– local polynomial, 104, 105, 107, 110, 413
– models, 4–6, 25, 43, 89, 236–239, 256, 269, 281, 286, 287, 323, 358, 408, 418, 431–435, 467, 485, 556
rejective sampling, 504, 511, 512–515
relative error, 58, 59, 232
repeated measures, 318–319
repeated surveys, 214, 235, 264, 289–313, 316, 325, *see also* Index of Vol. 29A
resampling, 96, 121–151, 473, 490, 495–499
residuals, 16, 21, 22, 27, 28, 69, 107, 221, 228–230, 234, 258, 340–342, 362, 381, 485
respondent distribution, 458
response, *see also* Index of Vol. 29A
– indicators, 116, 117, 324, 456, 457
– informative, 455, 456
– model, 76, 79, 269, 270, 325, 369
– probabilities, 99, 116, 117, 323, 427, 456, 461, 463–465, 467–469, 471, 475
retrospective measurement, 214, 317
retrospective sampling, 317
reweighting, 39, 436–440, 441–442
risk, 527, 528, 529, 536, 538, 541, 546, 553, 554
robust consistent variance estimator, 346
robust estimator, 27, 28, 92, 381
robustness, 18–20, 21, 44, 99, 115, 186, 212, 213, 216, 377, 383, 440, 441–442, 552
rolling estimates, 291, 294, 304
rotation, *see also* Index of Vol. 29A
– group, 290, 291, 296–301, 303
– patterns, 214, 279, 280, 290–294, 296–298, 300–306, 309, 310, 312, 313

S

Sampford design, 59, 504, 514, *see also* Index of Vol. 29A
sample-based, 4, 5, 7, 204, 228, 229, 247, *see also* Index of Vol. 29A
sample-complement distribution (model), 475–478
sample distribution (model), 455–487
sampled scatterplot, 217, 398, 402–405, 407–410, 412
sample empirical distribution function, 375
sample indicators, 131, 456, 457, 474
sampling, *see also* Index of Vol. 29A
– complex, 9, 129, 215, 367, 427, 435, 436, 438, 449
– multistage, 11, 15, 59, 126–127, 135, 164, 168–174, 292, 428, 435, 452, 518
– two-stage, 24, 26, 112, 139, 169, 179, 239, 273, 458, 480, 509, 557
– uniform, 515
sampling design, 3, 4, 6, 7, 9, 40, 55, 67, 70–72, 86–89, 91, 108, 115–117, 128, 129, 131, 137–138, 155, 164, 167, 190, 194, 195, 204, 213, 215, 221–223, 227, 231, 252–254, 290, 325, 326, 330, 425–429, 447, 457, 458, 461, 463, 503, 514, 524, 528, 547, 552
sampling error, 38, 140, 212, 255, 264, 273, 276–278, 284, 287, 289, 292–295, 301–304, 305–313
sampling fraction, 16, 17, 26, 27, 33, 45–48, 50, 54, 127, 130, 131, 139, 166, 190, 191, 194, 268, 294, 351, 434, 475, 476, 478, 489
– non-negligible, 47–48, 130, 131

sampling weight, 6, 55, 64, 65, 123, 126, 186, 330, 345, 350–353, 362, 366–368, 384, 409, 426, 427, 455, 463–465, 468–472, 479, 481, 482–484
sandwich
– covariance, 358, 438, 446
– estimator, 21, 28, 50, 92, 436
– variance estimation, 21, 27, 29, 50, 52, 436
scale-load, 95, 190–192
score function, 83, 101, 186, 215, 334, 338, 346, 474, *see also* Index of Vol. 29A
seasonal adjustment, 289, 291, 293, 297, 299, 307, 309–313
– X11, 308, 310–313
secondary domain, *see* unplanned domain
second order unbiased, 142, 143, 145, 146, 150, 151, 252, 261–263, 266, 269, 278, 281, 284, 285, 287
selection probability, 55, 59, 62, 63, 71, 78, 115, 153, 167, 191, 292, 299, 325, 330, 391, 425–428, 435, 436, 455, 456, 464, 465, 469, 470, 478, 481, 472, 487, 502
semiparametric framework, 215, 329, 345
semiparametric regression, 323
Sen–Yates–Grundy estimator, 226, 227
separate approach to survey data integration, 52–53
sequential sampling tagging, 502
simple random sampling (SRS), 6, 18, 28, 34, 40, 59, 123, 136, 140, 191–194, 238, 330, 425, 462, 468, 490–495, 497, 525, 527, 536, 552–557, *see also* Index of Vol. 29A
– conditional, 530–532, 534, 537, 544, 552
simple random sampling without replacement (SRSWOR), 16, 26, 123, 130, 131, 137, 157, 190, 226, 227, 249, 357, 428, 490, 494, 515, 525
simple random sampling with replacement (SRSWR), 123, 124, 128, 130, 192, 200, 428, 490, 492, 493, 494, 498
Singh's $Q^{(T)}$ test, 357
single-stage sampling, 16, 25, 48, 303, 464, 471, 475, 478
small area estimation, 118–119, 123, 140–150, 183, 187, 213, 220, 224, 251–288, 358, 423, 468, 476, 482–485, 521–522, *see also* Index of Vol. 29A
small domain, 176, 212, 223, 225, 232, 233, 235, 236, 239–241, 244, 251, 358
software, 20, 213, 223, 227, 237, 248–249
sparse tables, 368
spline function, 105, 106, 119, 417, 418
spline regression, 105–107, 112, 113, 118, 417–419
state-space models (SSMs), 214, 215, 279, 292, 305–308, 310, 311, 320, 323
strategy, 69, 146, 225, 247, 299, 334, 402, 424, 428, 526–528, 537, 542, 544, 546, 549, 550, 553
– minimax, 428, 529, 531, 540, 555, 557
– unbiased, 429, 531–533, 544, 547, 549
stratified expansion estimator, 16, 17, 23
stratified sampling, 131, 174–179, 203, 205, 223, 492, 498, 499, 519, 521, *see also* Index of Vol. 29A
structural equation model (SEM), 50, 214, 321, *see also* Index of Vol. 29A
structural submodel, marginal, 318
structured antedependence models, 215, 322
subject-specific approach, 318
successive sampling, 489, 505, 506, 508, 514, 515, 518
successive subsampling, 490, 518–521
sufficiency principle, 154, 155, 157, 186
superpopulation, 84–85, 87, 91, 93, 100, 114, 121, 153, 221, 257, 282, 350, 432, 438, 442–443, 510, *see also* Index of Vol. 29A
superpopulation models, 7, 9, 11–14, 31, 39, 87, 93, 100, 101, 108, 109, 111, 113, 121, 138–140, 154, 157, 177, 215, 228, 325, 429, 546–552, 554–557
survey, complex, 45, 51, 53, 96, 103–119, 126–134, 329–369
survey producer's target population, 34, 35, 51
survey weights, 109
survival analysis, 290, 322, 444
symmetric, 192, 204, 344, 384, 412, 491, 524, 526, 527, 532, 534, 536
– estimator, 526, 530, 534–535, 540, 542, 546
– parameter, 524, 527, 531, 533, 534, 536
synthetic estimator, 141, 212, 229, 235, 249, 254, 270, *see also* Index of Vol. 29A
systematic sampling, 59, 434, 510, *see also* Index of Vol. 29A

T

tail error rates, 192, 193, 204
target inclusion probabilities, 517, 518
target population, 34–40, 50–52, 131, 206, 425, 431, 433, 434, 460, 482
test inversion for interval estimation, 331, 344, 345, 350, 358, 367, *see also* inverse testing
test of sampling ignorability, 484–486
time series, 101, 214, 215, 252, 264–267, 278–280, 284, 289, 292, 295, 305–307, 309–312, 320, 323
– models, 212–214, 264, 265, 305, 307, 309, 311, 319, 320, 325
totals, 219, *see also* Index of Vol. 29A
– estimating, 11–31

trend estimation, 289, 291, 295, 306, 309–313, *see also* Index of Vol. 29A
truth function, 371
tuning constant, 106, 382, *see also* Index of Vol. 29A
two-level model, 146, 147, 281, 319, 320, 427, 466, 470, 471, 480, 481, 482
two-stage sampling, 24, 26, 112, 139, 169, 179, 239, 273, 458, 480, 509, 557, *see also* Index of Vol. 29A
two-step estimation method, 326
two-way model with interaction, 26

U

ultimate cluster variance estimator, 28, 122
unbiased strategy, 429, 531–533, 544, 547, 549
unequal probability sampling, 59, 81, 131, 191, 193, 194, 204, 207, 227, 248, 330, 352, 398, 428, 497–499, 505, *see also* Index of Vol. 29A
uniform sampling, 515
unit-level models, 141, 142, 213, 216, 220, 252, 255, 256, 267, 286, 330, 332, 361–367
unit nonresponse, 57, 58, 76–78, 80, 81, 99, 116, 135, 217, 231, 368, 407, 455, *see also* Index of Vol. 29A
units of analysis, 37
univariate, 55, 63, 91, 254–256, 262, 263, 266, 267, 270–275, 278, 284, 306
unplanned domain, 222–225, 227, 233, 244, 239–241, 248
urn model, 391
U-statistics, 428, 490, 491–497

V

variance estimation, 9, 20–22, 26–28, 50–52, 69–73, 122, 126–127, 134–137, 216, 222, 226, 248, 249, 292, 301, 311–313, 392–393, 415, 442–443, 464, 467, 473, 478, 479, *see also* Index of Vol. 29A
– jackknife, 22, 29, 52, 127–133, 135, 138, 139, 149, 227, 323, 415, 497, 498
– robust, 20–22, 26–29, 31
– sandwich, 21, 27, 29, 50, 52, 436

W

Wald's test, 331, 334, 339–340, 344, 349, 356
Wavelet, 106
weak model (also "weakening the model"), 383–384, 389
weighting cell estimator, 116, 117
weights, 53, 55, *see also* Index of Vol. 29A
– estimation, 17–18
– pseudo maximum likelihood, 439, 441–443, 451
Woodruff (confidence intervals for quantiles), 393–394
working covariance, 215, 332, 346, 351, 353, 354, 357–360, 365–367
working model, 4, 5, 18–22, 24, 27, 29, 212, 221, 284, 373, 377–382, 386, 389

X

X11, 308, 310–313

Y

Yates-Grundy estimator, 514

Subject Index: Index of Vol. 29A

100-bank, 127
1948 election, 570, 581
3-class EM, 360, 361

A

abutting panels, 100
accessibility, 164, 165, 172, 173, 517
accuracy and coverage evaluation (ACE) survey, 544, 552, 558
adaptive allocation, 114, 115, 123
adaptive sampling, 115–123, 497–498
– cluster, 118–123
address matching, 138
administrative data, 403, 427, 446–450, 482, 565, *see also* Index of Vol. 29B
advance letter, 138, 166
aerial photography, 476, 488
aggregation method, 208, 210, 330
agreement model, 286
agricultural surveys, 399–401, 403, 471–486
Akaike Information Criterion (AIC), 301
amenability, 164, 172, 173, 545
American Association of Public Opinion Research (AAPOR), 143–145, 167, 573, 575
American Community Survey (ACS), 93, 94, 175, 402, 407, 421, 564
anomaly plot, 209, 331
anonymisation, 382
approximate string comparison, 362
area sampling, 473–476
area sampling frame, 73–75, 85–87, 399, 400, 409, 441, 444, 445, 473–476, 479, 480, 490, 492, 493
associated persons, 100
asymmetric distribution, 250–251, 256
attribute disclosure, 382, 384
attrition
– bias, 532
– nonresponse, 101, 105, 107
authoring language, 319–322
automatic editing, 198–207, 211, 212, 328–329
automatic error localization, 198, 199, 202
autonomous independence, 543
auxiliary information, 4, 7, 8, 19–26, 60–65, 216, 217, 221–223, 231, 246, 259, 337, 338, 403–404, *see also* Index of Vol. 29B
auxiliary variable, 6, 10, 14, 19, 22, 26–27, 30, 34, 35, 171, 175, 178–180, 184, 216, 230, 234, 274, 332, 333, 337, 338, 340, 429, 430, 493, *see also* Index of Vol. 29B

B

balance edit, 189, 198, 202
balanced repeated replication (BRR), 63, 182, 342, 347, 434, *see also* Index of Vol. 29B
balanced sampling, 6, 19, 51–54, 493, 503
Bayesian Information Criterion (BIC), 301, 305
Behavioral Risk Factors Surveillance System, U.S., 86
Bernoulli sampling, 391, 393, 458, 467
best linear unbiased predictor (BLUP), 5, 504, 506, *see also* Index of Vol. 29B
between-trial correlations, 287
bias-variance trade-off, 256, 261, 278
biological specimen collections, 291
birth cohort studies, 99–100, *see also* Index of Vol. 29B
Blaise, 190, 191, 321, 322, 330, 342
BLUP, 5, 504, 506, *see also* Index of Vol. 29B
bootstrap, 63, 118, 124, 182, 239, 241–242, 263, 268, *see also* Index of Vol. 29B
bounding, 96, 173, 469
Brewer sampling, 45
bridging survey, 92
business
– register, 265, 442, 444, 445, 459
– surveys, 270–272, 401, 403, 441–470

C

CADI (computer assisted data input), 327
CAI (computer assisted interviewing), 317, 319, 327, 328

calendarization, 448–450
calibration, 20, 87, 152, 175, 177, 179, 216, 338, 341–342, 447, 464, 468, 562, *see also* Index of Vol. 29B
– estimation, 87, 178, 183, 184, 249, 259–261, 264, 265, 269, 270, 272, 338, 341–342, 403, 429–431
– robust estimator, 260, 264, 268, 269
– variance estimation, 48–51, 62–64, 83–84, 153–154, 235–243, 263, 345, 433–435, 503
call centers, 126
call outcome codes, 140, 142, 143
campaigns, 534, 568, 574, 575, 578, 583, 588–590, 592
CAPI (computer assisted personal interviewing), 93, 126, 191, 192, 549, 564
capture -mark-release- recapture (capture-recapture), 75, 498–499, 542, 545, *see also* Index of Vol. 29B
case disposition, 142–143
case management, 139
CASI (computer assisted self interviewing), 191, 192, 319
CATI (computer assisted telephone interviewing), 93, 125, 191, 192, 319, 525, 551, 578
causal independence, 543, 544
CAWI (computer assisted web interviewing), 191, 192
cellular phones, 591–592
census
– adjusted census counts, 541, 552, 565
– integrated census (IC), 547, 549, 550, 552
– intercensal estimates, 552
– intercensal years, 541, 559, 564
– long form, 541, 562–565
– one number census (ONC), 544, 546, 565
– register, 540, 542, 553, 550, 557, 565
– rolling, 541, 558, 559–560, 564
Census Coverage Survey (CCS), 546
census day
– population, 539, 554
– residence, 540, 557, 558
central limit theorem, 14, 64–65, *see also* Index of Vol. 29B
Chicago Record, 568
Chicago Tribune, 529, 569
classification errors, 285, 291, 296
classification probability model, 282, 289
classification rule, 357, 366
cluster, 106, 119, 120, 128, 241, 409–412, 445, 519
coefficient of variation (CV), 227, 412, 455, *see also* Index of Vol. 29B
cohabitants, 100, 105
Cohen's kappa, 286
cohort studies, birth, 99–100
collection mode, 165, 191, 192, 293–294, 310, 404, 405, 443
collection unit, 443, 444
completeness error, 326
complex survey, 305–306, 412, 423, 433
composite
– estimation, 95, 422, 431–433, 507, 508
– weight, 175, 432
computer assisted data input (CADI), 327
computer assisted interviewing (CAI), 317, 319, 327, 328, 425–426
computer assisted personal interviewing (CAPI), 93, 126, 191, 192, 549, 564
computer assisted self interviewing (CASI), 191, 192, 319
computer assisted telephone interviewing (CATI), 93, 125, 191, 192, 319, 525, 551, 578
computer assisted web interviewing (CAWI), 191, 192
conditional independence, 356
confidence interval, 17–19, 235, *see also* Index of Vol. 29B
– estimation, 123–124, 181–183
confidentiality, 382, 473, 485
consistency, 49–51, *see also* Index of Vol. 29B
– errors, 326, 329
consistent record, 189, 202, 465
consumer panels, 529–534
context effects, 92, 103, 595
controlled tabular adjustment, 387
coordination, 458–460
correlation bias, 544, 545, 552
coverage bias, 87
creative editing, 188
critical stream, 192, 211, 212, 330
cross-sectional estimation, 97, 101, 105
cross-sectional survey, 89, 98, 101, 102, 104–106, 108
CSPro, 190
cube method, 6, 19, 51–53

