# handbook of statistics 29A

## Sample Surveys: Design, Methods and Applications

Edited by
D. Pfeffermann
C.R. Rao

N·H

# 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

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

# Preface to Handbook 29A

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.

The present 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.

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. 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).

# 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).

# Introduction to Part 1

Sharon L. Lohr

## 1. Importance of survey design

Sample surveys have many possible objectives: to estimate changes in unemployment rates over time, to study through election polls how the public views political candidates, or to estimate the number of gila monsters in Arizona. In all surveys, however, the major goal is to estimate characteristics of a static or dynamic population using data from a sample. Mahalanobis (1965, p. 45) summarized the advantages of sample surveys: "…large scale sample surveys, when conducted in the proper way with a satisfactory survey design, can supply with great speed and at low cost information of sufficient accuracy for practical purposes and with the possibility of ascertainment of the margin of uncertainty on an objective basis." The key to attaining these advantages is the "satisfactory survey design."

Part 1 of this Handbook focuses on issues in survey design. For the purposes of this book, survey design means the procedure used to select units from the population for inclusion in the sample. Designing a survey is the most important stage of a survey since design deficiencies cannot always be compensated for when editing and analyzing the data. A sample that consists entirely of volunteers, such as a web-based poll that instructs visitors to "click here" if they wish to express opinions about a political candidate or issue, is usually useless for the purpose of estimating how many persons in a population of interest share those opinions.

The classical building blocks of survey design for probability samples, including simple random sampling, stratification, and multistage cluster sampling, were all developed with the goal of minimizing the survey cost while controlling the uncertainty associated with key estimates. Much of the research on these designs was motivated by methods used to collect survey data in the 1930s and 1940s. Data for many surveys were collected in person, which necessitated cluster sampling to reduce travel costs. At the same time, auxiliary information that could be used to improve design efficiency was sometimes limited, which reduced potential gains from stratification. Mahalanobis (1946) also emphasized the need for designs and estimators with straightforward computations so that additional errors would not be introduced by the people who served as computers.

Stratification and multistage sampling are still key design features for surveys. New methods of data collection and more available information for population units, however, can and should be factored into design choices. In addition, new uses of survey data lead to new demands for survey designs. While straightforward computations are less essential now than in 1946, conceptual simplicity of designs and estimators is still valuable for accuracy as well as public acceptance of survey estimates.

Section 2 of this introduction reviews the underlying framework of survey design and outlines how inferential approaches influence design choice. Section 3 then presents contemporary design challenges that are discussed in Part 1 of the Handbook.

## 2. Framework and approaches to design and inference

A finite population $\mathcal{U}$ is a set of $N$ units; we write $\mathcal{U} = \{1, 2, \ldots, N\}$. A sample $\mathcal{S}$ is a subset of $\mathcal{U}$. Unit $i$ has an associated $k$-vector of measurements $y_i$. One wishes to estimate or predict functions of $y_1, \ldots, y_N$ using the data in $\mathcal{S}$. Of particular interest is the population total, $Y = \sum_{i=1}^{N} y_i$.

Sometimes auxiliary information is available for units in the population before the sample is selected. Some countries have population registers with detailed information about the population; in other cases, information may be available from administrative records or previous data collection efforts. Let $x_i$ denote the vector of auxiliary information available for unit $i$. The auxiliary information may be used in the survey design, in the estimators, or in both. The fundamental design problem is to use the available auxiliary information to achieve as much precision as possible when estimating population quantities of interest.

Although Part 1 concerns survey design, it begins with a chapter by Brewer and Gregoire on philosophies of survey inference. This is appropriate because the approach that will be taken for inference has implications for the choice of design in the survey. Approaches to inference are treated in more detail in Chapters 23 and 24, but here we give a brief outline to show the relation to survey design.

Neyman (1934) promoted stratified random sampling in association with randomization-based, or design-based, inference. In randomization-based inference, the values $y_i$ are considered to be fixed, but unknown, quantities. The random variables used for inference are $Z_1, \ldots, Z_N$, where $Z_i$ represents the number of times that unit $i$ is selected to be in the sample. If sampling is done without replacement, $Z_i = 1$ if unit $i$ is included in the sample, and $Z_i = 0$ if unit $i$ is not included in the sample. The inclusion probability is $\pi_i = P(Z_i = 1)$ and the probability that a particular sample $\mathcal{S}$ is selected is $P(\mathcal{S}) = P(Z_i = 1, i \in \mathcal{S} \text{ and } Z_j = 0, j \notin \mathcal{S}\}$. The Horvitz–Thompson (1952) estimator of the population total is

$$\hat{Y}_{\mathrm{HT}} = \sum_{i \in \mathcal{S}} \frac{y_i}{\pi_i} = \sum_{i=1}^{N} Z_i \frac{y_i}{\pi_i}$$

with

$$V(\hat{Y}_{\mathrm{HT}}) = \sum_{i=1}^{N} \sum_{j=1}^{N} (\pi_{ij} - \pi_i \pi_j) \frac{y_i}{\pi_i} \frac{y_j}{\pi_j},$$

where $\pi_{ij} = P(Z_i = 1, Z_j = 1)$. The variance of $\hat{Y}_{\mathrm{HT}}$ depends on the joint probability function of the $Z_i$—the actual measurement of interest, $y_i$, is considered to be a constant for inferential purposes.

In model-based inference, also called prediction-based inference, the values $y_1, \ldots, y_N$ in the finite population are assumed to be realizations of random vectors that follow a stochastic model. Adopting the notation in Chapter 1, we let $Y_i$ represent the random variable generating the response for unit $i$. (Note that following standard usage $Y = \sum_{i=1}^{N} y_i$ is still the finite population total.) For a univariate response, the ratio model

$$Y_i = \beta x_i + U_i \tag{1}$$

is occasionally adopted, where the errors $U_i$ are assumed to be independently distributed with mean 0 and variance $x_i\sigma^2$. A prediction estimator of the population total using this model is

$$\hat{Y}_{\text{PRED}} = \sum_{i \in \mathcal{S}} y_i + \sum_{i \notin \mathcal{S}} \hat{\beta} x_i, \tag{2}$$

where $\hat{\beta} = \sum_{i \in \mathcal{S}} Y_i / \sum_{i \in \mathcal{S}} x_i$ is the best linear unbiased estimator of $\beta$ under the model. In a model-based approach, the variance of $\hat{Y}_{\text{PRED}}$ depends on the joint probability distribution, specified by the model, of the $U_i$ for units in the sample: the method used to select the sample is irrelevant for inference because presumably all relevant information is incorporated in the model.

What are the design implications of the inferential approach chosen? For the prediction estimator in (2), the model-based optimal design is that which minimizes the variance of $\hat{\beta}$ under the assumed model, namely a design that purposively selects the $n$ population units with the largest $x$ values to be the sample.

For randomization-based inference, one approach would be to incorporate the auxiliary information into the design through stratification based on the $x$ variable. If $y$ is positively correlated with $x$ and the variability increases with $x$, consistent with the model in (1), then the optimal stratification design will have larger sampling fractions in the strata with large $x$ and smaller sampling fractions in the strata with small $x$. Alternatively, with probability proportional to $x$ sampling, the inclusion probability $\pi_i$ is defined to be proportional to $x_i$; methods for selecting such a sample are described in Chapter 2. Both of these designs exploit the assumed population model structure in (1) and will reduce the randomization-based variance of $\hat{Y}_{\text{HT}}$ if the model approximately holds. They both lead to samples that are likely to contain proportionately more units with large values of $x_i$ than a simple random sample would contain, and in that sense are similar to the optimal design under the prediction approach. Stratification and unequal probability sampling are often used in tandem. For example, in business surveys, discussed in Chapter 17, it is common to first stratify by establishment size and then to sample with probability proportional to size within each stratum.

The optimal designs using stratification or unequal probability sampling have an important difference from the optimal design under the model-based approach: the randomization-based designs have positive probability of inclusion for every unit in the population. Although the stratified design has small sampling fraction in the stratum with the smallest values of $x$, it does prescribe taking observations from that stratum. The optimal model-based design, by contrast, takes no observations from that stratum, and data from that design are inadequate for checking the model assumptions. As Brewer and Gregoire point out in Chapter 1, if the model does not hold for unsampled population units, estimates using data from the optimal model-based design may be biased. For that reason, Royall and Herson (1973) suggested using balanced sampling designs, in which sample moments of auxiliary variables approximately equal the population moments of those variables. This provides a degree of robustness against the model assumptions for the variables included in the balancing. To achieve additional robustness with respect to other, perhaps unavailable, potential covariates, one of the possible balanced samples can be selected using randomization methods.

A probability sampling design is balanced on an auxiliary variable $x$ if the Horvitz–Thompson estimator of the total for $x$ equals the true population total for $x$. Berger and Tillé, in Chapter 2, describe methods for designing samples that are approximately balanced with respect to multiple auxiliary variables. These auxiliary variables can include stratum indicators so that stratified sampling is a special case of balanced sampling; they can also include continuous variables from a population register such as age or educational attainment that cut across the strata. The cube method for selecting samples presents an elegant geometric view of the balanced design problem. The balanced sampling methods presented in Chapter 2 yield probability sampling designs; randomization methods are used to select one of the many possible samples that satisfy the balancing constraints.

With stratification and unequal probability sampling, auxiliary information is used in the design. Alternatively, or additionally, auxiliary information about units or groups of units in the population can be incorporated into the estimator. For example, the ratio estimator $\hat{Y}_{\mathrm{R}} = \hat{Y}_{\mathrm{HT}}(X/\hat{X}_{\mathrm{HT}})$ adjusts the Horvitz–Thompson estimator of $Y$ by the ratio $X/\hat{X}_{\mathrm{HT}}$. If a simple random sample is taken, $\hat{Y}_{\mathrm{R}}$ has the same form as $\hat{Y}_{\mathrm{PRED}}$ from (2); the ratio estimator is motivated by the model in (1), but inference about $\hat{Y}_{\mathrm{R}}$ is based on the distribution of the design variables $Z_i$, while inference about $\hat{Y}_{\mathrm{PRED}}$ depends on the distribution of the model errors $U_i$. The ratio estimator calibrates (see Chapter 25) the Horvitz–Thompson estimator so that the estimated population total of the auxiliary variable coincides with the true total, $X = \sum_{i=1}^{N} x_i$. A stratified design achieves such calibration automatically for the auxiliary variables indicating stratum membership; in stratified random sampling, the Horvitz–Thompson estimator of each stratum size is exact. Balanced sampling extends this precalibration to other variables.

Note that data from a randomization-based design may later be analyzed using model-based inference, provided that relevant design features are incorporated in the model. Indeed, models are essential for treating nonresponse and measurement errors, as will be discussed in Part 2. But data that have been collected using a model-based design must be analyzed with a model-based approach; if no randomization is employed, randomization-based inference cannot be used.

Brewer and Gregoire, in Chapter 1, argue that the prediction and randomization approaches should be used together. In survey design, they can be used together by tentatively adopting a model when designing a randomization-based probability sample. The resulting design will use the auxiliary information to improve efficiency but will be robust to model misspecification. This approach is largely the one adopted in the chapters in Part 1 on specific design problems.

## 3. Challenges in survey design

The framework given in Section 2 is, in a sense, an idealized version of survey design. We assumed that a complete sampling frame exists, that auxiliary information useful for design is available for all units, and that any desired design can be implemented. Chapters 3–7 in the Handbook treat specific problems in survey design in which some of these assumptions are not met. The designs are all developed from the randomization-based perspective but strive to use auxiliary information as efficiently as possible.

Sampling designs are most efficient if they exploit high-quality auxiliary information. Sometimes, though, highly correlated auxiliary information is not available before sampling but can be collected relatively inexpensively in a preliminary survey. In a health survey, for example, one might wish to oversample persons at high risk for coronary heart disease but it is unknown who those persons are before the sample is collected. A phase I sample can be collected in which respondents are asked over the telephone about risk factors and grouped into risk strata on the basis of the verbal responses. In a phase II sample, subsamples of the original respondents are given medical examinations, with higher sampling fractions in the high-risk strata. The efficiency gained by using a two-phase sample depends on the relative costs of sampling in the two phases as well as the efficiency of the stratification of the phase-I respondents. Legg and Fuller, in Chapter 3, discuss recent results in two-phase sampling, including methods for incorporating additional auxiliary information in the estimators and methods for variance estimation. For two-phase samples, designs need to be specified for both phases, and the proportion of resources to be devoted to each phase needs to be determined.

With the introduction of new modes for collecting survey data, in some situations it is difficult to find one sampling frame that includes the entire population. Random digit dialing frames, for example, do not include households without telephones. In other situations, a complete sampling frame exists but is expensive to sample from; another frame, consisting of a list of some of the units in the population, is much cheaper to sample but does not cover the entire population. Chapter 4 discusses the theory and challenges of multiple-frame surveys, in which the union of two or more sampling frames is assumed to cover the population of interest. Sometimes the incomplete frames can be combined, omitting duplicates, to construct a complete sampling frame for the population. Alternatively, independent samples can be selected from the frames, and the information from the samples can be combined to obtain general population estimates. Often, one frame has more auxiliary information available for design purposes than other frames. A list of farms from a previous agricultural census may also have information on farm size, types of crops grown at the census time, and other information that may be used in stratifying or balancing the survey design. If independent samples are taken from the frames, each sample design can fully exploit the auxiliary information available for that frame. As with two-phase sample design, the design of a multiple-frame survey needs to include designs for each frame as well as the relative resources to be devoted to each sample.

Design decisions for surveys in which we are interested in changes over time are discussed in Chapter 5. Kalton distinguishes between surveys designed to estimate changes in population characteristics over time, for example, the change in the national unemployment rate between year 1 and year 2, and surveys designed to estimate gross changes, for example, how many persons move from unemployed status at time 1 to employed status at time 2. A repeated cross-sectional survey, sampling different persons each year, can be used to estimate the change in unemployment from 2010 to 2011, but it cannot be used to answer questions about persistence in unemployment among individuals. A longitudinal survey, following the same persons through repeated interviews, can be used to estimate yearly trends as well as persistence. A longitudinal survey design needs to consider possible attrition and measurement errors that may change over time.

Rare populations, the subject of Chapter 6, are those in which the individuals of interest are a small part of the population, for example, persons with a rare medical condition, or a special type of flower in a forest. In many situations, the auxiliary information that would be most useful for designing the sample, namely, information identifying which units of the sampling frame are in the rare population, is unfortunately unavailable. Thus, as in two-phase sampling, auxiliary information that could greatly improve the efficiency of the survey is unknown before sampling. Christman summarizes several methods that can be used to design surveys for estimating the size and characteristics of a rare population. Auxiliary information that can be used to predict membership in the rare population may be used for stratification. The units can be stratified by their likelihood of belonging to the rare population, and the strata with higher expected membership rates can then be sampled with higher sampling fractions. If that information is not available in advance, two-phase sampling can be used to collect information about rare population membership in phase I, as discussed in Chapter 3.

Christman also describes adaptive sampling designs, in which sampling is done sequentially. An initial sample is used to modify the inclusion probabilities of subsequently selected units. Adaptive sampling designs are particularly useful when the rare group is clustered within the population. In adaptive cluster sampling, clusters adjacent to those with high concentrations or counts of the population of interest receive higher probabilities for inclusion in subsequent sampling. In these adaptive designs, auxiliary information is collected sequentially.

Wolter, Chowdhury, and Kelly, in Chapter 7, update the uses and challenges of random-digit dialing surveys. Since auxiliary information may be limited to demographic summary statistics for the area codes (and even that may not be available if a survey of cellular telephone numbers is taken, where an individual may reside outside of the area code assigned to his/her cell number), the efficiency gained by stratification may be limited and much of the auxiliary information about the population can only be used in the estimation stage. Random-digit dialing surveys face new challenges as landline telephones are being replaced by other technology, but many of the methods used to design a random-digit dialing survey can carry over to newer modes such as cellular telephones and internet surveys.

Many design features described in Part 1 can be used together to improve the efficiency and quality of samples. Wolter, Chowdhury, and Kelly describe how random-digit dialing can be used as one sample in a multiple-frame survey; additional frames, such as a frame of cellular telephone users, can improve the coverage of the population. Multiple-frame surveys can also be used to combine information from surveys taken with different designs and for different purposes. For sampling rare populations, one frame might be a list of persons thought to belong to the rare population, and another frame might be that used for an adaptive cluster sample. In two-phase sampling, the auxiliary information gathered in phase I can be used to design a balanced sample for phase II.

Mahalanobis (1946) and Biemer and Lyberg (2003) emphasized the importance of designing surveys to minimize errors from all sources. The chapters in Part 1 discuss strategies to meet this challenge in new settings. Chapters 1–3 concentrate primarily on using auxiliary information to reduce the sampling variability of estimators. Chapters 4–7 discuss in addition how to handle anticipated effects of nonresponse and measurement errors in the survey design.

*Sample Surveys: Design, Methods and Applications, Vol. 29A*
ISSN: 0169-7161

DOI: 10.1016/S0169-7161(08)00001-1

# 1

# Introduction to Survey Sampling

*Ken Brewer and Timothy G. Gregoire*

## 1. Two alternative approaches to survey sampling inference

### *1.1. Laplace and his ratio estimator*

At some time in the mid-1780s (the exact date is difficult to establish), the eminent mathematician Pierre Laplace started to press the ailing French government to conduct an enumeration of the population in about 700 communes scattered over the Kingdom (Bru, 1988), with a view to estimating the total population of France. He intended to use for this purpose the fact that there was already a substantially complete registration of births in all communes, of which there would then have been of the order of 10,000. He reasoned that if he also knew the populations of those sample communes, he could estimate the ratio of population to annual births, and apply that ratio to the known number of births in a given year, to arrive at what we would now describe as a ratio estimate of the total French population (Laplace, 1783[1], 1814a and 1814b). For various reasons, however, notably the ever-expanding borders of the French empire during Napoleon's early years, events militated against him obtaining a suitable total of births for the entire French population, so his estimated ratio was never used for its original purpose (Bru, 1988; Cochran, 1978; Hald, 1998; Laplace, 1814a and 1814b, p. 762). He did, however, devise an ingenious way for estimating the precision with which that ratio was measured. This was less straightforward than the manner in which it would be estimated today but, at the time, it was a very considerable contribution to the theory of survey sampling.

### *1.2. A prediction model frequently used in survey sampling*

The method used by Laplace to estimate the precision of his estimated ratio was not dependent on the knowledge of results for the individual sample communes, which

[1] This paper is the text of an address given to the Academy on 30 October 1785, but appears to have been incorrectly dated back to 1783 while the Memoirs were being compiled. A virtually identical version of this address also appears in Laplace's *Oeuvres Complètes* 11 pp. 35–46. This version also contains three tables of vital statistics not provided in the Memoirs' version. They should, however, be treated with caution, as they contain several arithmetical inconsistencies.

would normally be required these days for survey sampling inference. The reason why it was not required there is chiefly that a particular model was invoked, namely one of drawing balls from an urn, each black ball representing a French citizen counted in Laplace's sample, and each white ball representing a birth within those sample communes in the average of the three preceding years. As it happens, there is another model frequently used in survey sampling these days, which leads to the same ratio estimator. That model is

$$Y_i = \beta X_i + U_i, \tag{1a}$$

which together with

$$E(U_i) = 0, \tag{1b}$$

$$E\left(U_i^2\right) = \sigma^2 X_i \tag{1c}$$

and

$$E(U_i U_j) = 0 \tag{1d}$$

for all $j \neq i$ can also be used for the same purpose.

Equation (1a) describes a survey variable value $Y_i$ (for instance the population of commune $i$) as generated by a survey parameter, $\beta$, times an auxiliary value, $X_i$, (that commune's average annual births) plus a random variable, $U_i$. Equation (1b) stipulates that this random variable has zero mean, Eq. (1c) that its variance is proportional to the auxiliary variable (in this case, annual births), and Eq. (1d) that there is no correlation between any pair of those random variables.

Given this model, the minimum variance unbiased estimator of $\beta$ is given by

$$\hat{\beta} = \frac{\sum\limits_{i=1}^{n} Y_i}{\sum\limits_{i=1}^{n} X_i}, \tag{2}$$

which in this instance is simply the ratio of black to white balls in Laplace's urn.