D

data
– collection, 72, 85, 93, 106, 162, 188, 191–192, 273, 318–325, 425–427, 446, 480–481, 548, 554
– disclosure, 107
– quality, 162, 383
decision rule, 353, 362
deductive imputation, 332, 465
deliberative polls, 589
demographic
– data, 199, 545, 547, 550
– estimates, 545, 557
– variables, 430, 518, 530, 531, 562
dependent interviewing, 97, 103, 104, 108

design, *see also* Index of Vol. 29B
– consistent, 61, 63, 81, 263, 391
– effect, 130, 132, 412, 423
– multiphase, 22
– multistage, 3, 22, 40, 73, 91, 97, 240, 242, 347, 400, 402, 409, 428, 444, 492
– variable, 216
– weight, 217
design-based inference, 4, 20, 258, 259, 391, *see also* Index of Vol. 29B
deterministic checking rules, 198
deterministic model, 170
disagreement rate, 286
disclosure
– avoidance, 485
– control, 348–350, 381–396
– risk, 382, 384–386, 388, 390, 485
disposition codes, 140, 143, 148, 149
distribution method, 208, 331
domain, 246, *see also* Index of Vol. 29B
– error, 326
– estimation, 77, 83
dominance rule, 386
Do Not Call Registry, 165
double expansion estimator (DEE), 56–58, 60, 61, 63–66, 68, 69, *see also* Index of Vol. 29B
double sampling, 55, 501, 502
doubly robust, 226, 234, 240
draw by draw, 45, 46
dual frame survey, 73, 74, 76–78, 84, 85, 87
dual system estimator (DSE)
– extended DSE, 553–554, 556
– model, 542–545, 547–548
duplicate, 73, 328, 351, 355, 365–367, 518, 552, 558

E

empirical best linear unbiased predictor (EBLUP), 506, *see also* Index of Vol. 29B
editing, 158–159, 187–214, 325–331, 447, 460–462
– hard edit, 189, 200
– interactive, 189–191, 194–197, 208, 211, 212
– macro, 207–211, 326, 328, 330–331, 460, 461
– micro, 188, 207, 211, 325, 326, 328, 330, 460, 461
– non-negativity edit, 189
– overediting, 188, 189, 191, 200, 210, 211, 328
– ratio edit, 189, 211
– selective editing, 159, 188, 192–197, 207, 212, 329–330, 461
– significance editing, 192, 461
– soft edit, 189, 200
eligibility, 167–169
– local, 540, 555
– national, 539, 555
Elmo Roper, 569, 571
EM algorithm, 293, 301, 336, 358–361, 364, 365, 371
embedded replicate measurements, 287
enterprise, 189, 191, 192, 399, 400, 447, 472, 477, 479–480, 486
entropy, 33, 34, 41, 45, *see also* Index of Vol. 29B
– maximum entropy, 33, 41, 45–47, 49
environmental surveys, 399–404, 472, 487–512
equal probability sampling, 22–27, 40, 345, 519, 520
erroneous enumerations, 542, 547, 554–558
erroneous fields, 187, 213
erroneous records, 187, 189, 206, 211
error
– localization, 159, 187, 198–200, 202–204, 206, 207, 328
– nonsampling, 15, 88, 216, 405, 420, 424, 435–436, 439, 447
– rate, 158, 297, 302, 305, 352, 357–358, 366, 370, 371–373, 546, 552
– sampling, 88, 93, 95, 116, 117, 176, 188, 195, 224, 259, 385, 387, 424, 435, 447, 473, 536, 552, 558, 575, 584, 586
error components, 554–557
– coverage errors, 216, 400, 401, 403, 445, 540–542, 550, 551, 554
– total error model, 556
error-free measurements, 281, 291, 315
errors-in-variables modeling, 281
establishment, 442, 450
– survey, 18, 21, 30, 165, 167, 168, 381, 385–386, 441, 472
estimating equations, 261, 262, *see also* Index of Vol. 29B
estimation, *see also* Index of Vol. 29B
– calibration, 87, 178, 183, 184, 249, 259–261, 264, 265, 269, 270, 272, 338, 341–342, 403, 429–431
– for change, 469–470
– composite, 95, 422, 431–433, 507, 508
– for level, 467, 469
– regression, 34–36, 60–63, 65, 66, 68, 69, 338, 339, 342, 343, 348, 402, 412, 431, 468
– of total, 39, 47–48, 256–270
– variance, 20, 23–24, 33–34, 48–51, 62–64, 83–84, 153–154, 235–243, 263, 345, 433–435, 503
European Social Survey, 169, 573
evaluation
– Canada, 550–551
– UK, 552
– USA, 552
evaluation follow-up (EFU), 541, 557–558

exit polls, 567, 576, 577, 582, 583, 584–587, 591, 593
expansion estimator, 12–13, 22–27, 30, 56, 60, 65, 216–217, 236–237, 256–258, 467, 469, 549, 561, *see also* Index of Vol. 29B
– reweighted expansion estimator (REE), 60–61
exploratory data analysis, 208–209
exponential distribution, 47, 250, 252
external validity, 306

F

face-to-face interviewing, 95–96, 106, 125, 191, 281, 318, 319, 525, 546, 576, 588
false match, 352, 353, 364, 371, 378, 556
false nonmatch, 353, 371, 548, 556, 557
farm surveys, 464
fax surveys, 527
Fellegi–Holt methodology, 199–200, 202, 203, 206, 212, 329, 464
first-phase sample, 55–59, 61, 63–66, 68, 69, 113–114, 437, 469, 480, 549
flight phase, 53–54
focus groups, 574, 588–589
fractional imputation, 245–246, 484
frame, 73–74, 154, 399–401, 408, *see also* Index of Vol. 29B
– area sampling, 73–75, 85–87, 399, 400, 409, 441, 444, 445, 473–476, 479, 480, 490, 492, 493
– grid, 400, 474, 491
– list sampling, 73–75, 85–87, 92, 119, 399–401, 408–409, 444–445, 457, 477–479, 490, 493, 521–522
– out-of-date, 453–454, 462
– survey, 351, 352, 365, 447
– telephone, 87, 409–410, 526
frequency weight, 162, 344–345
Fuller's preliminary test estimator, 253
function of means, 50–51

G

Gallup (polls), 569–574, 580–583, 594
generalized M-estimation, 260
generalized regression estimator (GREG), 34–36, 178, 216, 217, 249, 338, 339, 342, 343, 432, 565, *see also* Index of Vol. 29B
general response model, 282–284
geocode, 352, 550, 556
geographic information systems (GIS), 400, 476, 481, 484, 485, 491, 492, 508
global score function, 193
gold standard, 281, 291, 309, 481
government regulation, 447, 592–594
graph sampling designs, 117
GREG (generalized regression estimator), 34–36, 178, 216, 217, 249, 338, 339, 342, 343, 432, 565, *see also* Index of Vol. 29B
gross change, 7, 90, 97, 98, 103–104, 105, 480, 499
gross difference rate, 286, 287
g-weight, 468, 469, *see also* Index of Vol. 29B

H

Hájek estimator, 47–48, 50, 468, *see also* Index of Vol. 29B
Hansen–Hurwitz estimator, 42, 121, 123
hard edit, 189, 200
hard to count (HtC) score, 546
heterogeneous independence, 543, 544
Hidiroglou–Berthelot method, 463–464
hierarchical model, 300, 508
homogeneity assumption, 255, 290, 305
Horvitz–Thompson estimator (HTE), 4, 6, 32–35, 42, 56–57, 60, 76, 121, 146, 174, 176, 258, 261–262, 270, 272, 273, 337, 339, 344, 402, 498, 502, 503, *see also* Index of Vol. 29B
household
– panel survey, 100–101, 105, 179
– survey, 30, 91, 165, 168, 169, 171, 226, 270, 317, 348, 399–404, 407–439, 472, 508, 517
Huber function, 260–262, 264, 269
Hui–Walter method, 291–294, 296–297
hybrid dialing, 136

I

identifiability, 292, 298, 307, 315
identification, 319, 320, 348, 382, 388, 389, 390, 392, 442
identity disclosure, 382, 384, 390
imputation, 92, 103–105, 158–159, 164, 187, 188, 198–200, 211, 213, 215–246, 264, 265, 318, 329, 332–336, 346, 394, 447, 464–467, 483, *see also* Index of Vol. 29B
– auxiliary value, 219, 226–227
– classes, 216, 222, 231–234, 465
– deterministic, 218, 224, 230, 231, 238, 334, 465, 466
– EM, 293, 334–335
– mean, 219, 220, 229, 334
– model approach, 223
– nearest neighbor, 199, 219, 220, 227, 229, 230, 243, 333, 466
– random, 218–221, 225, 230, 231, 244, 333, 334
– random hot-deck, 219, 221, 227, 228–230, 232, 233, 244, 245
– ratio, 219, 220, 229, 236, 246
– regression, 219–221, 223, 225–226, 230, 236, 237, 241, 244, 246, 333, 465, 466
– stochastic, 220, 465, 466
– variance, 231–233, 243, 245

incentives, 102, 108, 118, 166, 349
inclusion probability, 4, 28, 30–34, 44, 46, 47, 76, 217, 227, 337, 344, 391, 393, 428, 496, 502, *see also* Index of Vol. 29B
– first-order, 33, 39, 40, 43, 45, 217, 337
– joint, 32, 39–41, 43, 46, 48–50, 503
– second-order, 32, 33, 217, 345
inclusion weight, 337, 344, 345
inconsistency ratio, 284, 286, 295
independent classification error (ICE) assumption, 304
index of inconsistency, 285, 287, 288, 290, 296, 298
ineligible units, 168
influential error, 188, 192, 194–196, 210, 212, 330
influential units (points), 248, 259, 265, 268, 269, 278
integrated census (IC), 547, 549, 550, 552
interactive editing, 189–191, 194–197, 208, 211, 212
interactive voice response (IVR), 573, 591
internet surveys, 85, 402, 534–538, 578, 580
interview completion rate, 133, 146, 148
interviewer, 73, 86, 92, 126, 135–142, 164, 177, 318–320, 323, 327, 426, 523–524, 571, 576, 579, 587
– effects, 576–579, 585, 595
– in-person interview, 480, 572, 580
interview–reinterview, 285, 293
intruder, 349, 352, 382–393
invalidity, 310
item nonresponse, 51, 104, 163–164, 215–216, 240, 318, 332, 336, 579
iterative proportional fitting, 340, 563, *see also* Index of Vol. 29B

J

jackknife, 34, 36, 63, 182, 239, 241–243, 306, 510, 547, 563, *see also* Index of Vol. 29B
– generalized, 50–51
– variance estimation, 50–51, 84, 153–154

K

kriging, 504–506

L

labor force surveys, 94–98, 103, 169, 172, 306, 422–423
land, sampling, 66–67, 473–476, 480
landing phase, 54
latent class, 306, 359
latent class model (LCM), 289, 294–303
– classical, 295
latent variable, 289, 295, 301, 303, 307, 309, 310, 312, *see also* Index of Vol. 29B
ℓ EM software, 293, 300, 307
leverage-salience theory, 166, 172
life cycle stage, 139–140
likelihood, *see also* Index of Vol. 29B
– kernel, 295, 299, 304
– ratio, 301, 351, 354, 355, 362, 364
likely voters, 581, 582–583
linear calibration estimator, 259
linearization, 61, 66, 84, 183, 238–240, 263, 347
– variance estimation, 62, 84
linkage error, 352, 371, 373–378, 556
link-tracing designs, 115, 117–118
list sampling, 477–479
list update, 351, 367, 368
Literary Digest, 568, 569
local independence, 287, 291, 295, 296, 301, 304, 305
local score function, 193
log-linear model, 306, 393, 545, *see also* Index of Vol. 29B
– representations of LCM, 298–301
lognormal distribution, 250, 510
longitudinal analysis, 92, 98, 99, 101, 103, 419, 486
longitudinal survey, 7, 97–107, 164, 265, 415, 427, 428, 439, 479–480, 481, 483–484, 536, *see also* Index of Vol. 29B

M

macro-editing, 207–211, 326, 328, 330–331, 460, 461
mail surveys, 126, 157, 165, 169, 319, 527, 536, 577
Markov assumption, 303, 305, 306, 309
Markov latent class model (MLCM), 303–305, 306, 308
Mass Observation, 570
master sample, 30, 51, 91, 437
match, 352–358, 362, 364, 366, 371–373, 376–378
matching address, 138, 377
matching errors, 103, 373, 376, 554, 557–558, 565
matching information, 433, 556
mean imputation (MI), 219, 220, 229, 334
mean squared error (MSE), 34, 117, 162, 252, 255–258, 262, 263–264, 269, 270, 273, *see also* Index of Vol. 29B
measurement error, 103–104, 108, 162, 281–285, 290, 309, 310, 314, 343, 389–392, 509–512, 536, 554, 575
M-estimation, 259–261, 278–279
method effect, 311, 312
microaggregation, 395
microdata, 161, 208, 209, 212, 348, 349, 382–383, 388–396, 483
micro-editing, 188, 207, 211, 325, 326, 328, 330, 460, 461
minimum variance design, 10, 25

misclassification, 87, 289, 290, 305, 306, 394, 551
missing at random (MAR), 222, 332, 335, 336, *see also* Index of Vol. 29B
missing completely at random (MCAR), 222, 332, 335, 336, *see also* Index of Vol. 29B
missing data, 59, 64, 104, 105, 146, 163, 164, 174, 215, 242, 243, 246, 332, 335, 464–466
Mitofsky–Waksberg method, 129, 130, 522, 577
mixed-mode, 169, 537–538
– data collection, 95, 192, 537
– surveys, 126, 537, 538, 580
mobile computers, 481
mode, 165, 166, 174, 281
– of collection, 165, 191, 192, 293–294, 310, 404, 405, 443
– effects in multiple frame surveys, 86
multinomial distribution, 123, 291, 292, 295, 545
multiphase design, 22, 437
multiple frame survey, 7, 8, 71–88, 175, 445, 479, 489
– design, 71, 75, 85, 479
– overlapping, 75–76, 85
– screening, 73–75, 78, 83, 85, 87
multiple imputation, 215, 243–246, 335–336, 394, 484
multiple mode survey, 480, 580
multiple occasions, 59
multiplicative weighting, 338, 340–341
multiplicity sampling, 115–119
multistage, 3, 15, 31–32, 21–22, 40, 73, 91, 97, 240, 242, 347, 400, 402, 409, 428, 444, 492–493, 528
multitrait-multimethod (MTMM) approach, 312, 313
multivariate outliers, 209, 248