### 1.3. The prediction model approach to survey sampling inference

While, given the model of Eqns. (1), the logic behind the ratio estimator might appear to be straightforward, there are in fact two very different ways of arriving at it, one obvious and one somewhat less obvious but no less important. We will examine the obvious one first.

It is indeed obvious that there is a close relationship between births and population. To begin with, most of the small geographical areas (there are a few exceptions such as military barracks and boarding schools) have approximately equal numbers of males and females. The age distribution is not quite so stable, but with a high probability different areas within the same country are likely to have more or less the same age distribution, so the proportion of females of child-bearing age to total population is also more or less constant. So, also with a reasonable measure of assurance, one might expect the ratio of births in a given year to total population to be more or less constant, which makes the ratio estimator an attractive choice.

We may have, therefore, a notion in our minds that the number in the population in the $i$th commune, $Y_i$, is proportional to the number of births there in an average year, $X_i$, plus a random error, $U_i$. If we write that idea down in mathematical form, we arrive at a set of equations similar to (1) above, though possibly with a more general variance structure than that implied by Eqns. (1c) and (1d), and that set would enable us to predict the value of $Y_i$ given only the value of $X_i$ together with an estimate of the ratio $\beta$. Laplace's estimate of $\beta$ was a little over 28.35.

The kind of inference that we have just used is often described as "model-based," but because it is a prediction model and because we shall meet another kind of model very shortly, it is preferable to describe it as "prediction-based," and this is the term that will be used here.

## *1.4. The randomization approach to survey sampling inference*

As already indicated, the other modern approach to survey sampling inference is more subtle, so it will take a little longer to describe. It is convenient to use a reasonably realistic scenario to do so.

The hypothetical country of Oz (which has a great deal more in common with Australia than with Frank L. Baum's mythical Land of Oz) has a population of 20 million people geographically distributed over 10,000 postcodes. These postcodes vary greatly among themselves in population, with much larger numbers of people in a typical urban than in a typical rural postcode.

Oz has a government agency named Centrifuge, which disburses welfare payments widely over the entire country. Its beneficiaries are in various categories such as Age Pensioners, Invalid Pensioners, and University Students. One group of its beneficiaries receives what are called Discretionary Benefits. These are paid to people who do not fall into any of the regular categories but are nevertheless judged to be in need of and/or deserving of financial support.

Centrifuge staff, being human, sometimes mistakenly make payments over and above what their beneficiaries are entitled to. In the Discretionary Benefits category, it is more difficult than usual to determine when such errors (known as overpayments) have been made, so when Centrifuge wanted to arrive at a figure for the amounts of Overpayments to Discretionary Beneficiaries, it decided to do so on a sample basis. Further, since it keeps its records in postcode order, it chose to select 1000 of these at random (one tenth of the total) and to spend considerable time and effort in ensuring that the Overpayments in these sample postcodes were accurately determined. (In what follows, the number of sample postcodes, in this case 1000, will be denoted by $n$ and the number of postcodes in total, in this case 10,000, denoted by $N$.)

The original intention of the Centrifuge sample designers had been to use the same kind of ratio estimator as Laplace had used in 1802, namely

$$\hat{Y} = \frac{\sum_{i=1}^{N} \delta_i Y_i}{\sum_{i=1}^{N} \delta_i X_i} \sum_{i=1}^{N} X_i, \tag{3}$$

with $Y_i$ being the amount of overpayments in the $i$th postcode and $X_i$ the corresponding postcode population. In (3), $\delta_i$ is a binary (1/0) indicator of inclusion into the sample of size $n$: for any particular sample, all but $n$ of the $N$ elements of the population will have a value of $\delta = 0$ so that the sum of $\delta_i Y_i$ over $i = 1 \ldots N$ yields the sum of just the $n$ values of $Y_i$ on those elements selected into the sample.

However, when this proposal came to the attention of a certain senior Centrifuge officer who had a good mathematical education, he queried the use of this ratio estimator on the grounds that the relationship between Overpayments (in this particular category) and Population in individual postcodes was so weak that the use of the model (1) to justify it was extremely precarious. He suggested that the population figures for the selected postcodes should be ignored and that the ratio estimator should be replaced by the simpler expansion estimator, which was

$$\hat{Y} = (N/n) \sum_{i=1}^{N} \delta_i Y_i. \tag{4}$$

When this suggestion was passed on to the survey designers, they saw that it was needed to be treated seriously, but they were still convinced that there was a sufficiently strong relationship between Overpayments and Population for the ratio estimator also to be a serious contender. Before long, one of them found a proof, given in several standard sampling textbooks, that without reliance on any prediction model such as Eqns. (1), the ratio estimator was more efficient than the expansion estimator provided (a) that the sample had been selected randomly from the parent population and (b) that the correlation between the $Y_i$ and the $X_i$ exceeded a certain value (the exact nature of which is irrelevant for the time being). The upshot was that when the sample data became available, that requirement was calculated to be met quite comfortably, and in consequence the ratio estimator was used after all.

### *1.5. A comparison of these two approaches*

The basic lesson to be drawn from the above scenario is that there are two radically different sources of survey sampling inference. The first is prediction on the basis of a mathematical model, of which (1), or something similar to it, is the one most commonly postulated. The other is randomized sampling, which can provide a valid inference regardless of whether the prediction model is a useful one or not. Note that a model can be useful even when it is imperfect. The famous aphorism of G.E.P. Box, "All models are wrong, but some are useful." (Box, 1979), is particularly relevant here.

There are also several other lessons that can be drawn. To begin with, models such as that of Eqns. (1) have parameters. Equation (1a) has the parameter $\beta$, and Eq. (1c) has the parameter $\sigma^2$ that describes the extent of variability in the $Y_i$. By contrast, the randomization-based estimator (4) involves no estimation of any parameter. All the quantities on the right hand side of (4), namely $N$, $n$, and the sample $Y_i$, are known, if not without error, at least without the need for any separate estimation or inference.

In consequence, we may say that estimators based on prediction inference are parametric, whereas those based on randomization inference are nonparametric. Parametric estimators tend to be more accurate than nonparametric estimators when the model on which they are based is sufficiently close to the truth as to be useful, but they are also sensitive to the possibility of model breakdown. By contrast, nonparametric estimators tend to be less efficient than parametric ones, but (since there is no model to break down) they are essentially robust. If an estimator is supported by both parametric and nonparametric inference, it is likely to be both efficient and robust. When the correlation between the sample $Y_i$ and the sample $X_i$ is sufficiently large to meet the relevant condition, mentioned but not defined above in the Oz scenario, the estimator is also likely to be both efficient and robust, but when the correlation fails to meet that condition, another estimator has a better randomization-based support, so the ratio estimator is no longer robust, and the indications are that the expansion estimator, which does not rely upon the usefulness of the prediction model (1), would be preferable.

It could be argued, however, that the expansion estimator itself could be considered as based on the even simpler prediction model

$$Y_i = \alpha + U_i, \tag{5}$$

where the random terms $U_i$ have zero means and zero correlations as before. In this case, the parameter to be estimated is $\alpha$, and it is optimally estimated by the mean of the sample observations $Y_i$. However, the parametrization used here is so simple that the parametric estimator based upon it coincides with the nonparametric estimator provided by randomization inference. This coincidence appears to have occasioned some considerable confusion, especially, but not exclusively, in the early days of survey sampling.

Moreover, it is also possible to regard the randomization approach as implying its own quite different model. Suppose we had a sample in which some of the units had been selected with one chance in ten, others with one chance in two, and the remainder with certainty. (Units selected with certainty are often described as "completely enumerated.") We could then make a model of the population from which such a sample had been selected by including in it (a) the units that had been selected with one chance in ten, together with nine exact copies of each such unit, (b) the units that had been selected with one chance in two, together with a single exact copy of each such unit, and (c) the units that had been included with certainty, but in this instance without any copies. Such a model would be a "randomization model." Further, since it would be a nonparametric model, it would be intrinsically robust, even if better models could be built that did use parameters.

In summary, the distinction between parametric prediction inference and nonparametric randomization inference is quite a vital one, and it is important to bear it in mind as we consider below some of the remarkable vicissitudes that have beset the history of survey sampling from its earliest times and have still by no means come to a definitive end.

## 2. Historical approaches to survey sampling inference

### *2.1. The development of randomization-based inference*

Although, as mentioned above, Laplace had made plans to use the ratio estimator as early as the mid-1780s, modern survey sampling is more usually reckoned as dating from the work of Anders Nicolai Kiaer, the first Director of the Norwegian Central Bureau of Statistics. By 1895, Kiaer, having already conducted sample surveys successfully in his own country for fifteen years or more, had found to his own satisfaction that it was not always necessary to enumerate an entire population to obtain useful information about it. He decided that it was time to convince his peers of this fact and attempted to do so first at the session of the International Statistical Institute (ISI) that was held in Berne that year. He argued there that what he called a "partial investigation," based on a subset of the population units, could indeed provide such information, provided only that the subset had been carefully chosen to reflect the whole of that population in miniature. He described this process as his "representative method," and he was able to gain some initial support for it, notably from his Scandinavian colleagues. Unfortunately, however, his idea of representation was too subjective and lacking in probabilistic rigor to make headway against the then universally held belief that only complete enumerations, "censuses," could provide any useful information (Lie, 2002; Wright, 2001).

It was nevertheless Kiaer's determined effort to overthrow that universally held belief that emboldened Lucien March, at the ISI's Berlin meeting in 1903, to suggest that randomization might provide an objective basis for such a partial investigation (Wright, 2001). This idea was further developed by Arthur Lyon Bowley, first in a theoretical paper (Bowley, 1906) and later by a practical demonstration of its feasibility in a pioneering survey conducted in Reading, England (Bowley, 1912).

By 1925, the ISI at its Rome meeting was sufficiently convinced (largely by the report of a study that it had itself commissioned) to adopt a resolution giving acceptance to the idea of sampling. However, it was left to the discretion of the investigators whether they should use randomized or purposive sampling. With the advantage of hindsight, we may conjecture that, however vague their awareness of the fact, they were intuiting that purposive sampling was under some circumstances capable of delivering accurate estimates, but that under other circumstances, the underpinning of randomization inference would be required.