N

National Council on Public Polls (NCPP), 574, 575, 584
National Health Interview Survey (NHIS), 86, 93, 133, 293, 410, 591, 592
National Resources Inventory (NRI), 66, 68, 69, 474–476, 479–480, 481, 483–485, 486, 492
National Survey of America's Families, 74, 87
National Survey of College Graduates, 74, 83
natural resource surveys, 66, 471–486
nearest neighbor imputation (NNI), 199, 219, 220, 227, 229, 230, 243, 333, 466, *see also* Index of Vol. 29B
net change, 90, 91, 95, 98, 103, 499
net difference rate, 291
network sampling, 115–117
news polls, 569
Neyman, 4, 14–20, 51, 55, 56
– allocation, 28–29, 114, 452, 453, 457
noise addition, 349, 388, 392, 394
noncontact, 134, 142, 147, 165, 169, 178, 522, 576, 590
noncoverage weight adjustment, 174, 175
noncritical stream, 192, 211, 212, 330
nonmatch, 138, 353, 356, 360, 364, 371–373, 376, 377, 548, 556, 557
nonresponse, 101, 147–148, 163–167, 170, 172, 175, 178–181, 183, 215, 221–223, 318, 332, 336–337, 413, 430, 431, 453, 488, 536, 576, 590, *see also* Index of Vol. 29B
– bias, 144, 166, 170–174, 180, 184, 215, 222, 224–225, 234, 431, 522, 590–591
– item, 51, 104, 163–164, 215–216, 240, 318, 332, 336, 579
– mechanism, 150, 217, 221–222, 223, 224, 235
– missing at random (MAR), 222, 332, 335, 336
– missing completely at random (MCAR), 222, 332, 335, 336
– model approach, 223
– in multiple frame surveys, 86, 87
– not missing at random (NMAR), 222, 323, 332, 336
– persistent nonrespondents, 174
– variance, 216, 231–233, 235
– unit, 160, 164, 170, 215, 336, 464
– wave, 101, 102, 104, 105, 164
nonsampling error, 15, 88, 216, 405, 420, 424, 435–436, 439, 447
not missing at random (NMAR), 222, 323, 332, 336, *see also* Index of Vol. 29B

O

online
– panels, 85, 86, 538, 580
– surveys, 579–580
operating structure, 442, 443
opinion polls, 399, 567–595
optimal allocation, 28–29, 74, 114, 452–453
optimal cutoff, 254, 255, 261
ordered systematic sampling, 24–25, 43
order sampling, 36, 46–47, *see also* Index of Vol. 29B
outliers, 157, 160, 177, 189, 209, 210, 214, 270, 328, 331, 403, 461, 464, 593, *see also* Index of Vol. 29B
– detection, 160, 193, 198, 209, 230, 247–249, 271, 462–463
– influential, 462
– nonrepresentative, 160, 247, 279, 462
– representative, 160, 247, 248, 462
– weights, 270
over-allocation, 453, 454
overcount, 540–542, 547–555
overcoverage follow-up survey, 553, 554
oversampling, 106, 305, 500

P

p% rule, 386
panel attrition, 96
panel conditioning, 96, 97, 101, 103, 108, 420, 533
panel survey, 66, 90, 94–107, 313–315, *see also* Index of Vol. 29B
– cross-national, 101
– freshening a panel sample, 99, 107, 108
– household, 100–101, 105, 179
– overlapping panel, 100
– rotating, 94–97, 101
paradata, 185, 408, 425, 439
parallel assumption, 287–289, 291
Pareto distribution, 250, 252
parse, 369, 370
path analysis, 309, *see also* Index of Vol. 29B
Peano key, 495, 496
period estimate, 94
permanent random numbers (PRNs), 21, 28, 30, 36, 92, 458, 459, 479
perturbation, 394–395, 500, 505, 506
physical support, 506, 507
plot design, 474, 478, 506–508
Poisson distribution, 111, 392, 545, 547
Poisson sampling, 33, 40–41, 62, 64, 66, 391, 393, 458, 459, *see also* Index of Vol. 29B
– conditional, 33, 45–46
population, *see also* Index of Vol. 29B
– inference, 473, 502
– proportion, 285, 390, 429
– register, 4, 6, 408, 409, 427, 550, 565
– size, estimation of, 75, 77, 110, 119, 123, 428, 498, 503, 543, 544
population-based, 177, 178, 180, 184, 526
ported numbers, 138
postenumeration survey (PES), 540–542, 544–546, 552, 555–558, 565
postrandomization method (PRAM), 394
poststratification, 86, 87, 151, 152, 175, 182, 183, 338–340, 347, 431
potential match, 352, 392
prediction, *see also* Index of Vol. 29B
– disclosure, 382, 385–387
– inference, 9–13, 16–17, 19–23, 25–27, 35–37
predictive dialing, 135, 136
primary sampling unit (PSUs), 31–32, 63, 67, 74, 91, 110, 129, 388, 411, 473, 490, 520, 545, 579
prior-posterior rule, 350, 386
probabilities of selection, 23, 127, 134, 146, 150–151, 167, 174, 182, 523, 582, 585
proportional allocation, 58, 112, 411, 452, 453
pseudo maximum likelihood (PML) estimation, 81–83, 306

Q

quasi-identifier, 351
quasi-simplex model, 313–315
questionnaire design, 162, 166, 281, 408
quota sampling, 571, 581

R

raking, 80, 83, 105, 152, 179, 180, 181, 183, 340, 432, 455, 469, 485, 563, *see also* Index of Vol. 29B
random-digit dialing (RDD), 7, 8, 74, 111, 125–154, 293, 401, 402, 513, 520–522, 572, 577
random error, 188, 191, 198–200, 202
random hot-deck imputation (RHDI), 219, 221, 227, 228–230, 232, 233, 244, 245
random imputation, 218–221, 225, 230, 231, 244, 333, 334
randomization inference, 4–6, 12–16, 21, 22
random sampling, 4, 6, 21–24, 28, 110–115, 117, 123, 125, 296, 400, 412, 446, 451, 452, 490, 492, 528
ranked set sampling (RSS), 500–502
Rao–Sampford design, 42, 44, 227
rare population, 8, 72, 85, 86, 90, 91, 109–124, 168, 426
ratio estimation, 484
ratio estimator, 6, 9–13, 26–27, 29–30, 80, 83, 265, 275, 277, 412, 463, 468, 469, 484, 502, 544, 586, *see also* Index of Vol. 29B
rationality of public opinion, 594
RDD (random-digit dialing), 7, 8, 74, 111, 125–154, 293, 401, 402, 513, 520–522, 572, 577
record linkage, 160, 161, 351–380, 389, 540, 552, 556, 565
record swapping, 395
reference period, 97, 291, 419, 448, 449, 467
refusal conversion, 140, 166
registration-based sampling (RBS), 578
regression, 18, 20, 22, 35, 49, 50, 178, 179, 196, 197, 209, 219, 223, 225, 232, 333, 427, 468, 499, 508, 510, *see also* Index of Vol. 29B
– estimation, 34–36, 60–63, 65, 66, 68, 69, 338, 339, 342, 343, 348, 402, 412, 431, 468
– imputation, 219–221, 223, 225–226, 230, 236, 237, 241, 244, 246, 333, 465, 466
rejective procedure, 45, 46, 129, 130
reliability ratio, 284, 309, 311
reliability weight, 199, 200, 204, 206
remote sensing, 474–475, 481, 488, 492
repeated observations, 479–480
repeated survey, 90, 91–94, 95, 108, 415–418, 420–425, 431, *see also* Index of Vol. 29B
replication, 93, 137–139, 142, 154, 182, 183, 240–242, 269, 313, 396, 434, 484
– variance, 484
– variance estimator, 62, 63, 65, 66, 240

resolution completion rate, 132, 145, 147
response, *see also* Index of Vol. 29B
– burden, 90, 92, 101, 102, 104, 151, 402, 404, 420, 448, 450, 459–460
– probabilities, 170, 174, 178, 216, 221–223, 225, 226, 229, 232–234, 238, 239, 431
– propensities, 171–174, 177, 184
– rates, 86, 101, 126, 137, 144–146, 165, 166, 167–170, 184, 536, 590–591
– stochastic model of, 170
reverse framework, 235, 238–240
rolling sample design, 541, 558, 559, 564
Roper Center, 574
rotating panel survey, 94–97, 101
rotation, 457–460, 470, *see also* Index of Vol. 29B
– group bias, 96
– scheme, 95, 416, 419–420
rounding, 387
– error, 198, 202, 203
route instruction, 320, 326
routing error, 326

S

safe data, 382–383
safe setting, 382, 394
Sampford design, 42, 44, 227, *see also* Index of Vol. 29B
sample allocation, 28–29, 114, 401–402, 411, 413, 501, 546, 551
sample-based, 177, 184, 542, *see also* Index of Vol. 29B
Sample Survey of Retail Stores, 74
sampling, 9–37, 39, 41, 55, 106, 109, 441, 473–480, *see also* Index of Vol. 29B
– adaptive, 115–123, 497–498
– design, 6, 21–36, 40, 43–45, 110, 115–123, 257, 488–490
– error, 88, 93, 95, 116, 117, 176, 188, 195, 224, 259, 385, 387, 424, 435, 447, 473, 536, 552, 558, 575, 584, 586
– list assisted, 130, 578
– multistage, 3, 15, 21–22, 31–32, 74, 168, 409, 473, 492
– with replacement, 21–23, 30–32, 41–42, 47
– without replacement, 21, 23–24, 32–34, 42, 47–48, 236, 446
– telephone, 128, 523–524, 525, 578
– two-stage, 40, 74, 129, 283, 284, 492
– unit, 62, 91, 109, 400, 442, 450–451, 517, 545
– weight, 68, 154, 174, 228, 563
sampling frame, 6–8, 30–31, 73–75, 125–131, 133–134, 146, 150, 173, 175, 270, 442, 444–445, 478, 487, 490, 492, 500–501, 507
– incomplete, 74, 75
sampling, two-phase 113–115
sampling variance (SV), 117, 231, 243, 283
satellite imagery, 475, 476, 484
Scientists and Engineers Statistical Data System, 74, 91
score function, 461, *see also* Index of Vol. 29B
– global, 193
– local, 193
score model, 312, 313
score variance, 284, 314, 315
screener completion rate, 132, 146, 148
screening, 73–75, 109–111, 113, 147, 525
seam effect, 104
second-phase sample, 56, 58, 64, 67, 437
second-phase strata, 55, 68
selective editing, 159, 188, 192–197, 207, 212, 329–330, 461
self-efficiency, 245
self interviewing, 319
self-representing strata, 455–457
self-weighting design, 238, 414, 415
semisupervised learning, 371, 372
sensitivity measure, 386
shopping center sampling, 401, 527–529
sign error, 201, 203
simple random sampling (SRS), 3, 21–24, 64, 110, 282, 296, 412, 451, 489, 492–495, 501, 519, *see also* Index of Vol. 29B
simple response variance (SRV), 283, 287
skewed distribution, 249, 251, 264, 403
skip instruction, 320
small area estimation, 87, 385, 403, 435, 439, 483–484, 561, *see also* Index of Vol. 29B
Social Science Research Council, 571, 581
social survey, 384, 407, 441, 573
spatial designs, 492–495, 503–506
spatially balanced, 493, 496, 500, 503
spatially constrained, 492–493, 503
splitting method, 44–45
standardization, 175, 368–370, 436
stationary transitions, 304
statistical unit, 442, 443, 450
stochastic imputation, 220, 465, 466
stochastic nonresponse model, 170
strategy, minimax, 504, 505
stratification, 21, 28–30, 74, 85, 111, 131–132, 257, 401–402, 411, 450, 479, 493–497, 546
– microstrata, 460
– optimal allocation, 28–29, 74, 114, 452–453
– self-representing strata, 455–457
– take-all stratum, 247, 451, 456
stratified sampling, 4, 6, 21, 28–30, 74, 111–115, 117, 120, 122, 123, 257–258, 265, 479, *see also* Index of Vol. 29B
– disproportionate, 92, 106, 111, 112, 123
– proportional allocation, 58, 112, 411, 452, 453
stratum jumpers, 36, 250, 270–278
straw polls, 568

string comparator, 362–364, 370
structural equation model (SEM), 98, 162, 281, 309–311, *see also* Index of Vol. 29B
structural submodel, 300, 310
successive difference replication (SDR), 563
superpopulation, 47, 55, 489, *see also* Index of Vol. 29B
suppression, 107, 349, 383, 387, 394
surveys
– aerial, 476, 481
– agricultural, 399–401, 403, 472, 473, 475, 477–480, 484
– business, 270–272, 401, 403, 441–470
– complex, 305–306, 412, 423
– design, 3–4, 6, 106, 408–415, 487
– environmental, 399–404, 472, 487–512
– establishment, 18, 21, 165, 167, 168, 381, 385–386, 441
– farm, 464
– fax, 527
– household, 29–32, 91, 165, 168, 169, 171, 226, 270, 317, 348, 399–404, 407–439, 472, 508, 517
– integration, 427, 436–437
– internet, 85, 402, 534–538, 578, 580
– mail, 126, 157, 165, 169, 319, 527, 536, 577
– multiple mode, 480, 580
– online, 579–580
– panel, 66, 90, 94–107, 313–315
– repeated, 90, 91–94, 95, 108, 415–418, 420–425, 431
– social, 384, 407, 441, 573
– telephone, 87, 110, 111, 439, 517–526, 577–578, 591
– web, 126, 480, 513, 534–538
survey feedback, 442, 445
Survey of Consumer Finances, U.S., 73, 82, 86
Survey of Doctorate Recipients, 74
synchronized sampling, 459
synthetic data, 394
synthetic estimator, 545, 561, *see also* Index of Vol. 29B
systematic error, 162, 198, 200, 212
systematic sampling, 21–22, 24–25, 42–43, 49, 494, *see also* Index of Vol. 29B
– randomized, 33–34, 43

T

target unit, 443, 450
Taylor series linearization, 61, 84, 181, 182, 238, 239
telephone
– banks, 128, 130, 517
– frame, 87, 409–410, 526
– interviewing, 96, 139, 318, 407, 525
– sampling, 128, 523–524, 525, 578
– surveys, 87, 134, 439, 517–526, 577–578, 591
telescoping effect, 90, 96–97
tessellation, 400, 474, 476, 489, 495–497
test-retest reinterview, 283, 285, 286, 293, 296
thousand-error, 198, 200, 211
three-phase sample, 65–66, 142, 474
time-in-sample bias, 96, 100, 101, 500
top coding, 107, 393
top-down method, 462–463
topic salience, 166
total, 6, 9–11, 21–37, 39, 42, 56, 76, 106, 152, 175, 227, 235, 240, 256, 381, 427, 543, *see also* Index of Vol. 29B
tracking polls, 574, 583–584
training data, 196, 352, 355, 366, 371
trend estimation, 92, 100, 422, 480, *see also* Index of Vol. 29B
trimming of weights, 153, 177
truth data, 362, 373
tuning constant, 261, 262–263, *see also* Index of Vol. 29B
two-phase framework (TPF), 59, 235, 236–238
two-phase random sampling, 55, 59, 113, 114, 182, 467, 469, 501
two-stage sampling, 40, 74, 129, 283, 492, *see also* Index of Vol. 29B
typographical error, 352, 355, 357, 361

U

undercount, 540, 542, 546, 547, 555
undercoverage, 71, 128–129, 133, 146, 151–152, 352, 445, 553
– survey, 553, 554
unequal probability sampling, 5, 20–22, 30–34, 40–47, 305, *see also* Index of Vol. 29B
uniqueness, 390, 393
unit nonresponse, 160, 164, 170, 215, 336, 464, *see also* Index of Vol. 29B
unlabeled, 371, 372
unlisted numbers, 128, 130, 518, 519
unresolved telephone numbers, 134, 145, 147
unsupervised learning, 255, 357, 372
utility, 383, 384

V

validation, 212, 230, 272, 448
variance estimation, 23–24, 26–27, 48–51, 62–64, 83–84, 153–154, 235–243, 263, 345, 433–435, 503, *see also* Index of Vol. 29B
– jackknife, 50–51, 84, 153–154
virgin cases, 136, 139

W

wave nonresponse, 101, 102, 104, 105, 164
Web surveys, 125, 126, 480, 513, 534–538

Weibull distribution, 250, 252
weighting, 104–106, 146, 163, 174–181, 338, 339, 340, 343, 427–435
– adjustment, 87, 104, 164, 174, 177–178, 183, 318, 336–343
– class adjustment, 163, 178, 179
– integrated, 430
– linear, 338–341
– multiplicative, 338, 340–341
weights, 80, 146–147, 150, 152–154, 174, 175, 181, 250, 262, 274, 318, 337, 340, 344, *see also* Index of Vol. 29B
– smoothing of, 272–274
Winsorization, 257–258, 274, 278
– cutoff, 251, 252, 255, 257, 274
– once-Winsorized mean, 252–253, 255, 256
– Searl's Winsorized mean, 251–252, 254
– type I and type II, 257, 260
– of weights, 250, 274
working residential numbers (WRNs), 129, 131, 522, 524
– cell-only households, 133, 173, 526

Z

zero banks, 130

Handbook of Statistics
Contents of Previous Volumes

Volume 1. Analysis of Variance
Edited by P.R. Krishnaiah
1980 xviii + 1002 pp.

1. Estimation of Variance Components by C.R. Rao and J. Kleffe
2. Multivariate Analysis of Variance of Repeated Measurements by N.H. Timm
3. Growth Curve Analysis by S. Geisser
4. Bayesian Inference in MANOVA by S.J. Press
5. Graphical Methods for Internal Comparisons in ANOVA and MANOVA by R. Gnanadesikan
6. Monotonicity and Unbiasedness Properties of ANOVA and MANOVA Tests by S. Das Gupta
7. Robustness of ANOVA and MANOVA Test Procedures by P.K. Ito
8. Analysis of Variance and Problems under Time Series Models by D.R. Brillinger
9. Tests of Univariate and Multivariate Normality by K.V. Mardia
10. Transformations to Normality by G. Kaskey, B. Kolman, P.R. Krishnaiah and L. Steinberg
11. ANOVA and MANOVA: Models for Categorical Data by V.P. Bhapkar
12. Inference and the Structural Model for ANOVA and MANOVA by D.A.S. Fraser
13. Inference Based on Conditionally Specified ANOVA Models Incorporating Preliminary Testing by T.A. Bancroft and C.-P. Han
14. Quadratic Forms in Normal Variables by C.G. Khatri
15. Generalized Inverse of Matrices and Applications to Linear Models by S.K. Mitra
16. Likelihood Ratio Tests for Mean Vectors and Covariance Matrices by P.R. Krishnaiah and J.C. Lee
17. Assessing Dimensionality in Multivariate Regression by A.J. Izenman
18. Parameter Estimation in Nonlinear Regression Models by H. Bunke
19. Early History of Multiple Comparison Tests by H.L. Harter
20. Representations of Simultaneous Pairwise Comparisons by A.R. Sampson
21. Simultaneous Test Procedures for Mean Vectors and Covariance Matrices by P.R. Krishnaiah, G.S. Mudholkar and P. Subbaiah
22. Nonparametric Simultaneous Inference for Some MANOVA Models by P.K. Sen

23. Comparison of Some Computer Programs for Univariate and Multivariate Analysis of Variance by R.D. Bock and D. Brandt
24. Computations of Some Multivariate Distributions by P.R. Krishnaiah
25. Inference on the Structure of Interaction Two-Way Classification Model by P.R. Krishnaiah and M. Yochmowitz

Volume 2. Classification, Pattern Recognition and Reduction of Dimensionality

Edited by P.R. Krishnaiah and L.N. Kanal

1982 xxii + 903 pp.

1. Discriminant Analysis for Time Series by R.H. Shumway
2. Optimum Rules for Classification into Two Multivariate Normal Populations with the Same Covariance Matrix by S. Das Gupta
3. Large Sample Approximations and Asymptotic Expansions of Classification Statistics by M. Siotani
4. Bayesian Discrimination by S. Geisser
5. Classification of Growth Curves by J.C. Lee
6. Nonparametric Classification by J.D. Broffitt
7. Logistic Discrimination by J.A. Anderson
8. Nearest Neighbor Methods in Discrimination by L. Devroye and T.J. Wagner
9. The Classification and Mixture Maximum Likelihood Approaches to Cluster Analysis by G.J. McLachlan
10. Graphical Techniques for Multivariate Data and for Clustering by J.M. Chambers and B. Kleiner
11. Cluster Analysis Software by R.K. Blashfield, M.S. Aldenderfer and L.C. Morey
12. Single-link Clustering Algorithms by F.J. Rohlf
13. Theory of Multidimensional Scaling by J. de Leeuw and W. Heiser
14. Multidimensional Scaling and its Application by M. Wish and J.D. Carroll
15. Intrinsic Dimensionality Extraction by K. Fukunaga
16. Structural Methods in Image Analysis and Recognition by L.N. Kanal, B.A. Lambird and D. Lavine
17. Image Models by N. Ahuja and A. Rosenfield
18. Image Texture Survey by R.M. Haralick
19. Applications of Stochastic Languages by K.S. Fu
20. A Unifying Viewpoint on Pattern Recognition by J.C. Simon, E. Backer and J. Sallentin
21. Logical Functions in the Problems of Empirical Prediction by G.S. Lbov
22. Inference and Data Tables and Missing Values by N.G. Zagoruiko and V.N. Yolkina
23. Recognition of Electrocardiographic Patterns by J.H. van Bemmel
24. Waveform Parsing Systems by G.C. Stockman
25. Continuous Speech Recognition: Statistical Methods by F. Jelinek, R.L. Mercer and L.R. Bahl
26. Applications of Pattern Recognition in Radar by A.A. Grometstein and W.H. Schoendorf

27. White Blood Cell Recognition by F.S. Gelsema and G.H. Landweerd
28. Pattern Recognition Techniques for Remote Sensing Applications by P.H. Swain
29. Optical Character Recognition – Theory and Practice by G. Nagy
30. Computer and Statistical Considerations for Oil Spill Identification by Y.T. Chien and T.J. Killeen
31. Pattern Recognition in Chemistry by B.R. Kowalski and S. Wold
32. Covariance Matrix Representation and Object-Predicate Symmetry by T. Kaminuma, S. Tomita and S. Watanabe
33. Multivariate Morphometrics by R.A. Reyment
34. Multivariate Analysis with Latent Variables by P.M. Bentler and D.G. Weeks
35. Use of Distance Measures, Information Measures and Error Bounds in Feature Evaluation by M. Ben-Bassat
36. Topics in Measurement Selection by J.M. Van Campenhout
37. Selection of Variables Under Univariate Regression Models by P.R. Krishnaiah
38. On the Selection of Variables Under Regression Models Using Krishnaiah's Finite Intersection Tests by J.L. Schmidhammer
39. Dimensionality and Sample Size Considerations in Pattern Recognition Practice by A.K. Jain and B. Chandrasekaran
40. Selecting Variables in Discriminant Analysis for Improving upon Classical Procedures by W. Schaafsma
41. Selection of Variables in Discriminant Analysis by P.R. Krishnaiah

Volume 3. Time Series in the Frequency Domain
Edited by D.R. Brillinger and P.R. Krishnaiah
1983 xiv + 485 pp.

1. Wiener Filtering (with emphasis on frequency-domain approaches) by R.J. Bhansali and D. Karavellas
2. The Finite Fourier Transform of a Stationary Process by D.R. Brillinger
3. Seasonal and Calendar Adjustment by W.S. Cleveland
4. Optimal Inference in the Frequency Domain by R.B. Davies
5. Applications of Spectral Analysis in Econometrics by C.W.J. Granger and R. Engle
6. Signal Estimation by E.J. Hannan
7. Complex Demodulation: Some Theory and Applications by T. Hasan
8. Estimating the Gain of a Linear Filter from Noisy Data by M.J. Hinich
9. A Spectral Analysis Primer by L.H. Koopmans
10. Robust-Resistant Spectral Analysis by R.D. Martin
11. Autoregressive Spectral Estimation by E. Parzen
12. Threshold Autoregression and Some Frequency-Domain Characteristics by J. Pemberton and H. Tong
13. The Frequency-Domain Approach to the Analysis of Closed-Loop Systems by M.B. Priestley
14. The Bispectral Analysis of Nonlinear Stationary Time Series with Reference to Bilinear Time-Series Models by T. Subba Rao
15. Frequency-Domain Analysis of Multidimensional Time-Series Data by E.A. Robinson

16. Review of Various Approaches to Power Spectrum Estimation by P.M. Robinson
17. Cumulants and Cumulant Spectra by M. Rosenblatt
18. Replicated Time-Series Regression: An Approach to Signal Estimation and Detection by R.H. Shumway
19. Computer Programming of Spectrum Estimation by T. Thrall
20. Likelihood Ratio Tests on Covariance Matrices and Mean Vectors of Complex Multivariate Normal Populations and their Applications in Time Series by P.R. Krishnaiah, J.C. Lee and T.C. Chang

Volume 4. Nonparametric Methods
Edited by P.R. Krishnaiah and P.K. Sen
1984 xx + 968 pp.

1. Randomization Procedures by C.B. Bell and P.K. Sen
2. Univariate and Multivariate Multisample Location and Scale Tests by V.P. Bhapkar
3. Hypothesis of Symmetry by M. Hušková
4. Measures of Dependence by K. Joag-Dev
5. Tests of Randomness against Trend or Serial Correlations by G.K. Bhattacharyya
6. Combination of Independent Tests by J.L. Folks
7. Combinatorics by L. Takács
8. Rank Statistics and Limit Theorems by M. Ghosh
9. Asymptotic Comparison of Tests – A Review by K. Singh
10. Nonparametric Methods in Two-Way Layouts by D. Quade
11. Rank Tests in Linear Models by J.N. Adichie
12. On the Use of Rank Tests and Estimates in the Linear Model by J.C. Aubuchon and T.P. Hettmansperger
13. Nonparametric Preliminary Test Inference by A.K.Md.E. Saleh and P.K. Sen
14. Paired Comparisons: Some Basic Procedures and Examples by R.A. Bradley
15. Restricted Alternatives by S.K. Chatterjee
16. Adaptive Methods by M. Hušková
17. Order Statistics by J. Galambos
18. Induced Order Statistics: Theory and Applications by P.K. Bhattacharya
19. Empirical Distribution Function by F. Csáki
20. Invariance Principles for Empirical Processes by M. Csörgő
21. M-, L- and R-estimators by J. Jurečková
22. Nonparametric Sequential Estimation by P.K. Sen
23. Stochastic Approximation by V. Dupač
24. Density Estimation by P. Révész
25. Censored Data by A.P. Basu
26. Tests for Exponentiality by K.A. Doksum and B.S. Yandell
27. Nonparametric Concepts and Methods in Reliability by M. Hollander and F. Proschan
28. Sequential Nonparametric Tests by U. Müller-Funk
29. Nonparametric Procedures for some Miscellaneous Problems by P.K. Sen
30. Minimum Distance Procedures by R. Beran

31. Nonparametric Methods in Directional Data Analysis by S.R. Jammalamadaka
32. Application of Nonparametric Statistics to Cancer Data by H.S. Wieand
33. Nonparametric Frequentist Proposals for Monitoring Comparative Survival Studies by M. Gail
34. Meteorological Applications of Permutation Techniques Based on Distance Functions by P.W. Mielke Jr
35. Categorical Data Problems Using Information Theoretic Approach by S. Kullback and J.C. Keegel
36. Tables for Order Statistics by P.R. Krishnaiah and P.K. Sen
37. Selected Tables for Nonparametric Statistics by P.K. Sen and P.R. Krishnaiah

Volume 5. Time Series in the Time Domain
Edited by E.J. Hannan, P.R. Krishnaiah and M.M. Rao
1985 xiv + 490 pp.

1. Nonstationary Autoregressive Time Series by W.A. Fuller
2. Non-Linear Time Series Models and Dynamical Systems by T. Ozaki
3. Autoregressive Moving Average Models, Intervention Problems and Outlier Detection in Time Series by G.C. Tiao
4. Robustness in Time Series and Estimating ARMA Models by R.D. Martin and V.J. Yohai
5. Time Series Analysis with Unequally Spaced Data by R.H. Jones
6. Various Model Selection Techniques in Time Series Analysis by R. Shibata
7. Estimation of Parameters in Dynamical Systems by L. Ljung
8. Recursive Identification, Estimation and Control by P. Young
9. General Structure and Parametrization of ARMA and State-Space Systems and its Relation to Statistical Problems by M. Deistler
10. Harmonizable, Cramér, and Karhunen Classes of Processes by M.M. Rao
11. On Non-Stationary Time Series by C.S.K. Bhagavan
12. Harmonizable Filtering and Sampling of Time Series by D.K. Chang
13. Sampling Designs for Time Series by S. Cambanis
14. Measuring Attenuation by M.A. Cameron and P.J. Thomson
15. Speech Recognition Using LPC Distance Measures by P.J. Thomson and P. de Souza
16. Varying Coefficient Regression by D.F. Nicholls and A.R. Pagan
17. Small Samples and Large Equations Systems by H. Theil and D.G. Fiebig

Volume 6. Sampling
Edited by P.R. Krishnaiah and C.R. Rao
1988 xvi + 594 pp.