In the following year, Bowley published a substantial monograph in which he presented what was then known concerning the purposive and randomizing approaches to sample selection and also made suggestions for further developments in both of them (Bowley, 1926). These included the notion of collecting similar units into groups called "strata," including the same proportion of units from each stratum in the sample, and an attempt to make purposive sampling more rigorous by taking into account the correlations between, on the one hand, the variables of interest for the survey and, on the other, any auxiliary variables that could be helpful in the estimation process.

### 2.2. *Neyman's establishment of a randomization orthodoxy*

A few years later, Corrado Gini and Luigi Galvani selected a purposive sample of 29 out of 214 districts (circondari) from the 1921 Italian Population Census (Gini and Galvani, 1929). Their sample was chosen in such a way as to reflect almost exactly the whole-of-Italy average values for seven variables chosen for their importance, but it was shown by Jerzy Neyman (1934) that it exhibited substantial differences from those averages for other important variables.

Neyman went on to attack this study with a three pronged argument. His criticisms may be summarized as follows:

(1) Because randomization had not been used, the investigators had not been able to invoke the Central Limit Theorem. Consequently, they had been unable to use

the normality of the estimates to construct the confidence intervals that Neyman himself had recently invented and which appeared in English for the first time in his 1934 paper.

(2) On the investigators' own admission, the difficulty of achieving their "purposive" requirement (that the sample match the population closely on seven variables) had caused them to limit their attention to the 214 districts rather than to the 8354 communes into which Italy had also been divided. In consequence, their 15% sample consisted of only 29 districts (instead of perhaps 1200 or 1300 communes). Neyman further showed that a considerably more accurate set of estimates could have been expected had the sample consisted of this larger number of smaller units. Regardless of whether the decision to use districts had required the use of purposive sampling, or whether the causation was the other way round, it was evident that purposive sampling and samples consisting of far too few units went hand in hand.

(3) The population model used by the investigators was demonstrably unrealistic and inappropriate. Models by their very nature were always liable to represent the actual situation inadequately. Randomization obviated the need for population modeling.[2] With randomization-based inference, the statistical properties of an estimator are reckoned with respect to the distribution of its estimates from all samples that might possibly be drawn using the design under consideration. The same estimator under different designs will admit to differing statistical properties. For example, an estimator that is unbiased under an equal probability design (see Section 3 of this chapter for an elucidation of various designs that are in common use) may well be biased under an unequal probability design.

In the event, the ideas that Neyman had presented in this paper, though relevant for their time and well presented, caught on only gradually over the course of the next decade. W. Edwards Deming heard Neyman in London in 1936 and soon arranged for him to lecture, and his approach to be taught, to U.S. government statisticians. A crucial event in its acceptance was the use in the 1940 U.S. Population and Housing Census of a one-in-twenty sample designed by Deming, along with Morris Hansen and others, to obtain answers to additional questions. Once accepted, however, Neyman's arguments swept all other considerations aside for at least two decades.

Those twenty odd years were a time of great progress. In the terms introduced by Kuhn (1996), finite population sampling had found a universally accepted "paradigm" (or "disciplinary matrix") in randomization-based inference, and an unusually fruitful period of normal science had ensued. Several influential sampling textbooks were published, including most importantly those by Hansen et al. (1953) and by Cochran (1953, 1963). Other advances included the use of self-weighting, multistage, unequal probability samples by Hansen and Hurwitz at the U.S. Bureau of the Census, Mahalanobis's invention of interpenetrating samples to simplify the estimation of variance for complex survey designs and to measure and control the incidence of nonsampling errors, and the beginnings of what later came to be described as "model-assisted survey sampling."

[2] The model of Eqns. (1) above had not been published at the time of Neyman's presentation. It is believed first to have appeared in Fairfield Smith (1938) in the context of a survey of agricultural crops. Another early example of its use is in Jessen (1942).

A lone challenge to this orthodoxy was voiced by Godambe (1955) with his proof of the nonexistence of any uniformly best randomization-based estimator of the population mean, but few others working in this excitingly innovative field seemed to be concerned by this result.

### *2.3. Model-assisted or model-based? The controversy over prediction inference*

It therefore came as a considerable shock to the finite population sampling establishment when Royall (1970b) issued his highly readable call to arms for the reinstatement of purposive sampling and prediction-based inference. To read this paper was to read Neyman (1934) being stood on its head. The identical issues were being considered, but the opposite conclusions were being drawn.

By 1973, Royall had abandoned the most extreme of his recommendations. This was that the best sample to select would be the one that was optimal in terms of a model closely resembling Eqns. (1). (That sample would typically have consisted of the largest $n$ units in the population, asking for trouble if the parameter $\beta$ had not in fact been constant over the entire range of sizes of the population units.) In Royall and Herson (1973a and 1973b), the authors suggested instead that the sample should be chosen to be "balanced", in other words that the moments of the sample $X_i$ should be as close as possible to the corresponding moments of the entire population. (This was very similar to the much earlier notion that samples should be chosen purposively to resemble the population in miniature, and the samples of Gini and Galvani (1929) had been chosen in much that same way!)

With that exception, Royall's original stand remained unshaken. The business of a sampling statistician was to make a model of the relevant population, design a sample to estimate its parameters, and make all inferences regarding that population in terms of those parameter estimates. The randomization-based concept of defining the variance of an estimator in terms of the variability of its estimates over all possible samples was to be discarded in favor of the prediction variance, which was sample-specific and based on averaging over all possible realizations of the chosen prediction model.

Sampling statisticians had at no stage been slow to take sides in this debate. Now the battle lines were drawn. The heat of the argument appears to have been exacerbated by language blocks; for instance, the words "expectation" and "variance" carried one set of connotations for randomization-based inference and quite another for prediction-based inference. Assertions made on one side would therefore have appeared as unintelligible nonsense by the other.

A major establishment counterattack was launched with Hansen et al. (1983). A small (and by most standards undetectable) divergence from Royall's model was shown nevertheless to be capable of distorting the sample inferences substantially. The obvious answer would surely have been "But this distortion would not have occurred if the sample had been drawn in a balanced fashion. Haven't you read Royall and Herson (1973a and b)?" Strangely, it does not seem to have been presented at the time.

Much later, a third position was also offered, the one held by the present authors, namely that since there were merits in both approaches, and that it was possible to combine them, the two should be used together. For the purposes of this Handbook volume, it is necessary to consider all three positions as dispassionately as possible. Much can be gained by asking the question as to whether Neyman (1934) or Royall

(1970b) provided the more credible interpretation of the facts, both as they existed in 1934 or 1970 and also at the present day (2009).

### *2.4. A closer look at Neyman's criticisms of Gini and Galvani*

The proposition will be presented here that Neyman's criticisms and prescriptions were appropriate for his time, but that they have been overtaken by events. Consider first his contention that without randomization, it was impossible to use confidence intervals to measure the accuracy of the sample estimates.

This argument was received coolly enough at the time. In moving the vote of thanks to Neyman at the time of the paper's presentation, Bowley wondered aloud whether confidence intervals were a "confidence trick." He asked "Does [a confidence interval] really lead us to what we need—the chance that within the universe which we are sampling the proportion is within these certain limits? I think it does not. I think we are in the position of knowing that *either* an improbable event had occurred *or* the proportion in the population is within these limits... The statement of the theory is not convincing, and until I am convinced I am doubtful of its validity."

In his reply, Neyman pointed out that Bowley's question in the first quoted sentence above "contain[ed] the statement of the problem in the form of Bayes" and that in consequence its solution "*must* depend upon the probability law *a priori*." He added "In so far as we keep to the old form of the problem, any further progress is impossible." He thus concluded that there was a need to stop asking Bowley's "Bayesian" question and instead adopt the stance that the "*either...or*" statement contained in his second quoted sentence "form[ed] a basis for the practical work of a statistician concerned with problems of estimation." There can be little doubt but that Neyman's suggestion was a useful prescription for the time, and the enormous amount of valuable work that has since been done using Neyman and Pearson's confidence intervals is witness to this.

However, the fact remains that confidence intervals are not easy to understand. A confidence interval is in fact a sample-specific range of potentially true values of the parameter being estimated, which has been constructed so as to have a particular property. This property is that, over a large number of sample observations, the proportion of times that the true parameter value falls inside that range (constructed for each sample separately) is equal to a predetermined value known as the confidence level. This confidence level is conventionally written as $(1-\alpha)$, where $\alpha$ is small compared with unity. Conventional choices for $\alpha$ are 0.05, 0.01, and sometimes 0.001. Thus, if many samples of size $n$ are drawn independently from a normal distribution and the relevant confidence interval for $\alpha = 0.05$ is calculated for each sample, the proportion of times that the true parameter value will lie within any given sample's own confidence interval will, before that sample is selected, be 0.95, or 95%.

It is not the case, however, that the probability of this true parameter value lying within the confidence interval as calculated for any individual sample of size $n$ will be 95%. The confidence interval calculated for any individual sample of size $n$ will, in general, be wider or narrower than average and might be centered well away from the true parameter value, especially if $n$ is small. It is also sometimes possible to recognize when a sample is atypical and, hence, make the informed guess that in this particular case, the probability of the true value lying in a particular 95% confidence interval differs substantially from 0.95.

If, however, an agreement is made beforehand that a long succession of wagers is to be made on the basis that (say) Fred will give Harry \$1 every time the true value lies inside any random sample's properly calculated 95% confidence interval, and Harry will give Fred \$19 each time it does not; then at the end of that long sequence, those two gamblers would be back close to where they started. In those circumstances, the 95% confidence interval would also be identical with the 95% Bayesian credibility interval that would be obtained with a flat prior distribution over the entire real line ranging from minus infinity to plus infinity. In that instance, Bowley's "Bayesian question" could be given an unequivocally affirmative answer.

The result of one type of classical hypothesis test is also closely related to the confidence interval. Hypothesis tests are seldom applied to data obtained from household or establishment surveys, but they are frequently used in other survey sampling contexts.