1. A Brief History of Random Sampling Methods by D.R. Bellhouse
2. First Course in Survey Sampling by T. Dalenius
3. Optimality of Sampling Strategies by A. Chaudhuri
4. Simple Random Sampling by P.K. Pathak

5. On Single Stage Unequal Probability Sampling by V.P. Godambe and M.E. Thompson
6. Systematic Sampling by D.R. Bellhouse
7. Systematic Sampling with Illustrative Examples by M.N. Murthy and T.J. Rao
8. Sampling in Time by D.A. Binder and M.A. Hidiroglou
9. Bayesian Inference in Finite Populations by W.A. Ericson
10. Inference Based on Data from Complex Sample Designs by G. Nathan
11. Inference for Finite Population Quantiles by J. Sedransk and P.J. Smith
12. Asymptotics in Finite Population Sampling by P.K. Sen
13. The Technique of Replicated or Interpenetrating Samples by J.C. Koop
14. On the Use of Models in Sampling from Finite Populations by I. Thomsen and D. Tesfu
15. The Prediction Approach to Sampling Theory by R.M. Royall
16. Sample Survey Analysis: Analysis of Variance and Contingency Tables by D.H. Freeman Jr
17. Variance Estimation in Sample Surveys by J.N.K. Rao
18. Ratio and Regression Estimators by P.S.R.S. Rao
19. Role and Use of Composite Sampling and Capture-Recapture Sampling in Ecological Studies by M.T. Boswell, K.P. Burnham and G.P. Patil
20. Data-based Sampling and Model-based Estimation for Environmental Resources by G.P. Patil, G.J. Babu, R.C. Hennemuth, W.L. Meyers, M.B. Rajarshi and C. Taillie
21. On Transect Sampling to Assess Wildlife Populations and Marine Resources by F.L. Ramsey, C.E. Gates, G.P. Patil and C. Taillie
22. A Review of Current Survey Sampling Methods in Marketing Research (Telephone, Mall Intercept and Panel Surveys) by R. Velu and G.M. Naidu
23. Observational Errors in Behavioural Traits of Man and their Implications for Genetics by P.V. Sukhatme
24. Designs in Survey Sampling Avoiding Contiguous Units by A.S. Hedayat, C.R. Rao and J. Stufken

Volume 7. Quality Control and Reliability
Edited by P.R. Krishnaiah and C.R. Rao
1988 xiv + 503 pp.

1. Transformation of Western Style of Management by W. Edwards Deming
2. Software Reliability by F.B. Bastani and C.V. Ramamoorthy
3. Stress–Strength Models for Reliability by R.A. Johnson
4. Approximate Computation of Power Generating System Reliability Indexes by M. Mazumdar
5. Software Reliability Models by T.A. Mazzuchi and N.D. Singpurwalla
6. Dependence Notions in Reliability Theory by N.R. Chaganty and K. Joagdev
7. Application of Goodness-of-Fit Tests in Reliability by B.W. Woodruff and A.H. Moore
8. Multivariate Nonparametric Classes in Reliability by H.W. Block and T.H. Savits

9. Selection and Ranking Procedures in Reliability Models by S.S. Gupta and S. Panchapakesan
10. The Impact of Reliability Theory on Some Branches of Mathematics and Statistics by P.J. Boland and F. Proschan
11. Reliability Ideas and Applications in Economics and Social Sciences by M.C. Bhattacharjee
12. Mean Residual Life: Theory and Applications by F. Guess and F. Proschan
13. Life Distribution Models and Incomplete Data by R.E. Barlow and F. Proschan
14. Piecewise Geometric Estimation of a Survival Function by G.M. Mimmack and F. Proschan
15. Applications of Pattern Recognition in Failure Diagnosis and Quality Control by L.F. Pau
16. Nonparametric Estimation of Density and Hazard Rate Functions when Samples are Censored by W.J. Padgett
17. Multivariate Process Control by F.B. Alt and N.D. Smith
18. QMP/USP – A Modern Approach to Statistical Quality Auditing by B. Hoadley
19. Review About Estimation of Change Points by P.R. Krishnaiah and B.Q. Miao
20. Nonparametric Methods for Changepoint Problems by M. Csörgő and L. Horváth
21. Optimal Allocation of Multistate Components by E. El-Neweihi, F. Proschan and J. Sethuraman
22. Weibull, Log-Weibull and Gamma Order Statistics by H.L. Herter
23. Multivariate Exponential Distributions and their Applications in Reliability by A.P. Basu
24. Recent Developments in the Inverse Gaussian Distribution by S. Iyengar and G. Patwardhan

Volume 8. Statistical Methods in Biological and Medical Sciences
Edited by C.R. Rao and R. Chakraborty
1991 xvi + 554 pp.

1. Methods for the Inheritance of Qualitative Traits by J. Rice, R. Neuman and S.O. Moldin
2. Ascertainment Biases and their Resolution in Biological Surveys by W.J. Ewens
3. Statistical Considerations in Applications of Path Analytical in Genetic Epidemiology by D.C. Rao
4. Statistical Methods for Linkage Analysis by G.M. Lathrop and J.M. Lalouel
5. Statistical Design and Analysis of Epidemiologic Studies: Some Directions of Current Research by N. Breslow
6. Robust Classification Procedures and their Applications to Anthropometry by N. Balakrishnan and R.S. Ambagaspitiya
7. Analysis of Population Structure: A Comparative Analysis of Different Estimators of Wright's Fixation Indices by R. Chakraborty and H. Danker-Hopfe
8. Estimation of Relationships from Genetic Data by E.A. Thompson
9. Measurement of Genetic Variation for Evolutionary Studies by R. Chakraborty and C.R. Rao

10. Statistical Methods for Phylogenetic Tree Reconstruction by N. Saitou
11. Statistical Models for Sex-Ratio Evolution by S. Lessard
12. Stochastic Models of Carcinogenesis by S.H. Moolgavkar
13. An Application of Score Methodology: Confidence Intervals and Tests of Fit for One-Hit-Curves by J.J. Gart
14. Kidney-Survival Analysis of IgA Nephropathy Patients: A Case Study by O.J.W.F. Kardaun
15. Confidence Bands and the Relation with Decision Analysis: Theory by O.J.W.F. Kardaun
16. Sample Size Determination in Clinical Research by J. Bock and H. Toutenburg

Volume 9. Computational Statistics
Edited by C.R. Rao
1993 xix + 1045 pp.

1. Algorithms by B. Kalyanasundaram
2. Steady State Analysis of Stochastic Systems by K. Kant
3. Parallel Computer Architectures by R. Krishnamurti and B. Narahari
4. Database Systems by S. Lanka and S. Pal
5. Programming Languages and Systems by S. Purushothaman and J. Seaman
6. Algorithms and Complexity for Markov Processes by R. Varadarajan
7. Mathematical Programming: A Computational Perspective by W.W. Hager, R. Horst and P.M. Pardalos
8. Integer Programming by P.M. Pardalos and Y. Li
9. Numerical Aspects of Solving Linear Least Squares Problems by J.L. Barlow
10. The Total Least Squares Problem by S. van Huffel and H. Zha
11. Construction of Reliable Maximum-Likelihood-Algorithms with Applications to Logistic and Cox Regression by D. Böhning
12. Nonparametric Function Estimation by T. Gasser, J. Engel and B. Seifert
13. Computation Using the OR Decomposition by C.R. Goodall
14. The EM Algorithm by N. Laird
15. Analysis of Ordered Categorial Data through Appropriate Scaling by C.R. Rao and P.M. Caligiuri
16. Statistical Applications of Artificial Intelligence by W.A. Gale, D.J. Hand and A.E. Kelly
17. Some Aspects of Natural Language Processes by A.K. Joshi
18. Gibbs Sampling by S.F. Arnold
19. Bootstrap Methodology by G.J. Babu and C.R. Rao
20. The Art of Computer Generation of Random Variables by M.T. Boswell, S.D. Gore, G.P. Patil and C. Taillie
21. Jackknife Variance Estimation and Bias Reduction by S. Das Peddada
22. Designing Effective Statistical Graphs by D.A. Burn
23. Graphical Methods for Linear Models by A.S. Hadi
24. Graphics for Time Series Analysis by H.J. Newton
25. Graphics as Visual Language by T. Selkar and A. Appel

26. Statistical Graphics and Visualization by E.J. Wegman and D.B. Carr
27. Multivariate Statistical Visualization by F.W. Young, R.A. Faldowski and M.M. McFarlane
28. Graphical Methods for Process Control by T.L. Ziemer

Volume 10. Signal Processing and its Applications
Edited by N.K. Bose and C.R. Rao
1993 xvii + 992 pp.

1. Signal Processing for Linear Instrumental Systems with Noise: A General Theory with Illustrations from Optical Imaging and Light Scattering Problems by M. Bertero and E.R. Pike
2. Boundary Implication Results in Parameter Space by N.K. Bose
3. Sampling of Bandlimited Signals: Fundamental Results and Some Extensions by J.L. Brown Jr
4. Localization of Sources in a Sector: Algorithms and Statistical Analysis by K. Buckley and X.-L. Xu
5. The Signal Subspace Direction-of-Arrival Algorithm by J.A. Cadzow
6. Digital Differentiators by S.C. Dutta Roy and B. Kumar
7. Orthogonal Decompositions of 2D Random Fields and their Applications for 2D Spectral Estimation by J.M. Francos
8. VLSI in Signal Processing by A. Ghouse
9. Constrained Beamforming and Adaptive Algorithms by L.C. Godara
10. Bispectral Speckle Interferometry to Reconstruct Extended Objects from Turbulence-Degraded Telescope Images by D.M. Goodman, T.W. Lawrence, E. M. Johansson and J.P. Fitch
11. Multi-Dimensional Signal Processing by K. Hirano and T. Nomura
12. On the Assessment of Visual Communication by F.O. Huck, C.L. Fales, R. Alter-Gartenberg and Z. Rahman
13. VLSI Implementations of Number Theoretic Concepts with Applications in Signal Processing by G.A. Jullien, N.M. Wigley and J. Reilly
14. Decision-level Neural Net Sensor Fusion by R.Y. Levine and T.S. Khuon
15. Statistical Algorithms for Noncausal Gauss Markov Fields by J.M.F. Moura and N. Balram
16. Subspace Methods for Directions-of-Arrival Estimation by A. Paulraj, B. Ottersten, R. Roy, A. Swindlehurst, G. Xu and T. Kailath
17. Closed Form Solution to the Estimates of Directions of Arrival Using Data from an Array of Sensors by C.R. Rao and B. Zhou
18. High-Resolution Direction Finding by S.V. Schell and W.A. Gardner
19. Multiscale Signal Processing Techniques: A Review by A.H. Tewfik, M. Kim and M. Deriche
20. Sampling Theorems and Wavelets by G.G. Walter
21. Image and Video Coding Research by J.W. Woods
22. Fast Algorithms for Structured Matrices in Signal Processing by A.E. Yagle

Volume 11. Econometrics

Edited by G.S. Maddala, C.R. Rao and H.D. Vinod

1993 xx + 783 pp.

1. Estimation from Endogenously Stratified Samples by S.R. Cosslett
2. Semiparametric and Nonparametric Estimation of Quantal Response Models by J.L. Horowitz
3. The Selection Problem in Econometrics and Statistics by C.F. Manski
4. General Nonparametric Regression Estimation and Testing in Econometrics by A. Ullah and H.D. Vinod
5. Simultaneous Microeconometric Models with Censored or Qualitative Dependent Variables by R. Blundell and R.J. Smith
6. Multivariate Tobit Models in Econometrics by L.-F. Lee
7. Estimation of Limited Dependent Variable Models under Rational Expectations by G.S. Maddala
8. Nonlinear Time Series and Macroeconometrics by W.A. Brock and S.M. Potter
9. Estimation, Inference and Forecasting of Time Series Subject to Changes in Time by J.D. Hamilton
10. Structural Time Series Models by A.C. Harvey and N. Shephard
11. Bayesian Testing and Testing Bayesians by J.-P. Florens and M. Mouchart
12. Pseudo-Likelihood Methods by C. Gourieroux and A. Monfort
13. Rao's Score Test: Recent Asymptotic Results by R. Mukerjee
14. On the Strong Consistency of M-Estimates in Linear Models under a General Discrepancy Function by Z.D. Bai, Z.J. Liu and C.R. Rao
15. Some Aspects of Generalized Method of Moments Estimation by A. Hall
16. Efficient Estimation of Models with Conditional Moment Restrictions by W.K. Newey
17. Generalized Method of Moments: Econometric Applications by M. Ogaki
18. Testing for Heteroscedasticity by A.R. Pagan and Y. Pak
19. Simulation Estimation Methods for Limited Dependent Variable Models by V.A. Hajivassiliou
20. Simulation Estimation for Panel Data Models with Limited Dependent Variable by M.P. Keane
21. A Perspective Application of Bootstrap Methods in Econometrics by J. Jeong and G.S. Maddala
22. Stochastic Simulations for Inference in Nonlinear Errors-in-Variables Models by R.S. Mariano and B.W. Brown
23. Bootstrap Methods: Applications in Econometrics by H.D. Vinod
24. Identifying Outliers and Influential Observations in Econometric Models by S.G. Donald and G.S. Maddala
25. Statistical Aspects of Calibration in Macroeconomics by A.W. Gregory and G.W. Smith
26. Panel Data Models with Rational Expectations by K. Lahiri
27. Continuous Time Financial Models: Statistical Applications of Stochastic Processes by K.R. Sawyer

Volume 12. Environmental Statistics
Edited by G.P. Patil and C.R. Rao
1994 xix + 927 pp.

1. Environmetrics: An Emerging Science by J.S. Hunter
2. A National Center for Statistical Ecology and Environmental Statistics: A Center Without Walls by G.P. Patil
3. Replicate Measurements for Data Quality and Environmental Modeling by W. Liggett
4. Design and Analysis of Composite Sampling Procedures: A Review by G. Lovison, S.D. Gore and G.P. Patil
5. Ranked Set Sampling by G.P. Patil, A.K. Sinha and C. Taillie
6. Environmental Adaptive Sampling by G.A.F. Seber and S.K. Thompson
7. Statistical Analysis of Censored Environmental Data by M. Akritas, T. Ruscitti and G.P. Patil
8. Biological Monitoring: Statistical Issues and Models by E.P. Smith
9. Environmental Sampling and Monitoring by S.V. Stehman and W. Scott Overton
10. Ecological Statistics by B.F.J. Manly
11. Forest Biometrics by H.E. Burkhart and T.G. Gregoire
12. Ecological Diversity and Forest Management by J.H. Gove, G.P. Patil, B.F. Swindel and C. Taillie
13. Ornithological Statistics by P.M. North
14. Statistical Methods in Developmental Toxicology by P.J. Catalano and L.M. Ryan
15. Environmental Biometry: Assessing Impacts of Environmental Stimuli Via Animal and Microbial Laboratory Studies by W.W. Piegorsch
16. Stochasticity in Deterministic Models by J.J.M. Bedaux and S.A.L.M. Kooijman
17. Compartmental Models of Ecological and Environmental Systems by J.H. Matis and T.E. Wehrly
18. Environmental Remote Sensing and Geographic Information Systems-Based Modeling by W.L. Myers
19. Regression Analysis of Spatially Correlated Data: The Kanawha County Health Study by C.A. Donnelly, J.H. Ware and N.M. Laird
20. Methods for Estimating Heterogeneous Spatial Covariance Functions with Environmental Applications by P. Guttorp and P.D. Sampson
21. Meta-analysis in Environmental Statistics by V. Hasselblad
22. Statistical Methods in Atmospheric Science by A.R. Solow
23. Statistics with Agricultural Pests and Environmental Impacts by L.J. Young and J.H. Young
24. A Crystal Cube for Coastal and Estuarine Degradation: Selection of End-points and Development of Indices for Use in Decision Making by M.T. Boswell, J.S.O'Connor and G.P. Patil
25. How Does Scientific Information in General and Statistical Information in Particular Input to the Environmental Regulatory Process? by C.R. Cothern
26. Environmental Regulatory Statistics by C.B. Davis
27. An Overview of Statistical Issues Related to Environmental Cleanup by R. Gilbert
28. Environmental Risk Estimation and Policy Decisions by H. Lacayo Jr

Volume 13. Design and Analysis of Experiments
Edited by S. Ghosh and C.R. Rao
1996 xviii + 1230 pp.

1. The Design and Analysis of Clinical Trials by P. Armitage
2. Clinical Trials in Drug Development: Some Statistical Issues by H.I. Patel
3. Optimal Crossover Designs by J. Stufken
4. Design and Analysis of Experiments: Nonparametric Methods with Applications to Clinical Trials by P.K. Sen
5. Adaptive Designs for Parametric Models by S. Zacks
6. Observational Studies and Nonrandomized Experiments by P.R. Rosenbaum
7. Robust Design: Experiments for Improving Quality by D.M. Steinberg
8. Analysis of Location and Dispersion Effects from Factorial Experiments with a Circular Response by C.M. Anderson
9. Computer Experiments by J.R. Koehler and A.B. Owen
10. A Critique of Some Aspects of Experimental Design by J.N. Srivastava
11. Response Surface Designs by N.R. Draper and D.K.J. Lin
12. Multiresponse Surface Methodology by A.I. Khuri
13. Sequential Assembly of Fractions in Factorial Experiments by S. Ghosh
14. Designs for Nonlinear and Generalized Linear Models by A.C. Atkinson and L.M. Haines
15. Spatial Experimental Design by R.J. Martin
16. Design of Spatial Experiments: Model Fitting and Prediction by V.V. Fedorov
17. Design of Experiments with Selection and Ranking Goals by S.S. Gupta and S. Panchapakesan
18. Multiple Comparisons by A.C. Tamhane
19. Nonparametric Methods in Design and Analysis of Experiments by E. Brunner and M.L. Puri
20. Nonparametric Analysis of Experiments by A.M. Dean and D.A. Wolfe
21. Block and Other Designs in Agriculture by D.J. Street
22. Block Designs: Their Combinatorial and Statistical Properties by T. Calinski and S. Kageyama
23. Developments in Incomplete Block Designs for Parallel Line Bioassays by S. Gupta and R. Mukerjee
24. Row-Column Designs by K.R. Shah and B.K. Sinha
25. Nested Designs by J.P. Morgan
26. Optimal Design: Exact Theory by C.S. Cheng
27. Optimal and Efficient Treatment – Control Designs by D. Majumdar
28. Model Robust Designs by Y.-J. Chang and W.I. Notz
29. Review of Optimal Bayes Designs by A. DasGupta
30. Approximate Designs for Polynomial Regression: Invariance, Admissibility, and Optimality by N. Gaffke and B. Heiligers

Volume 14. Statistical Methods in Finance
Edited by G.S. Maddala and C.R. Rao
1996 xvi + 733 pp.

1. Econometric Evaluation of Asset Pricing Models by W.E. Person and R. Jegannathan
2. Instrumental Variables Estimation of Conditional Beta Pricing Models by C.R. Harvey and C.M. Kirby
3. Semiparametric Methods for Asset Pricing Models by B.N. Lehmann
4. Modeling the Term Structure by A.R. Pagan, A.D. Hall and V. Martin
5. Stochastic Volatility by E. Ghysels, A.C. Harvey and E. Renault
6. Stock Price Volatility by S.F. LeRoy
7. GARCH Models of Volatility by F.C. Palm
8. Forecast Evaluation and Combination by F.X. Diebold and J.A. Lopez
9. Predictable Components in Stock Returns by G. Kaul
10. Interset Rate Spreads as Predictors of Business Cycles by K. Lahiri and J.G. Wang
11. Nonlinear Time Series, Complexity Theory, and Finance by W.A. Brock and P.J.F. deLima
12. Count Data Models for Financial Data by A.C. Cameron and P.K. Trivedi
13. Financial Applications of Stable Distributions by J.H. McCulloch
14. Probability Distributions for Financial Models by J.B. McDonald
15. Bootstrap Based Tests in Financial Models by G.S. Maddala and H. Li
16. Principal Component and Factor Analyses by C.R. Rao
17. Errors in Variables Problems in Finance by G.S. Maddala and M. Nimalendran
18. Financial Applications of Artificial Neural Networks by M. Qi
19. Applications of Limited Dependent Variable Models in Finance by G.S. Maddala
20. Testing Option Pricing Models by D.S. Bates
21. Peso Problems: Their Theoretical and Empirical Implications by M.D.D. Evans
22. Modeling Market Microstructure Time Series by J. Hasbrouck
23. Statistical Methods in Tests of Portfolio Efficiency: A Synthesis by J. Shanken

Volume 15. Robust Inference
Edited by G.S. Maddala and C.R. Rao
1997 xviii + 698 pp.

1. Robust Inference in Multivariate Linear Regression Using Difference of Two Convex Functions as the Discrepancy Measure by Z.D. Bai, C.R. Rao and Y. H. Wu
2. Minimum Distance Estimation: The Approach Using Density-Based Distances by A. Basu, I.R. Harris and S. Basu
3. Robust Inference: The Approach Based on Influence Functions by M. Markatou and E. Ronchetti
4. Practical Applications of Bounded-Influence Tests by S. Heritier and M.-P. Victoria-Feser

5. Introduction to Positive-Breakdown Methods by P.J. Rousseeuw
6. Outlier Identification and Robust Methods by U. Gather and C. Becker
7. Rank-Based Analysis of Linear Models by T.P. Hettmansperger, J.W. McKean and S.J. Sheather
8. Rank Tests for Linear Models by R. Koenker
9. Some Extensions in the Robust Estimation of Parameters of Exponential and Double Exponential Distributions in the Presence of Multiple Outliers by A. Childs and N. Balakrishnan
10. Outliers, Unit Roots and Robust Estimation of Nonstationary Time Series by G.S. Maddala and Y. Yin
11. Autocorrelation-Robust Inference by P.M. Robinson and C. Velasco
12. A Practitioner's Guide to Robust Covariance Matrix Estimation by W.J. den Haan and A. Levin
13. Approaches to the Robust Estimation of Mixed Models by A.H. Welsh and A.M. Richardson
14. Nonparametric Maximum Likelihood Methods by S.R. Cosslett
15. A Guide to Censored Quantile Regressions by B. Fitzenberger
16. What Can Be Learned About Population Parameters When the Data Are Contaminated by J.L. Horowitz and C.F. Manski
17. Asymptotic Representations and Interrelations of Robust Estimators and Their Applications by J. Jurecková and P.K. Sen
18. Small Sample Asymptotics: Applications in Robustness by C.A. Field and M.A. Tingley
19. On the Fundamentals of Data Robustness by G. Maguluri and K. Singh
20. Statistical Analysis With Incomplete Data: A Selective Review by M.G. Akritas and M.P. La Valley
21. On Contamination Level and Sensitivity of Robust Tests by J.Á. Visšek
22. Finite Sample Robustness of Tests: An Overview by T. Kariya and P. Kim
23. Future Directions by G.S. Maddala and C.R. Rao

Volume 16. Order Statistics – Theory and Methods
Edited by N. Balakrishnan and C.R. Rao
1997 xix + 688 pp.

1. Order Statistics: An Introduction by N. Balakrishnan and C.R. Rao
2. Order Statistics: A Historical Perspective by H. Leon Harter and N. Balakrishnan
3. Computer Simulation of Order Statistics by Pandu R. Tadikamalla and N. Balakrishnan
4. Lorenz Ordering of Order Statistics and Record Values by Barry C. Arnold and Jose A. Villasenor
5. Stochastic Ordering of Order Statistics by Philip J. Boland, Moshe Shaked and J. George Shanthikumar
6. Bounds for Expectations of L-Estimates by T. Rychlik
7. Recurrence Relations and Identities for Moments of Order Statistics by N. Balakrishnan and K.S. Sultan

8. Recent Approaches to Characterizations Based on Order Statistics and Record Values by C.R. Rao and D.N. Shanbhag
9. Characterizations of Distributions via Identically Distributed Functions of Order Statistics by Ursula Gather, Udo Kamps and Nicole Schweitzer
10. Characterizations of Distributions by Recurrence Relations and Identities for Moments of Order Statistics by Udo Kamps
11. Univariate Extreme Value Theory and Applications by Janos Galambos
12. Order Statistics: Asymptotics in Applications by Pranab Kumar Sen
13. Zero-One Laws for Large Order Statistics by R.J. Tomkins and Hong Wang
14. Some Exact Properties of Cook's D_1 by D.R. Jensen and D.E. Ramirez
15. Generalized Recurrence Relations for Moments of Order Statistics from Non-Identical Pareto and Truncated Pareto Random Variables with Applications to Robustness by Aaron Childs and N. Balakrishnan
16. A Semiparametric Bootstrap for Simulating Extreme Order Statistics by Robert L. Strawderman and Daniel Zelterman
17. Approximations to Distributions of Sample Quantiles by Chunsheng Ma and John Robinson
18. Concomitants of Order Statistics by H.A. David and H.N. Nagaraja
19. A Record of Records by Valery B. Nevzorov and N. Balakrishnan
20. Weighted Sequential Empirical Type Processes with Applications to Change-Point Problems by Barbara Szyszkowicz
21. Sequential Quantile and Bahadur–Kiefer Processes by Miklós Csörgő and Barbara Szyszkowicz

Volume 17. Order Statistics: Applications
Edited by N. Balakrishnan and C.R. Rao
1998 xviii + 712 pp.

1. Order Statistics in Exponential Distribution by Asit P. Basu and Bahadur Singh
2. Higher Order Moments of Order Statistics from Exponential and Right-truncated Exponential Distributions and Applications to Life-testing Problems by N. Balakrishnan and Shanti S. Gupta
3. Log-gamma Order Statistics and Linear Estimation of Parameters by N. Balakrishnan and P.S. Chan
4. Recurrence Relations for Single and Product Moments of Order Statistics from a Generalized Logistic Distribution with Applications to Inference and Generalizations to Double Truncation by N. Balakrishnan and Rita Aggarwala
5. Order Statistics from the Type III Generalized Logistic Distribution and Applications by N. Balakrishnan and S.K. Lee
6. Estimation of Scale Parameter Based on a Fixed Set of Order Statistics by Sanat K. Sarkar and Wenjin Wang
7. Optimal Linear Inference Using Selected Order Statistics in Location-Scale Models by M. Masoom Ali and Dale Umbach
8. L-Estimation by J.R.M. Hosking
9. On Some L-estimation in Linear Regression Models by Soroush Alimoradi and A.K.Md. Ehsanes Saleh

10. The Role of Order Statistics in Estimating Threshold Parameters by A. Clifford Cohen
11. Parameter Estimation under Multiply Type-II Censoring by Fanhui Kong
12. On Some Aspects of Ranked Set Sampling in Parametric Estimation by Nora Ni Chuiv and Bimal K. Sinha
13. Some Uses of Order Statistics in Bayesian Analysis by Seymour Geisser
14. Inverse Sampling Procedures to Test for Homogeneity in a Multinomial Distribution by S. Panchapakesan, Aaron Childs, B.H. Humphrey and N. Balakrishnan
15. Prediction of Order Statistics by Kenneth S. Kaminsky and Paul I. Nelson
16. The Probability Plot: Tests of Fit Based on the Correlation Coefficient by R.A. Lockhart and M.A. Stephens
17. Distribution Assessment by Samuel Shapiro
18. Application of Order Statistics to Sampling Plans for Inspection by Variables by Helmut Schneider and Frances Barbera
19. Linear Combinations of Ordered Symmetric Observations with Applications to Visual Acuity by Marios Viana
20. Order-Statistic Filtering and Smoothing of Time-Series: Part I by Gonzalo R. Arce, Yeong-Taeg Kim and Kenneth E. Barner
21. Order-Statistic Filtering and Smoothing of Time-Series: Part II by Kenneth E. Barner and Gonzalo R. Arce
22. Order Statistics in Image Processing by Scott T. Acton and Alan C. Bovik
23. Order Statistics Application to CFAR Radar Target Detection by R. Viswanathan

Volume 18. Bioenvironmental and Public Health Statistics
Edited by P.K. Sen and C.R. Rao
2000 xxiv + 1105 pp.