The type of classical test contemplated here is often used in medical trials. The hypothesis to be tested is that a newly devised medical treatment is superior to an existing standard treatment, for which the effectiveness is known without appreciable error. In this situation, there can never be any reason to imagine that the two treatments are identically effective so that event can unquestionably be accorded the probability zero. The probability that the alternative treatment is the better one can then legitimately be estimated by the proportion of the area under the likelihood function that corresponds to values greater than the standard treatment's effectiveness. Moreover, if that standard effectiveness happens to be lower than that at the lower end of the one-sided 95% confidence interval, it can reasonably be claimed that the new treatment is superior to the standard one "with 95% confidence."

However, in that situation, the investigators might well wish to go further and quote the proportion of the area corresponding to all values less than standard treatment's effectiveness (Fisher's $p$-statistic). If, for instance, that proportion were 0.015, they might wish to claim that the new treatment was superior "with 98.5% confidence." To do so might invite the objection that the language used was inappropriate because Neyman's $\alpha$ was an arbitrarily chosen fixed value, whereas Fisher's $p$ was a realization of a random variable, but the close similarity between the two situations would be undeniable. For further discussions of this distinction, see Hubbard and Bayarri (2003) and Berger (2003).

The situation would have been entirely different, however, had the investigation been directed to the question as to whether an additional parameter was required for a given regression model to be realistic. Such questions often arise in contexts such as biodiversity surveys and sociological studies. It is then necessary to accord the null hypothesis value itself (which is usually but not always zero) a nonzero probability. It is becoming increasingly well recognized that in these circumstances, the face value of Fisher's $p$ can give a grossly misleading estimate of the probability that an additional parameter is needed. A relatively new concept, the "false discovery rate" (Benjamini and Hochberg, 1995; Benjamini and Yekutieli, 2001; Efron et al., 2001; Sorić, 1989), can be used to provide useful insights. To summarize the findings in these papers very briefly, those false discovery rates observed empirically have, more often than not, been found to exceed the corresponding $p$-statistic by a considerable order of magnitude.

It is also relevant to mention that the populations met with in finite population sampling, and especially those encountered in establishment surveys, are often far removed from obeying a normal distribution, and that with the smaller samples often selected from them, the assumption of normality for the consequent estimators is unlikely even to produce accurate confidence intervals!

Nevertheless, and despite the misgivings presented above, it is still the case that randomization does provide a useful basis for the estimation of a sample variance. The criterion of minimizing that variance is also a useful one for determining optimum estimators. However, we should not expect randomization alone to provide anything further.

Neyman's second contention was that purposive sampling and samples consisting of fewer than an adequate number of units went hand in hand. This was undoubtedly the case in the 1930s, but a similar kind of matching of sample to population (Royall and his co-authors use the expression "balanced sampling") can now be undertaken quite rapidly using third-generation computers, provided only that the matching is not made on too many variables simultaneously. Brewer (1999a) presents a case that it might be preferable to choose a sample randomly and use calibrated estimators to compensate for any lack of balance, rather than to go to the trouble of selecting balanced samples. However, those who prefer to use balanced sampling can now select randomly from among many balanced or nearly balanced samples using the "cube method" (Deville and Tillé, 2004). This paper also contains several references to earlier methods for selecting balanced samples.

Neyman's third contention was basically that population models were not to be trusted. It is difficult here to improve on the earlier quote from George Box that "All models are wrong, but some models are useful." Equations (1) above provide a very simple model that has been in use since 1938. It relates a variable of interest in a sample survey to an auxiliary variable, all the population values of which are conveniently known.

In its simplest form, the relationship between these variables is assumed to be basically proportional but with a random term modifying that proportional relationship for each population unit. (Admittedly, in some instances, it is convenient to add an intercept term, or to have more than one regressor variable, and/or an additional equation to model the variance of that equation's random term, but nevertheless that simple model can be adequate in a remarkably wide set of circumstances.)

As previously mentioned, such models have been used quite frequently in survey sampling. However, it is one thing to use a prediction model to improve on an existing randomization-based estimator (as was done in the Oz scenario above) and it is quite another thing actually to base one's sampling inference on that model. The former, or "model-assisted" approach to survey sampling inference, is clearly distinguished from prediction-based inference proper in the following quotation, taken from the Preface to the encyclopedic book, *Model Assisted Survey Sampling* by Särndal et al. (1992, also available in paperbook 2003):

> Statistical modeling has strongly influenced survey sampling theory in recent years. In this book, sampling theory is assisted by modeling. It becomes simple to explain how the auxiliary information in a given survey will lead to a particular estimation technique. The teaching of sampling and the style of presentation in journal articles have changed a great deal by this new emphasis. Readers of this book will become familiar with this new style.

> We use the randomization theory or design-based point of view. This is the traditional mode of inference in surveys, ever since the sampling breakthroughs in the 1930s and 1940s. The reasoning is familiar to survey statisticians in government and elsewhere.

As this quotation indicates, using a prediction model to form an estimator as Royall proposed, without regard to any justification in terms of randomization theory, is quite a different approach. It is often described as "model-based," or pejoratively as "model-dependent," but it appears preferable to use the expression, "prediction-based."

A seminal paper attacking the use of a prediction model for such purposes was that by Hansen et al. (1983), which has already been mentioned; but there can be no serious doubt attached to the proposition that this model provides a reasonable first approximation to many real situations. Once again, Neyman's contention has been overtaken by events.

### *2.5. Other recent developments in sample survey inference*

A similarly detailed assessment of the now classic papers written by Royall and his colleagues in the 1970s and early 1980s is less necessary, since there have been fewer changes since they were written, but it is worth providing a short summary of some of them. Royall (1970b) has already been mentioned as having turned Neyman (1934) on its head. Royall (1971) took the same arguments a stage further. In Royall and Herson (1973a and 1973b), there is an implicit admission that selecting the sample that minimized the prediction-based variance (prediction variance) was not a viable strategy. The suggestion offered there is to select balanced samples instead: ones that reflect the moments of the parent population. In this recommendation, it recalls the early twentieth-century preoccupation with finding a sample that resembled the population in miniature but, as has been indicated above, this does not necessarily count against it.

Royall (1976) provides a useful and entertaining introduction to prediction-based inference, written at a time when the early criticisms of it had been fully taken into account. Joint papers by Royall and Eberhardt (1975) and Royall and Cumberland (1978, 1981a and 1981b) deal with various aspects of prediction variance estimation, whereas Cumberland and Royall (1981) offer a prediction-based consideration of unequal probability sampling. The book by Valliant et al. (2000) provides a comprehensive account of survey sampling from the prediction-based viewpoint up to that date, and that by Bolfarine and Zacks (1992) presents a Bayesian perspective on it.

Significant contributions have also been made by other authors. Bardsley and Chambers (1984) offered ridge regression as an alternative to pure calibration when the number of regressor variables was substantial. Chambers and Dunstan (1986) and Chambers et al. (1992) considered the estimation of distribution functions from a prediction-based standpoint. Chambers et al. (1993) and Chambers and Kokic (1993) deal specifically with questions of robustness against model breakdown. A more considerable bibliography of important papers relating to prediction-inference can be found in Valliant et al. (2000).

The randomization-based literature over recent years has been far too extensive to reference in the same detail, and in any case comparatively little of it deals with the question of sampling inference. However, two publications already mentioned above are of especial importance. These are the polemical paper by Hansen et al. (1983) and the highly influential text-book by Särndal et al. (1992), which sets out explicitly to indicate what can be achieved by using model-assisted methods of sample estimation without the explicit use of prediction-based inference. Other recent papers of particular interest in this field include Deville and Särndal (1992) and Deville et al. (1993).

Publications advocating or even mentioning the use of both forms of inference simultaneously are few in number. Brewer (1994) would seem to be the earliest to appear in print. It was written in anticipation of and to improve upon Brewer (1995), which faithfully records what the author was advocating at the First International Conference on Establishment Surveys in 1993, but was subsequently found not to be as efficient or even as workable as the alternative provided in Brewer (1994). A few years later, Brewer (1999a) compared stratified balanced with stratified random sampling and Brewer (1999b) provided a detailed description of how the two inferences could be used simultaneously in unequal probability sampling; also Brewer's (2002) textbook has provided yet further details on this topic, including some unsought spin-offs that follow from their simultaneous use, and an extension to multistage sampling.

All three views are still held. The establishment view is that model-assisted randomization-based inference has worked well for several decades, and there is insufficient reason to change. The prediction-based approach continues to be presented by others as the only one that can consistently be held by a well-educated statistician. And a few say "Why not use both?" Only time and experience are likely to resolve the issue, but in the meantime, all three views need to be clearly understood.

## 3. Some common sampling strategies

### *3.1. Some ground-clearing definitions*

So far, we have only been broadly considering the options that the sampling statistician has when making inferences from the sample to the population from which it was drawn. It is now time to consider the specifics, and for that we will need to use certain definitions.

A *sample design* is a procedure for selecting a sample from a population in a specific fashion. These are some examples:

- simple random sampling with and without replacement;
- random sampling with unequal probabilities, again with and without replacement;
- systematic sampling with equal or unequal probabilities;
- stratified sampling, in which the population units are first classified into groups or "strata" having certain properties in common;
- two-phase sampling, in which a large sample is drawn at the first phase and a subsample from that large sample at the second phase;
- multistage sampling, usually in the context of area sampling, in which a sample of (necessarily large) first-stage units is selected first, samples within those first-stage sample units at the second stage, and so on for possibly third and fourth stages; and
- permanent random number sampling, in which each population unit is assigned a number, and the sample at any time is defined in terms of the ranges of those permanent random numbers that are to be in sample at that time.

This list is not exhaustive, and any given sample may have more than one of those characteristics. For instance, a sample could be of three stages, with stratification and unequal probability sampling at the first stage, unstratified unequal probability sampling at the second stage, and systematic random sampling with equal probabilities at the third stage. Subsequently, subsamples could be drawn from that sample, converting it into a multiphase multistage sample design.

A *sample estimate* is a statistic produced using sample data that can give users an indication as to the value of a population quantity. Special attention will be paid in this section to estimates of population total and population mean because these loom so large in the responsibilities of national statistical offices, but there are many sample surveys that have more ambitious objectives and may be set up so as to estimate small domain totals, regression and/or correlation coefficients, measures of dispersion, or even conceivably coefficients of heteroskedasticity (measures of the extent to which the variance of the $U_i$ can itself vary with the size of the auxiliary variable $X_i$).

A *sample estimator* is a prescription, usually a mathematical formula, indicating how estimates of population quantities are to be obtained from the sample survey data.

An *estimation procedure* is a specification as to what sample estimators are to be used in a given sample survey.

A *sample strategy* is a combination of a sample design and an estimation procedure. Given a specific sample strategy, it is possible to work out what estimates can be produced and how accurately those estimates can be made.

One consequence of the fact that two quite disparate inferential approaches can be used to form survey estimators is that considerable care needs to be taken in the choice of notation. In statistical practice generally, random variables are represented by uppercase symbols and fixed numbers by lowercase symbols, but between the two approaches, an observed value automatically changes its status. Specifically, in both approaches, a sample value can be represented as the product of a population value and the inclusion indicator, $\delta$, which was introduced in (3). However, in the prediction-based approach, the population value is a random variable and the inclusion indicator is a fixed number, whereas in the randomization-based approach, it is the inclusion indicator that is the random variable while the population value is a fixed number. There is no ideal way to resolve this notational problem, but we shall continue to denote population values by, say, $Y_i$ or $X_i$ and sample values by $\delta_i Y_i$ or $\delta_i X_i$, as we did in Eq. (3).

## 3.2. *Equal probability sampling with the expansion estimator*

In what follows, the sample strategies will first be presented in the context of randomization-based inference, then that of the nearest equivalent in prediction-based inference, and finally, wherever appropriate, there will be a note as to how they can be combined.

### 3.2.1. *Simple random sampling with replacement using the expansion estimator*

From a randomization-based standpoint, simple random sampling with replacement (srswr) is the simplest of all selection procedures. It is appropriate for use where (a) the population consists of units whose sizes are not themselves known, but are known not to differ too greatly amongst themselves, and (b) it has no geographical or hierarchical structure that might be useful for stratification or area sampling purposes. Examples are populations of easily accessible individuals or households, administrative records relating to individuals, households, or family businesses; and franchise holders in a large franchise.

The number of population units is assumed known, say $N$, and a sample is selected by drawing a single unit from this population, completely at random, $n$ times. Each time a unit is drawn, its identity is recorded, and the unit so drawn is returned to the population so that it stands exactly the same chance of being selected at any subsequent draw as it did at the first draw. At the end of the $n$ draws, the $i$th population unit appears in the sample $\nu_i$ times, where $\nu_i$ is a number between 0 and $n$, and the sum of the $\nu_i$ over the population is $n$.

The typical survey variable value on the $i$th population unit may be denoted by $Y_i$. The population total of the $Y_i$ may be written $Y$. A randomization-unbiased estimator of $Y$ is the *expansion estimator*, namely $\hat{Y} = (N/n)\sum_{i=1}^{N} \nu_i Y_i$. (To form the corresponding randomization-unbiased estimator of the population mean, $\bar{Y} = Y/N$, replace the expression $N/n$ in this paragraph by $1/n$.)

The randomization variance of the estimator $\hat{Y}$ is $V(\hat{Y}) = (N^2/n)S_{wr}^2$, where $S_{wr}^2 = N^{-1}\sum_{i=1}^{N}(Y_i - \bar{Y})^2$. $V(\hat{Y})$ is in turn estimated randomization-unbiasedly by $(N^2/n)\hat{S}_{wr}^2$, where $\hat{S}_{wr}^2 = N^{-1}\sum_{i=1}^{N}\nu_i(Y_i - \bar{Y})^2$. (To form the corresponding expressions for the population mean, replace the expression $N^2/n$ throughout this paragraph by $1/n$. Since these changes from population total to population mean are fairly obvious, they will not be repeated for other sampling strategies.) Full derivations of these formulae will be found in most sampling textbooks.

There is no simple prediction-based counterpart to srswr. From the point of view of prediction-based inference, multiple appearances of a population unit add no information additional to that provided by the first appearance. Even from the randomization standpoint, srswr is seldom called for, as simple random sampling without replacement (or srswor) is more efficient. Simple random sampling with replacement is considered here purely on account of its extremely simple randomization variance and variance estimator, and because (by comparison with it) both the extra efficiency of srswor and the extra complications involved in its use can be readily appreciated.

### *3.2.2. Simple random sampling without replacement using the expansion estimator*

This sample design is identical with srswr, except that instead of allowing selected population units to be selected again at later draws, units already selected are given no subsequent probabilities of selection. In consequence, the units not yet selected have higher conditional probabilities of being selected at later draws. Because the expected number of distinct units included in sample is always $n$ (the maximum possible number under srswr), the srswor estimators of population total and mean have smaller variances than their srswr counterparts. A randomization-unbiased estimator of $Y$ is again $\hat{Y} = (N/n)\sum_{i=1}^{N} \nu_i Y_i$, but since under srswor the $\nu_i$ take only the values 0 and 1, it will be convenient hereafter to use a different symbol, $\delta_i$, in its place.

The randomization variance of the estimator $\hat{Y}$ is $V(\hat{Y}) = (N - n)(N/n)S^2$, where $S^2 = (N-1)^{-1}\sum_{i=1}^{N}(Y_i - \bar{Y})^2$. The variance estimator $V(\hat{Y})$ is in turn estimated randomization-unbiasedly by $(N-n)(N/n)\hat{S}^2$, where $\hat{S}^2 = (n - 1)^{-1} \sum_{i=1}^{N} \delta_i (Y_i - \hat{\bar{Y}})^2$. The substitution of the factor $N^2$ (in the srswr formulae for the variance and the unbiased variance estimator) by the factor $N(N-n)$ (in the corresponding srswor formulae) is indicative of the extent to which the use of sampling without replacement reduces the variance.

Note, however, that the sampling fraction, $n/N$, is not particularly influential in reducing the variance, even for srswor, unless $n/N$ is an appreciable fraction of unity. An estimate of a proportion obtained from an srswor sample of 3000 people in, say, Wales, is not appreciably any more accurate than the corresponding estimate obtained from a sample of 3000 people in the United States; and this is despite the proportion of Welsh people in the first sample being about 1 in 1000 and the proportion of Americans in the second being only 1 in 100,000. For thin samples like these, such variances are to all intents and purposes inversely proportional to the sample size, and the percentage standard errors are inversely proportional to the square root of the sample size. Full derivations of these formulae will be again be found in most sampling textbooks.

Since srswor is both more efficient and more convenient than srswr, it will be assumed, from this point on, that sampling is without replacement unless otherwise specified. One important variant on srswor, which also results in sampling without replacement, is systematic sampling with equal probabilities, and this is the next sampling design that will be considered.

### *3.2.3. Systematic sampling with equal probabilities, using the expansion estimator*

Systematic sampling, by definition, is the selection of sample units from a comprehensive list using a constant skip interval between neighboring selections. If, for instance, the skip interval is 10, then one possible systematic sample from a population of 104 would consist of the second unit in order, then the 12th, the 22nd, etc. up to and including the 102nd unit in order. This sample would be selected if the starting point (usually chosen randomly as a number between 1 and the skip interval) was chosen to be 2. The sample size would then be 11 units with probability 0.4 and 10 units with probability 0.6, and the expected sample size would be 10.4, or more generally the population size divided by the skip interval.

There are two important subcases of such systematic selection. The first is where the population is deliberately randomized in order prior to selection. The only substantial difference between this kind of systematic selection and srswor is that in the latter case, the sample size is fixed, whereas in the former it is a random variable. Even from the strictest possible randomization standpoint, however, it is possible to consider the selection procedure as conditioned on the selection of the particular random start (in this case 2), in which case the sample size would be fixed at 10 and the srswor theory would then hold without any modification. This conditional randomization theory is used very commonly, and from a model-assisted point of view it is totally acceptable.

That is emphatically not true, however, for the second subcase, where the population is not deliberately randomized in order prior to selection. Randomization theory in that subcase is not appropriate and it could be quite dangerous to apply it. In an extreme case, the 104 units could be soldiers, and every 10th one from the 3rd onwards could be a sergeant, the remainder being privates. In that case, the sample selected above would consist entirely of privates, and if the random start had been three rather than two, the sample would have been entirely one of sergeants. This, however, is a rare and easily detectable situation within this nonrandomized subcase. A more likely situation would be one where the population had been ordered according to some informative characteristic, such as age. In that instance, the sample would in one sense be a highly desirable one, reflecting the age distribution of the population better than by chance. That would be the kind of sample that the early pioneers of survey sampling would have been seeking with their purposive sampling, one that reflected in miniature the properties of the population as a whole.

From the randomization standpoint, however, that sample would have had two defects, one obvious and one rather more subtle. Consider a sample survey aimed at estimating the level of health in the population of 104 persons as a whole. The obvious defect would be that although the obvious estimate based on the systematic sample would reflect that level considerably more accurately than one based on a random sample would have done, the randomization-based estimate of its variance would not provide an appropriate measure of its accuracy.

The more subtle defect is that the randomization-based estimate of its variance would in fact tend to overestimate even what the variance would have been if a randomized sample had been selected. So the systematic sample would tend to reduce the actual variance but slightly inflate the estimated variance! (This last point is indeed a subtle one, and most readers should not worry if they are not able to work out why this should be. It has to do with the fact that the average squared distance between sample units is slightly greater for a systematic sample than it is for a purely random sample.)

In summary, then, systematic sampling is temptingly easy to use and in most cases will yield a better estimate than a purely randomized sample of the same size, but the estimated variance would not reflect this betterment, and in some instances a systematic sample could produce a radically unsuitable and misleading sample. To be on the safe side, therefore, it would be advisable to randomize the order of the population units before selection and to use the srswor theory to analyze the sample.

### *3.2.4. Simple prediction inference using the expansion estimator*

Simple random sampling without replacement does have a prediction-based counterpart. The appropriate prediction model is the special case of Eqns. (1) in which all the $X_i$ take the value unity. The prediction variances of the $U_i$ in (1c) are in this instance all the same, at $\sigma^2$. Because this very simple model is being taken as an accurate refection of reality, it would not matter, in theory, how the sample was selected. It could (to take the extreme case) be a "convenience sample" consisting of all the people in the relevant defined category whom the survey investigator happened to know personally, but of course, in practice, the use of such a "convenience sample" would make the assumptions underlying the equality of the $X_i$ very hard to accept. It would be much more convincing if the sample were chosen randomly from a carefully compiled list, which would then be an srswor sample, and it is not surprising that the formulae relevant to this form of prediction sampling inference should be virtually identical to those for randomization sampling srswor.