1. Bioenvironment and Public Health: Statistical Perspectives by Pranab K. Sen
2. Some Examples of Random Process Environmental Data Analysis by David R. Brillinger
3. Modeling Infectious Diseases – Aids by L. Billard
4. On Some Multiplicity Problems and Multiple Comparison Procedures in Biostatistics by Yosef Hochberg and Peter H. Westfall
5. Analysis of Longitudinal Data by Julio M. Singer and Dalton F. Andrade
6. Regression Models for Survival Data by Richard A. Johnson and John P. Klein
7. Generalised Linear Models for Independent and Dependent Responses by Bahjat F. Qaqish and John S. Preisser
8. Hierarchial and Empirical Bayes Methods for Environmental Risk Assessment by Gauri Datta, Malay Ghosh and Lance A. Waller
9. Non-parametrics in Bioenvironmental and Public Health Statistics by Pranab Kumar Sen
10. Estimation and Comparison of Growth and Dose-Response Curves in the Presence of Purposeful Censoring by Paul W. Stewart
11. Spatial Statistical Methods for Environmental Epidemiology by Andrew B. Lawson and Noel Cressie

12. Evaluating Diagnostic Tests in Public Health by Margaret Pepe, Wendy Leisenring and Carolyn Rutter
13. Statistical Issues in Inhalation Toxicology by E. Weller, L. Ryan and D. Dockery
14. Quantitative Potency Estimation to Measure Risk with Bioenvironmental Hazards by A. John Bailer and Walter W. Piegorsch
15. The Analysis of Case-Control Data: Epidemiologic Studies of Familial Aggregation by Nan M. Laird, Garrett M. Fitzmaurice and Ann G. Schwartz
16. Cochran–Mantel–Haenszel Techniques: Applications Involving Epidemiologic Survey Data by Daniel B. Hall, Robert F. Woolson, William R. Clarke and Martha F. Jones
17. Measurement Error Models for Environmental and Occupational Health Applications by Robert H. Lyles and Lawrence L. Kupper
18. Statistical Perspectives in Clinical Epidemiology by Shrikant I. Bangdiwala and Sergio R. Muñoz
19. ANOVA and ANOCOVA for Two-Period Crossover Trial Data: New vs. Standard by Subir Ghosh and Lisa D. Fairchild
20. Statistical Methods for Crossover Designs in Bioenvironmental and Public Health Studies by Gail E. Tudor, Gary G. Koch and Diane Catellier
21. Statistical Models for Human Reproduction by C.M. Suchindran and Helen P. Koo
22. Statistical Methods for Reproductive Risk Assessment by Sati Mazumdar, Yikang Xu, Donald R. Mattison, Nancy B. Sussman and Vincent C. Arena
23. Selection Biases of Samples and their Resolutions by Ranajit Chakraborty and C. Radhakrishna Rao
24. Genomic Sequences and Quasi-Multivariate CATANOVA by Hildete Prisco Pinheiro, Françoise Seillier-Moiseiwitsch, Pranab Kumar Sen and Joseph Eron Jr
25. Statistical Methods for Multivariate Failure Time Data and Competing Risks by Ralph A. DeMasi
26. Bounds on Joint Survival Probabilities with Positively Dependent Competing Risks by Sanat K. Sarkar and Kalyan Ghosh
27. Modeling Multivariate Failure Time Data by Limin X. Clegg, Jianwen Cai and Pranab K. Sen
28. The Cost–Effectiveness Ratio in the Analysis of Health Care Programs by Joseph C. Gardiner, Cathy J. Bradley and Marianne Huebner
29. Quality-of-Life: Statistical Validation and Analysis An Example from a Clinical Trial by Balakrishna Hosmane, Clement Maurath and Richard Manski
30. Carcinogenic Potency: Statistical Perspectives by Anup Dewanji
31. Statistical Applications in Cardiovascular Disease by Elizabeth R. DeLong and David M. DeLong
32. Medical Informatics and Health Care Systems: Biostatistical and Epidemiologic Perspectives by J. Zvárová
33. Methods of Establishing In Vitro–In Vivo Relationships for Modified Release Drug Products by David T. Mauger and Vernon M. Chinchilli
34. Statistics in Psychiatric Research by Sati Mazumdar, Patricia R. Houck and Charles F. Reynolds III
35. Bridging the Biostatistics–Epidemiology Gap by Lloyd J. Edwards
36. Biodiversity – Measurement and Analysis by S.P. Mukherjee

Volume 19. Stochastic Processes: Theory and Methods
Edited by D.N. Shanbhag and C.R. Rao
2001 xiv + 967 pp.

1. Pareto Processes by Barry C. Arnold
2. Branching Processes by K.B. Athreya and A.N. Vidyashankar
3. Inference in Stochastic Processes by I.V. Basawa
4. Topics in Poisson Approximation by A.D. Barbour
5. Some Elements on Lévy Processes by Jean Bertoin
6. Iterated Random Maps and Some Classes of Markov Processes by Rabi Bhattacharya and Edward C. Waymire
7. Random Walk and Fluctuation Theory by N.H. Bingham
8. A Semigroup Representation and Asymptotic Behavior of Certain Statistics of the Fisher–Wright–Moran Coalescent by Adam Bobrowski, Marek Kimmel, Ovide Arino and Ranajit Chakraborty
9. Continuous-Time ARMA Processes by P.J. Brockwell
10. Record Sequences and their Applications by John Bunge and Charles M. Goldie
11. Stochastic Networks with Product Form Equilibrium by Hans Daduna
12. Stochastic Processes in Insurance and Finance by Paul Embrechts, Rüdiger Frey and Hansjörg Furrer
13. Renewal Theory by D.R. Grey
14. The Kolmogorov Isomorphism Theorem and Extensions to some Nonstationary Processes by Yûichirô Kakihara
15. Stochastic Processes in Reliability by Masaaki Kijima, Haijun Li and Moshe Shaked
16. On the supports of Stochastic Processes of Multiplicity One by A. Kłopotowski and M.G. Nadkarni
17. Gaussian Processes: Inequalities, Small Ball Probabilities and Applications by W.V. Li and Q.-M. Shao
18. Point Processes and Some Related Processes by Robin K. Milne
19. Characterization and Identifiability for Stochastic Processes by B.L.S. Prakasa Rao
20. Associated Sequences and Related Inference Problems by B.L.S. Prakasa Rao and Isha Dewan
21. Exchangeability, Functional Equations, and Characterizations by C.R. Rao and D.N. Shanbhag
22. Martingales and Some Applications by M.M. Rao
23. Markov Chains: Structure and Applications by R.L. Tweedie
24. Diffusion Processes by S.R.S. Varadhan
25. Itô's Stochastic Calculus and Its Applications by S. Watanabe

Volume 20. Advances in Reliability
Edited by N. Balakrishnan and C.R. Rao
2001 xxii + 860 pp.

1. Basic Probabilistic Models in Reliability by N. Balakrishnan, N. Limnios and C. Papadopoulos

2. The Weibull Nonhomogeneous Poisson Process by A.P Basu and S.E. Rigdon
3. Bathtub-Shaped Failure Rate Life Distributions by C.D. Lai, M. Xie and D.N.P. Murthy
4. Equilibrium Distribution – its Role in Reliability Theory by A. Chatterjee and S.P. Mukherjee
5. Reliability and Hazard Based on Finite Mixture Models by E.K. Al-Hussaini and K.S. Sultan
6. Mixtures and Monotonicity of Failure Rate Functions by M. Shaked and F. Spizzichino
7. Hazard Measure and Mean Residual Life Orderings: A Unified Approach by M. Asadi and D.N. Shanbhag
8. Some Comparison Results of the Reliability Functions of Some Coherent Systems by J. Mi
9. On the Reliability of Hierarchical Structures by L.B. Klebanov and G.J. Szekely
10. Consecutive k-out-of-n Systems by N.A. Mokhlis
11. Exact Reliability and Lifetime of Consecutive Systems by S. Aki
12. Sequential k-out-of-n Systems by E. Cramer and U. Kamps
13. Progressive Censoring: A Review by R. Aggarwala
14. Point and Interval Estimation for Parameters of the Logistic Distribution Based on Progressively Type-II Censored Samples by N. Balakrishnan and N. Kannan
15. Progressively Censored Variables-Sampling Plans for Life Testing by U. Balasooriya
16. Graphical Techniques for Analysis of Data From Repairable Systems by P.A. Akersten, B. Klefsjö and B. Bergman
17. A Bayes Approach to the Problem of Making Repairs by G.C. McDonald
18. Statistical Analysis for Masked Data by B.J. Flehinger†, B. Reiser and E. Yashchin
19. Analysis of Masked Failure Data under Competing Risks by A. Sen, S. Basu and M. Banerjee
20. Warranty and Reliability by D.N.P. Murthy and W.R. Blischke
21. Statistical Analysis of Reliability Warranty Data by K. Suzuki, Md. Rezaul Karim and L. Wang
22. Prediction of Field Reliability of Units, Each under Differing Dynamic Stresses, from Accelerated Test Data by W. Nelson
23. Step-Stress Accelerated Life Test by E. Gouno and N. Balakrishnan
24. Estimation of Correlation under Destructive Testing by R. Johnson and W. Lu
25. System-Based Component Test Plans for Reliability Demonstration: A Review and Survey of the State-of-the-Art by J. Rajgopal and M. Mazumdar
26. Life-Test Planning for Preliminary Screening of Materials: A Case Study by J. Stein and N. Doganaksoy
27. Analysis of Reliability Data from In-House Audit Laboratory Testing by R. Agrawal and N. Doganaksoy
28. Software Reliability Modeling, Estimation and Analysis by M. Xie and G.Y. Hong
29. Bayesian Analysis for Software Reliability Data by J.A. Achcar
30. Direct Graphical Estimation for the Parameters in a Three-Parameter Weibull Distribution by P.R. Nelson and K.B. Kulasekera

31. Bayesian and Frequentist Methods in Change-Point Problems by N. Ebrahimi and S.K. Ghosh
32. The Operating Characteristics of Sequential Procedures in Reliability by S. Zacks
33. Simultaneous Selection of Extreme Populations from a Set of Two-Parameter Exponential Populations by K. Hussein and S. Panchapakesan

Volume 21. Stochastic Processes: Modelling and Simulation
Edited by D.N. Shanbhag and C.R. Rao
2003 xxviii + 1002 pp.

1. Modelling and Numerical Methods in Manufacturing System Using Control Theory by E.K. Boukas and Z.K. Liu
2. Models of Random Graphs and their Applications by C. Cannings and D.B. Penman
3. Locally Self-Similar Processes and their Wavelet Analysis by J.E. Cavanaugh, Y. Wang and J.W. Davis
4. Stochastic Models for DNA Replication by R. Cowan
5. An Empirical Process with Applications to Testing the Exponential and Geometric Models by J.A. Ferreira
6. Patterns in Sequences of Random Events by J. Gani
7. Stochastic Models in Telecommunications for Optimal Design, Control and Performance Evaluation by N. Gautam
8. Stochastic Processes in Epidemic Modelling and Simulation by D. Greenhalgh
9. Empirical Estimators Based on MCMC Data by P.E. Greenwood and W. Wefelmeyer
10. Fractals and the Modelling of Self-Similarity by B.M. Hambly
11. Numerical Methods in Queueing Theory by D. Heyman
12. Applications of Markov Chains to the Distribution Theory of Runs and Patterns by M.V. Koutras
13. Modelling Image Analysis Problems Using Markov Random Fields by S.Z. Li
14. An Introduction to Semi-Markov Processes with Application to Reliability by N. Limnios and G. Oprişan
15. Departures and Related Characteristics in Queueing Models by M. Manoharan, M.H. Alamatsaz and D.N. Shanbhag
16. Discrete Variate Time Series by E. McKenzie
17. Extreme Value Theory, Models and Simulation by S. Nadarajah
18. Biological Applications of Branching Processes by A.G. Pakes
19. Markov Chain Approaches to Damage Models by C.R. Rao, M. Albassam, M.B. Rao and D.N. Shanbhag
20. Point Processes in Astronomy: Exciting Events in the Universe by J.D. Scargle and G.J. Babu
21. On the Theory of Discrete and Continuous Bilinear Time Series Models by T. Subba Rao and Gy. Terdik
22. Nonlinear and Non-Gaussian State-Space Modeling with Monte Carlo Techniques: A Survey and Comparative Study by H. Tanizaki
23. Markov Modelling of Burst Behaviour in Ion Channels by G.F. Yeo, R.K. Milne, B.W. Madsen, Y. Li and R.O. Edeson

Volume 22. Statistics in Industry
Edited by R. Khattree and C.R. Rao
2003 xxi + 1150 pp.

1. Guidelines for Selecting Factors and Factor Levels for an Industrial Designed Experiment by V. Czitrom
2. Industrial Experimentation for Screening by D.K.J. Lin
3. The Planning and Analysis of Industrial Selection and Screening Experiments by G. Pan, T.J. Santner and D.M. Goldsman
4. Uniform Experimental Designs and their Applications in Industry by K.-T. Fang and D.K.J. Lin
5. Mixed Models and Repeated Measures: Some Illustrative Industrial Examples by G.A. Milliken
6. Current Modeling and Design Issues in Response Surface Methodology: GLMs and Models with Block Effects by A.I. Khuri
7. A Review of Design and Modeling in Computer Experiments by V.C.P. Chen, K.-L. Tsui, R.R. Barton and J.K. Allen
8. Quality Improvement and Robustness via Design of Experiments by B.E. Ankenman and A.M. Dean
9. Software to Support Manufacturing Experiments by J.E. Reece
10. Statistics in the Semiconductor Industry by V. Czitrom
11. PREDICT: A New Approach to Product Development and Lifetime Assessment Using Information Integration Technology by J.M. Booker, T.R. Bement, M.A. Meyerand W.J. Kerscher III
12. The Promise and Challenge of Mining Web Transaction Data by S.R. Dalal, D. Egan, Y. Ho and M. Rosenstein
13. Control Chart Schemes for Monitoring the Mean and Variance of Processes Subject to Sustained Shifts and Drifts by Z.G. Stoumbos, M.R. Reynolds Jr and W.H. Woodall
14. Multivariate Control Charts: Hotelling T^2, Data Depth and Beyond by R.Y. Liu
15. Effective Sample Sizes for T^2 Control Charts by R.L. Mason, Y.-M. Chou and J.C. Young
16. Multidimensional Scaling in Process Control by T.F. Cox
17. Quantifying the Capability of Industrial Processes by A.M. Polansky and S.N.U.A. Kirmani
18. Taguchi's Approach to On-line Control Procedure by M.S. Srivastava and Y. Wu
19. Dead-Band Adjustment Schemes for On-line Feedback Quality Control by A. Luceño
20. Statistical Calibration and Measurements by H. Iyer
21. Subsampling Designs in Industry: Statistical Inference for Variance Components by R. Khattree
22. Repeatability, Reproducibility and Interlaboratory Studies by R. Khattree
23. Tolerancing – Approaches and Related Issues in Industry by T.S. Arthanari
24. Goodness-of-fit Tests for Univariate and Multivariate Normal Models by D.K. Srivastava and G.S. Mudholkar

25. Normal Theory Methods and their Simple Robust Analogs for Univariate and Multivariate Linear Models by D.K. Srivastava and G.S. Mudholkar
26. Diagnostic Methods for Univariate and Multivariate Normal Data by D.N. Naik
27. Dimension Reduction Methods Used in Industry by G. Merola and B. Abraham
28. Growth and Wear Curves by A.M. Kshirsagar
29. Time Series in Industry and Business by B. Abraham and N. Balakrishna
30. Stochastic Process Models for Reliability in Dynamic Environments by N.D. Singpurwalla, T.A. Mazzuchi, S. Özekici and R. Soyer
31. Bayesian Inference for the Number of Undetected Errors by S. Basu

Volume 23. Advances in Survival Analysis
Edited by N. Balakrishnan and C.R. Rao
2003 xxv + 795 pp.