The minimum-variance prediction-unbiased estimator of $Y$ under the simple prediction model described in the previous paragraph is identical with the randomization-unbiased estimator under srswor, namely $\hat{Y} = (N/n)\sum_{i=1}^{N} \delta_i Y_i$. Further, the prediction variance of $\hat{Y}$ is $V(\hat{Y}) = (N - n)(N/n)\sigma^2$. A prediction-unbiased estimator of $V(\hat{Y})$ is $\hat{V}(\hat{Y}) = (N - n)(N/n)\hat{\sigma}^2$, where $\hat{\sigma}^2 = (n - 1)^{-1} \sum_{i=1}^{N} (\delta_i Y_i - \hat{\hat{Y}})^2$ where $\hat{\hat{Y}} = \hat{Y}/N$. Note that although the prediction variance is typically sample-specific, in this instance it is the same for all samples. However, the estimated prediction variance does, as always, vary from sample to sample.

### *3.3. Equal probability sampling with the ratio estimator*

So far, we have been using estimators that depend only on the sample observations $Y_i$ themselves. More often than not, however, the sampling statistician has at hand relevant auxiliary information regarding most of the units in the population. We have already noted that Laplace, back at the turn of the 19th century, had access (at least in principle) to annual birth registration figures that were approximately proportional to the population figures that he was attempting to estimate. To take a typical modern example, the population for a Survey of Retail Establishments (shops) would typically consist mainly of shops that had already been in existence at the time of the most recent complete Census of Retail Establishments, and the principal information collected at that Census would have been the sales figures for the previous calendar or financial year. Current sales would, for most establishments and for a reasonable period, remain approximately proportional to those Census sales figures.

Returning to the model of Eqns. (1), we may equate the $Y_i$ with the current sales of the sample establishments, the $X_i$ with the Census sales of the sample and nonsample establishments, and the $X$ with the total Census sales over all sample and nonsample establishments combined. It may be remembered that "Centrifuge's" ratio estimators worked well both when the model of Eqns. (1) was a useful one and also in the weaker situation when there was a comparatively modest correlation between the $Y_i$ and the $X_i$. In a similar fashion, the corresponding ratio estimator for this Survey of Retail Establishments tends to outperform the corresponding expansion estimator, at least until it is time to conduct the next Census of Retail Establishments, which would typically be some time in the next 5–10 years.

It was stated above that the population for a Census of Retail Establishments would typically consist mainly of shops that had already been in existence at the time of the most recent complete Census. Such shops would make up the "Main Subuniverse" for the survey. In practice, there would usually be a substantial minority of shops of which the existence would be known, but which had not been in business at the time of that Census, and for these there would be a separate "New Business Subuniverse," which for want of a suitable auxiliary variable would need to be estimated using an expansion estimator, and in times of rapid growth there might even be an "Unlisted New Business Provision" to allow for the sales of shops that were so new that their existence was merely inferred on the basis of previous experience. Nevertheless, even then, the main core of the estimate of survey period sales would still be the sales of shops in the Main Subuniverse, these sales would be based on Ratio Estimation, and the relevant Ratio Estimator would be the product of the $\hat{\beta}$ of Eq. (2) and the Total of Census Sales $X$.

The modern way of estimating the variance of that ratio estimator depends on whether the relevant variance to be estimated is the randomization variance, which is based on the variability of the estimates over all possible samples, or whether it is the prediction variance, which is sample specific. (For a discussion of the difference between the randomization and prediction approaches to inference, the reader may wish to refer back to Sections 1.3 and 1.4.)

The most common practice at present is to estimate the randomization-variance, and for that the procedure is as follows: denote the population total of the $Y_i$ by $Y$, its expansion estimator by $\hat{Y}$, and its ratio estimator by $\hat{Y}_R$. Then the randomization variance of $\hat{Y}_R$ is approximated by

$$V(\hat{Y}_R) \approx V(\hat{Y}) + \beta^2 V(\hat{X}) - 2\beta C(\hat{Y}, \hat{X}), \tag{6}$$

where $\beta$ is the same parameter as in Eq. (1a), $V(\hat{Y})$ is the randomization variance of the expansion estimator of $Y$, $V(\hat{X})$ is the variance of the corresponding expansion estimator of $X$, based on the same sample size, and $C(\hat{Y}, \hat{X})$ is the covariance between those two estimators.

The approximate randomization-variance of $\hat{Y}_R$ can therefore be estimated by

$$\hat{V}(\hat{Y}_R) = \hat{V}(\hat{Y}) + \hat{\beta}^2 \hat{V}(\hat{X}) - 2\hat{\beta}\hat{C}(\hat{Y}, \hat{X}), \tag{7}$$

where $\hat{V}(\hat{Y})$ is the randomization-unbiased estimator of $V(\hat{Y})$, given in Subsection 3.2.2, $\hat{V}(\hat{X})$ is the corresponding expression in the $X$-variable, $\hat{C}(\hat{Y}, \hat{X})$ is the corresponding expression for the randomization-unbiased estimator of covariance between them, namely $(N - n)(N/n) \sum_{i=1}^{N} \delta_i(Y_i - \bar{Y})(X_i - \bar{X})$, and $\hat{\beta}$ is the sample estimator of $\beta$, as given in Eq. (2).

### *3.4. Simple balanced sampling with the expansion estimator*

An alternative to simple random sampling is simple balanced sampling, which has already been referred to in Section 2.3. When the sample has been selected in such a way as to be balanced on the auxiliary variables $X_i$, in the way described in that section, the expansion estimator is comparable in accuracy to that section's ratio estimator itself. This is because the expansion estimator based on the balanced sample is then "calibrated" on those $X_i$. That is to say, the expansion estimate of the total $X$ is necessarily without error; it is exactly equal to $X$. It is easy to see that in the situation described in the previous subsection, $\hat{Y}_R$ was similarly "calibrated" on the $X_i$, that is, $\hat{X}_R$ would have been exactly equal to $X$.

It is a matter of some contention as to whether it is preferable to use simple random sampling and the ratio estimator or simple balanced sampling and the expansion estimator. The choice is basically between a simple selection procedure and a relatively complex estimator on the one hand and a simple estimator with a relatively complex selection procedure on the other. The choice is considered at length in Brewer (1999a). It depends crucially on the prior choice of sampling inference. Those who hold exclusively to randomization for this purpose would necessarily prefer the ratio estimation option. It is only those who are prepared to accept prediction inference, either as an alternative or exclusively, for whom the choice between the two strategies described above would be a matter of taste.

For a further discussion of balanced sampling, see Sections 2.3 and 2.4.

### *3.5. Stratified random sampling with equal inclusion probabilities within strata*

If any kind of supplementary information is available that enables population units to be grouped together in such a way that they are reasonably similar within their groups and reasonably different from group to group, it will usually pay to treat these groups as separate subpopulations, or *strata*, and obtain estimates from each stratum separately. Examples of such groups include males and females, different descriptions of retail outlets (grocers, butchers, other food and drink, clothing, footwear, hardware, etc.), industries of nonretail businesses, dwellings in urban and in rural areas, or in metropolitan and nonmetropolitan areas.

It takes a great deal of similarity to obtain a poorer estimate by stratification, and the resulting increase in variance is almost always trivial, so the default rule is "Use all the relevant information that you have. When in doubt, stratify." There are, however, several exceptions to this rule.

The first is that if there are many such groups, and all the differences between all possible pairs of groups are known to be small, there is little to gain by stratification, and the business of dealing with lots of little strata might itself amount to an appreciable increase in effort. However, this is an extreme situation, so in most cases, it is safer to stick with the default rule. (In any case, do not worry. Experience will gradually give you the feel as to when to stratify and when not to do so.)

The remaining exceptions all relate to stratification by size. Size is an awkward criterion to stratify on because the boundaries between size strata are so obviously arbitrary. If stratification by size has already been decided upon, one useful rule of thumb is that size boundaries such as "under 10,000," "10,000–19,999," "20,000–49,999," "50,000–99,999," "100,000–199,999," and "over 200,000" (with appropriate adjustments to take account of the scale in which the units are measured) are difficult to improve on appreciably. Moreover, there is unlikely to be much gain in forming more than about six size strata.

Another useful rule of thumb is that each stratum should be of about the same order of magnitude in its total measure of size. This rule can be particularly helpful in choosing the boundary between the lowest two and that between the highest two strata. Dalenius (1957) does give formulae that enable optimum boundaries between size strata to be determined, but they are not recommended for general use, partly because they are complicated to apply and partly because rules of thumb and common sense will get sufficiently close to a very flat optimum. A more modern approach may be found in Lavallée and Hidiroglou (1988).

Finally, there is one situation where it might very well pay not to stratify by size at all, and that is where PRN sampling is being used. This situation will be seen later (in Section 3.9).

#### *3.5.1. Neyman and optimal allocations of sample units to strata*

Another important feature of stratification is that once the strata themselves have been defined, there are some simple rules for allocating the sample size efficiently among them. One is "Neyman allocation," which is another piece of sampling methodology recommended by Neyman in his famous 1934 paper that has already been mentioned several times. The other, usually known as "Optimum allocation," is similar to Neyman allocation but also allows for the possibility that the cost of observing the value of a sample unit can differ from stratum to stratum.

Neyman allocation minimizes the variance of a sample estimate subject to a given total sample size.[3] Basically, the allocation of sample units to a stratum $h$ should be proportional to $N_h S_h$, where $N_h$ is the number of population units in the $h$th stratum and $S_h$ is the relevant population standard deviation in that stratum.[4]

Optimum allocation is not very different. It minimizes the variance of a sample estimate subject to a given total cost and consequently allocates units in a stratum to sample proportionally to $N_h S_h/\sqrt{C_h}$, where $C_h$ is the cost of obtaining the value $Y_i$ for a single sample unit in the $h$th stratum. Since, however, it is typically more difficult to gather data from small businesses than from large ones, the effect of using Optimal rather than Neyman allocation for business surveys is to concentrate the sample toward the larger units.