1. Evaluation of the Performance of Survival Analysis Models: Discrimination and Calibration Measures by R.B. D'Agostino and B.-H. Nam
2. Discretizing a Continuous Covariate in Survival Studies by J.P. Klein and J.-T. Wu
3. On Comparison of Two Classification Methods with Survival Endpoints by Y. Lu, H. Jin and J. Mi
4. Time-Varying Effects in Survival Analysis by T.H. Scheike
5. Kaplan–Meier Integrals by W. Stute
6. Statistical Analysis of Doubly Interval-Censored Failure Time Data by J. Sun
7. The Missing Censoring-Indicator Model of Random Censorship by S. Subramanian
8. Estimation of the Bivariate Survival Function with Generalized Bivariate Right Censored Data Structures by S. Keleş, M.J. van der Laan and J.M. Robins
9. Estimation of Semi-Markov Models with Right-Censored Data by O. Pons
10. Nonparametric Bivariate Estimation with Randomly Truncated Observations by Ü. Gürler
11. Lower Bounds for Estimating a Hazard by C. Huber and B. MacGibbon
12. Non-Parametric Hazard Rate Estimation under Progressive Type-II Censoring by N. Balakrishnan and L. Bordes
13. Statistical Tests of the Equality of Survival Curves: Reconsidering the Options by G.P. Suciu, S. Lemeshow and M. Moeschberger
14. Testing Equality of Survival Functions with Bivariate Censored Data: A Review by P.V. Rao
15. Statistical Methods for the Comparison of Crossing Survival Curves by C.T. Le
16. Inference for Competing Risks by J.P. Klein and R. Bajorunaite
17. Analysis of Cause-Specific Events in Competing Risks Survival Data by J. Dignam, J. Bryant and H.S. Wieand
18. Analysis of Progressively Censored Competing Risks Data by D. Kundu, N. Kannan and N. Balakrishnan
19. Marginal Analysis of Point Processes with Competing Risks by R.J. Cook, B. Chen and P. Major
20. Categorical Auxiliary Data in the Discrete Time Proportional Hazards Model by P. Slasor and N. Laird

21. Hosmer and Lemeshow type Goodness-of-Fit Statistics for the Cox Proportional Hazards Model by S. May and D.W. Hosmer
22. The Effects of Misspecifying Cox's Regression Model on Randomized Treatment Group Comparisons by A.G. DiRienzo and S.W. Lagakos
23. Statistical Modeling in Survival Analysis and Its Influence on the Duration Analysis by V. Bagdonavičius and M. Nikulin
24. Accelerated Hazards Model: Method, Theory and Applications by Y.Q. Chen, N.P. Jewell and J. Yang
25. Diagnostics for the Accelerated Life Time Model of Survival Data by D. Zelterman and H. Lin
26. Cumulative Damage Approaches Leading to Inverse Gaussian Accelerated Test Models by A. Onar and W.J. Padgett
27. On Estimating the Gamma Accelerated Failure-Time Models by K.M. Koti
28. Frailty Model and its Application to Seizure Data by N. Ebrahimi, X. Zhang, A. Berg and S. Shinnar
29. State Space Models for Survival Analysis by W.Y. Tan and W. Ke
30. First Hitting Time Models for Lifetime Data by M.-L.T. Lee and G.A. Whitmore
31. An Increasing Hazard Cure Model by Y. Peng and K.B.G. Dear
32. Marginal Analyses of Multistage Data by G.A. Satten and S. Datta
33. The Matrix-Valued Counting Process Model with Proportional Hazards for Sequential Survival Data by K.L. Kesler and P.K. Sen
34. Analysis of Recurrent Event Data by J. Cai and D.E. Schaubel
35. Current Status Data: Review, Recent Developments and Open Problems by N.P. Jewell and M. van der Laan
36. Appraisal of Models for the Study of Disease Progression in Psoriatic Arthritis by R. Aguirre-Hernández and V.T. Farewell
37. Survival Analysis with Gene Expression Arrays by D.K. Pauler, J. Hardin, J.R. Faulkner, M. LeBlanc and J.J. Crowley
38. Joint Analysis of Longitudinal Quality of Life and Survival Processes by M. Mesbah, J.-F. Dupuy, N. Heutte and L. Awad
39. Modelling Survival Data using Flowgraph Models by A.V. Huzurbazar
40. Nonparametric Methods for Repair Models by M. Hollander and J. Set- huraman

Volume 24. Data Mining and Data Visualization
Edited by C.R. Rao, E.J. Wegman and J.L. Solka
2005 xiv + 643 pp.

1. Statistical Data Mining by E.J. Wegman and J.L. Solka
2. From Data Mining to Knowledge Mining by K.A. Kaufman and R.S. Michalski
3. Mining Computer Security Data by D.J. Marchette
4. Data Mining of Text Files by A.R. Martinez
5. Text Data Mining with Minimal Spanning Trees by J.L. Solka, A.C. Bryant and E.J. Wegman
6. Information Hiding: Steganography and Steganalysis by Z. Duric, M. Jacobs and S. Jajodia

7. Canonical Variate Analysis and Related Methods for Reduction of Dimensionality and Graphical Representation by C.R. Rao
8. Pattern Recognition by D.J. Hand
9. Multidimensional Density Estimation by D.W. Scott and S.R. Sain
10. Multivariate Outlier Detection and Robustness by M. Hubert, P.J. Rousseeuw and S. Van Aelst
11. Classification and Regression Trees, Bagging, and Boosting by C.D. Sutton
12. Fast Algorithms for Classification Using Class Cover Catch Digraphs by D.J. Marchette, E.J. Wegman and C.E. Priebe
13. On Genetic Algorithms and their Applications by Y.H. Said
14. Computational Methods for High-Dimensional Rotations in Data Visualization by A. Buja, D. Cook, D. Asimov and C. Hurley
15. Some Recent Graphics Templates and Software for Showing Statistical Summaries by D.B. Carr
16. Interactive Statistical Graphics: the Paradigm of Linked Views by A. Wilhelm
17. Data Visualization and Virtual Reality by J.X. Chen

Volume 25. Bayesian Thinking: Modeling and Computation
Edited by D.K. Dey and C.R. Rao
2005 xx + 1041 pp.

1. Bayesian Inference for Causal Effects by D.B. Rubin
2. Reference Analysis by J.M. Bernardo
3. Probability Matching Priors by G.S. Datta and T.J. Sweeting
4. Model Selection and Hypothesis Testing based on Objective Probabilities and Bayes Factors by L.R. Pericchi
5. Role of P-values and other Measures of Evidence in Bayesian Analysis by J. Ghosh, S. Purkayastha and T. Samanta
6. Bayesian Model Checking and Model Diagnostics by H.S. Stern and S. Sinharay
7. The Elimination of Nuisance Parameters by B. Liseo
8. Bayesian Estimation of Multivariate Location Parameters by A.C. Brandwein and W.E. Strawderman
9. Bayesian Nonparametric Modeling and Data Analysis: An Introduction by T.E. Hanson, A.J. Branscum and W.O. Johnson
10. Some Bayesian Nonparametric Models by P. Damien
11. Bayesian Modeling in the Wavelet Domain by F. Ruggeri and B. Vidakovic
12. Bayesian Nonparametric Inference by S. Walker
13. Bayesian Methods for Function Estimation by N. Choudhuri, S. Ghosal and A. Roy
14. MCMC Methods to Estimate Bayesian Parametric Models by A. Mira
15. Bayesian Computation: From Posterior Densities to Bayes Factors, Marginal Likelihoods, and Posterior Model Probabilities by M.-H. Chen
16. Bayesian Modelling and Inference on Mixtures of Distributions by J.-M. Marin, K. Mengersen and C.P Robert
17. Simulation Based Optimal Design by P. Müller

18. Variable Selection and Covariance Selection in Multivariate Regression Models by E. Cripps, C. Carter and R. Kohn
19. Dynamic Models by H.S. Migon, D. Gamerman, H.F. Lopes and M.A.R. Ferreira
20. Bayesian Thinking in Spatial Statistics by L.A. Waller
21. Robust Bayesian Analysis by F. Ruggeri, D. Ríos Insua and Jacinto Martin
22. Elliptical Measurement Error Models – A Bayesian Approach by H. Bolfarine and R.B. Arellano-Valle
23. Bayesian Sensitivity Analysis in Skew-elliptical Models by I. Vidal, P. Iglesias and M.D. Branco
24. Bayesian Methods for DNA Microarray Data Analysis by V. Baladandayuthapani, S. Ray and B.K. Mallick
25. Bayesian Biostatistics by D.B. Dunson
26. Innovative Bayesian Methods for Biostatistics and Epidemiology by P. Gustafson, S. Hossain and L. McCandless
27. Bayesian Analysis of Case-Control Studies by B. Mukherjee, S. Sinha and M. Ghosh
28. Bayesian Analysis of ROC Data by V.E. Johnson and T.D. Johnson
29. Modeling and Analysis for Categorical Response Data by S. Chib
30. Bayesian Methods and Simulation-Based Computation for Contingency Tables by J.H. Albert
31. Multiple Events Time Data: A Bayesian Recourse by D. Sinha and S.K. Ghosh
32. Bayesian Survival Analysis for Discrete Data with Left-Truncation and Interval Censoring by C.Z. He and D. Sun
33. Software Reliability by L. Kuo
34. Bayesian Aspects of Small Area Estimation by T. Maiti
35. Teaching Bayesian Thought to Nonstatisticians by D.K. Stangl

Volume 26. Psychometrics
Edited by C.R. Rao and S. Sinharay
2007 xx + 1169 pp.

1. A History and Overview of Psychometrics by Lyle V. Jones and David Thissen
2. Selected Topics in Classical Test Theory by Charles Lewis
3. Validity: Foundational Issues and Statistical Methodology by Bruno D. Zumbo
4. Reliability Coefficients and Generalizability Theory by Noreen M. Webb, Richard J. Shavelson and Edward H. Haertel
5. Differential Item Functioning and Item Bias by Randall D. Penfield and Gregory Camilli
6. Equating Test Scores by Paul W. Holland, Neil J. Dorans and Nancy S. Petersen
7. Electronic Essay Grading by Shelby J. Haberman
8. Some Matrix Results Useful in Psychometric Research by C. Radhakrishna Rao
9. Factor Analysis by Haruo Yanai and Masanori Ichikawa
10. Structural Equation Modeling by Ke-Hai Yuan and Peter M. Bentler
11. Applications of Multidimensional Scaling in Psychometrics by Yoshio Takane
12. Multilevel Models in Psychometrics by Fiona Steele and Harvey Goldstein

13. Latent Class Analysis in Psychometrics by C. Mitchell Dayton and George B. Macready
14. Random-Effects Models for Preference Data by Ulf Böckenholt and Rung-Ching Tsai
15. Item Response Theory in a General Framework by R. Darrell Bock and Irini Moustaki
16. Rasch Models by Gerhard H. Fischer
17. Hierarchical Item Response Theory Models by Matthew S. Johnson, Sandip Sinharay and Eric T. Bradlow
18. Multidimensional Item Response Theory by Mark D. Reckase
19. Mixture Distribution Item Response Models by Matthias von Davier and Jürgen Rost
20. Scoring Open Ended Questions by Gunter Maris and Timo Bechger
21. Assessing the Fit of Item Response Theory Models by Hariharan Swaminathan, Ronald K. Hambleton and H. Jane Rogers
22. Nonparametric Item Response Theory and Special Topics by Klaas Sijtsma and Rob R. Meijer
23. Automatic Item Generation and Cognitive Psychology by Susan Embretson and Xiangdong Yang
24. Statistical Inference for Causal Effects, with Emphasis on Applications in Psychometrics and Education by Donald B.Rubin
25. Statistical Aspects of Adaptive Testing by Wim J. van der Linden and Cees A.W. Glas
26. Bayesian Psychometric Modeling From An Evidence-Centered Design Perspective by Robert J. Mislevy and Roy Levy
27. Value-Added Modeling by Henry Braun and Howard Wainer
28. Three Statistical Paradoxes in the Interpretation of Group Differences: Il-lustrated with Medical School Admission and Licensing Data by Howard Wainer and Lisa M. Brown
29. Meta-Analysis by Larry V. Hedges
30. Vertical Scaling: Statistical Models for Measuring Growth and Achievement by Richard J. Patz and Lihua Yao
31. COGNITIVE DIAGNOSIS
    a. Review of Cognitively Diagnostic Assessment and a Summary of Psychometric Models by Louis V. DiBello, Louis A. Roussos and William Stout
    b. Some Notes on Models for Cognitively Based Skills Diagnosis by Shelby J. Haberman and Matthias von Davier
32. The Statistical Procedures Used in National Assessment of Educational Progress: Recent Developments and Future Directions by Matthias von Davier, Sandip Sinharay, Andreas Oranje and Albert Beaton
33. Statistical Procedures Used in College Admissions Testing by Jinghua Liu, Deborah J. Harris and Amy Schmidt
34. FUTURE CHALLENGES IN PSYCHOMETRICS
    a. Integration of Models by Robert L. Brennan
    b. Linking Scores Across Computer and Paper-Based Modes of Test Administration by Daniel R. Eignor

c. Linking Cognitively-Based Models and Psychometric Methods by Mark J. Gierl and Jacqueline P. Leighton
d. Technical Considerations in Equating Complex Assessments by Ida Lawrence
e. Future Challenges to Psychometrics: Validity, Validity, Validity by Neal Kingston
f. Testing with and without Computers by Piet Sanders
G. Practical Challenges to Psychometrics Driven by Increased Visibility of Assessment by Cynthia Board Schmeiser

Volume 27. Epidemiology and Medical Statistics
Edited by C.R. Rao, J.P. Miller, and D.C.Rao
2009 xviii + 812 pp.

1. Statistical Methods and Challenges in Epidemiology and Biomedical Research by Ross L. Prentice
2. Statistical Inference for Causal Effects, With Emphasis on Applications in Epidemiology and Medical Statistics by Donald B. Rubin
3. Epidemiologic Study Designs by Kenneth J. Rothman, Sander Greenland and Timothy L. Lash
4. Statistical Methods for Assessing Biomarkers and Analyzing Biomarker Data by Stephen W. Looney and Joseph L. Hagan
5. Linear and Non-Linear Regression Methods in Epidemiology and Biostatistics by Eric Vittinghoff, Charles E. McCulloch, David V. Glidden and Stephen C. Shiboski
6. Logistic Regression by Edward L. Spitznagel Jr.
7. Count Response Regression Models by Joseph M. Hilbe and William H. Greene
8. Mixed Models by Matthew J. Gurka and Lloyd J. Edwards
9. Survival Analysis by John P. Klein and Mei-Jie Zhang
10. A Review of Statistical Analyses for Competing Risks by Melvin L. Moeschberger, Kevin P. Tordoff and Nidhi Kochar
11. Cluster Analysis by William D. Shannon
12. Factor Analysis and Related Methods by Carol M. Woods and Michael C. Edwards
13. Structural Equation Modeling by Kentaro Hayashi, Peter M. Bentler and Ke-Hai Yuan
14. Statistical Modeling in Biomedical Research: Longitudinal Data Analysis by Chengjie Xiong, Kejun Zhu, Kai Yu and J. Philip Miller
15. Design and Analysis of Cross-Over Trials by Michael G. Kenward and Byron Jones
16. Sequential and Group Sequential Designs in Clinical Trials: Guidelines for Practitioners by Madhu Mazumdar and Heejung Bang
17. Early Phase Clinical Trials: Phases I and II by Feng Gao, Kathryn Trinkaus and J. Philip Miller
18. Definitive Phase III and Phase IV Clinical Trials by Barry R. Davis and Sarah Baraniuk
19. Incomplete Data in Epidemiology and Medical Statistics by Susanne Rässler, Donald B. Rubin and Elizabeth R. Zell
20. Meta-Analysis by Edward L. Spitznagel Jr.
21. The Multiple Comparison Issue in Health Care Research by Lemuel A. Moyé

22. Power: Establishing the Optimum Sample Size by Richard A. Zeller and Yan Yan
23. Statistical Learning in Medical Data Analysis by Grace Wahba
24. Evidence Based Medicine and Medical Decision Making by Dan Mayer, MD
25. Estimation of Marginal Regression Models with Multiple Source Predictors by Heather J. Litman, Nicholas J. Horton, Bernardo Hernández and Nan M. Laird
26. Difference Equations with Public Health Applications by Asha Seth Kapadia and Lemuel A. Moyé
27. The Bayesian Approach to Experimental Data Analysis by Bruno Lecoutre