Strangely, Optimum allocation seems seldom to have been used in survey practice. This is partly, perhaps, because it complicates the sample design, partly because (for any given level of accuracy) it results in the selection of a larger sample, and partly because it is not often known how much more expensive it is to collect data from smaller businesses.

### *3.5.2. Stratification with ratio estimation*

Since the effect of stratification is effectively to divide the population into a number of subpopulations, each of which can be sampled from and estimated for separately, it is theoretically possible to choose a different selection procedure and a different estimator for each stratum. However, the arguments for using a particular selection procedure and a particular estimator are usually much the same for each stratum, so this complication seldom arises.

A more important question that does frequently arise is whether or not there is any point in combining strata for estimation purposes. This leads to the distinction between "stratum-by-stratum estimation" (also known as "separate stratum estimation") and "across-stratum estimation" (also known as "combined stratum estimation"), which will be the principal topic of this subsection.

The more straightforward of these two options is stratum-by-stratum estimation, in which each stratum is regarded as a separate subpopulation, to which the observations in other strata are irrelevant. The problem with this approach, however, is that in the randomization approach the ratio estimator is biased, and the importance of that bias, relative to the corresponding standard error, can be large when the sample size is small. It is customary in some statistical offices to set a minimum (say six) to the sample size for any stratum, but even for samples of six, it is possible for the randomization bias

[3] We are indebted to Gad Nathan for his discovery that Tschuprow (or Chuprov) had actually published the same result in 1923, but his result was buried in a heap of less useful mathematics. Also, it was Neyman who brought it into prominence, and he would presumably have devised it independently of Tschuprow in any case.

[4] A special allowance has then to be made for those population units that need to be completely enumerated, and the question as to what is the relevant population standard deviation cannot be answered fully at this point, but readers already familiar with the basics of stratification are referred forward to Subsection 3.5.2.

to be appreciable, so the assumption is made that the estimation of the parameter $\beta$ in Eq. (1a) should be carried out over all size strata combined. That is to say, the value of $\beta$ is estimated as the ratio of the sum over the strata of the expansion estimates of the survey variable $y$ to the sum over the strata of the expansion estimates of the auxiliary variable $x$. This is termed the across-stratum ratio estimator of $\beta$, and the product of this with the known sum over all sampled size strata of the auxiliary variable $X$ is termed the across-stratum estimator of the total $Y$ of the survey variable $y$.

This across-stratum ratio estimator, being based on a larger effective sample size than that of any individual stratum, has a smaller randomization bias than the stratum-by-stratum ratio estimator, but because the ratio of $y$ to $x$ is being estimated over all size strata instead of separately for each, there is the strong probability that the randomization variance of the across-stratum ratio estimator will be greater than that of the stratum-by-stratum ratio estimator. Certainly, the estimators of variance yield larger estimates for the former than the latter. So there is a trade-off between unestimated (but undoubtedly real) randomization bias, and estimated randomization variance.

When looked at from the prediction approach, however, the conclusion is quite different. If the prediction models used for the individual size strata have different parameters $\beta_h$, say, where $h$ is a stratum indicator, then it is the across-stratum ratio estimator that is now biased (since it is estimating a nonexistent common parameter $\beta$) while the stratum-by-stratum ratio estimator (since it relies on small sample sizes for each) may have the larger prediction variance. If however, the prediction models for the different size strata have the same parameter $\beta$ in common, the stratum-by-stratum ratio estimator is manifestly imprecise, since it is not using all the relevant data for its inferences, and even the across-stratum ratio estimator, while prediction-unbiased, is not using the prediction-optimal weights to estimate the common parameter $\beta$.

It therefore appears that looked at from either approach, the choice between these two estimators is suboptimal, and if viewed from both approaches simultaneously, it would usually appear to be inconclusive. The underlying fact is that stratification by size is at best a suboptimal solution to the need for probabilities of inclusion in sample to increase with the size of the population unit. We shall see later (Section 3.9) that a more logical approach would be to avoid using size as an axis of stratification entirely and to use unequal probabilities of inclusion in sample instead. While this does involve certain complications, they are nothing that high-speed computers cannot cope with, whereas the complications brought about by frequent transitions from one size stratum to another within the framework of PRN sampling are distinctly less tractable.

### *3.6. Sampling with probabilities proportional to size with replacement*

As we have just seen, there are now serious arguments for using Unequal Probability Sampling within the context of surveys (chiefly establishment surveys) for which the norm has long been stratification by size and equal inclusion probabilities within strata. However, the genesis of unequal probability sampling, dating from Hansen and Hurwitz (1943), occurred in the very different context of area sampling for household surveys. The objective of Hansen and Hurwitz was to establish a master sample for the conduct of household surveys within the continental United States. It was unreasonable to contemplate the construction of a framework that included every household in the United States.[5]

Because of this difficulty, Hansen and Hurwitz instead constructed a multistaged framework. They started by dividing the United States into geographical strata, each containing roughly the same number of households. Within each stratum, each household was to have the same probability of inclusion in sample and to make this possible the selection was carried out in stages. The first stage of selection was of Primary Sampling Units (PSUs), which were relatively large geographical and administrative areas. These were sometimes counties, sometimes amalgamations of small counties, and sometimes major portions of large counties.

The important fact was that it was relatively easy to make a complete list of the PSUs within each stratum. However, it was not easy to construct a complete list of PSUs that were of more or less equal size in terms of numbers of households (or dwellings or individuals, whatever was the most accessible measure of size). Some were appreciably larger than others, but the intention remained that in the final sample, each household in the stratum would have the same probability of inclusion as every other household. So Hansen and Hurwitz decided that they would assign each PSU in a given stratum a measure of size; that the sum of those measures of size would be the product of the sample interval (or "spacing interval" or "skip interval") $i$ and the number of PSUs to be selected from that stratum, say $n$, which number was to be chosen beforehand. Then, a random number $r$ would be chosen between one and the sample interval, and the PSUs selected would be those containing the size measures numbered $r, r + i, r + 2i \dots r + (n - 1)i$ (see Table 1).

Clearly, the larger the size of a PSU, the larger would be its probability of inclusion in sample. To ensure that the larger probability of selection at the first stage did not translate into a larger probability of inclusion of households at the final stage, Hansen and Hurwitz then required that the product of the probabilities of inclusion at all subsequent stages was to be inversely proportional to the probability of selection at the first stage. So at the final stage of selection (Hansen and Hurwitz contemplated up to three such stages), the population units were individual households and each had the same eventual probability of inclusion in sample as every other household in the stratum.

To ease the estimation of variance, both overall and at each stage, Hansen and Hurwitz allowed it to proceed as though selection had been with replacement at each stage. Since the inclusion probabilities, even at each stage, were comparatively small, this was a reasonable approximation. One of the great simplifications was that the overall variance, the components from all stages combined, could be estimated as though there had been only a single stage of selection. Before the introduction of computers, this was a brilliant simplification, and even today the exact estimation of variance when sampling is without replacement still involves certain complications, considered in Section 3.7.

[5] Conceptually, it might be easier to think of this as a list of every dwelling. In fact, the two would have been identical since the definition of a dwelling was whatever a household was occupying, which might for instance be a share of a private house. A household in turn was defined as a group of people sharing meals on a regular basis.

Table 1
Example of PSU selection with randomized listing

| Sample fraction 1/147 | | Number of sample PSUs 2 | | Cluster size 32.8 | |
|---|---|---|---|---|---|
| PSU No. | No. of Dwellings | No. of Clusters | Cumulated Clusters | Selection Number | Within-PSU Sample Fraction |
| 1 | 1550 | 47 | 47 | | |
| 10 | 639 | 20 | 67 | | |
| 7 | 728 | 22 | 89 | | |
| 5 | 1055 | 32 | 121 | 103 | 1/32 |
| 9 | 732 | 22 | 143 | | |
| 2 | 911 | 28 | 171 | | |
| 6 | 553 | 17 | 188 | | |
| 3 | 1153 | 35 | 223 | | |
| 4 | 1457 | 44 | 267 | 250 | 1/44 |
| 8 | 873 | 27 | 294 | | |
| Total | 9651 | 294 | | | |

*Note*: The number of clusters in PSU number 10 has been rounded up from 19.48 to 20 in order for the total number of clusters to be divisible by 147. Note also that the selection number 103 lies in the interval between 90 and 121 while the selection number 250 lies in the interval between 224 and 267.

### *3.7. Sampling with unequal probabilities without replacement*

The transition from sampling with replacement to sampling without replacement was reasonably simple for simple random sampling but that was far from the case for sampling with unequal probabilities. The first into the field were Horvitz and Thompson (1952). Their estimator is appropriately named after them as the Horvitz-Thompson Estimator or HTE. It is simply the sum over the sample of the ratios of each unit's survey variable value ($y_i$ for the $i$th unit) to its probability of inclusion in sample ($\pi_i$). The authors showed that this estimator was randomization unbiased. They also produced a formula for its variance and a (usually unbiased) estimator of that variance. These last two formulae were functions of the "second-order inclusion probabilities," that is, the probabilities of inclusion in sample of all possible pairs of population units. If the number of units in the population is denoted by $N$, then the number of possible pairs is $N(N-1)/2$, so the variance formula involved a summation over $N(N-1)/2$ terms, and even the variance estimation formula required a sum over $n(n-1)/2$ pairs of sample units.

Papers by Sen (1953) and by Yates and Grundy (1953) soon followed. Both of these made use of the fact that when the selection procedure ensured a sample of predetermined size ($n$ units), the variance was both minimized in itself and capable of being estimated much more accurately than when the sample size was not fixed. Both papers arrived at the same formulae for the fixed-sample-size variance and for an estimator of that variance that was randomization unbiased, provided that the joint inclusion probabilities, $\pi_{ij}$, for all possible pairs of units were greater than zero. However, this Sen–Yates–Grundy variance estimator still depended on the $n(n-1)/2$ values of the $\pi_{ij}$ so that the variance could not be estimated randomization-unbiasedly without evaluating this large number of joint inclusion probabilities.

Many without-replacement selection schemes have been devised in attempts to minimize these problems. One of the earliest and simplest was randomized systematic sampling, or "RANSYS," originally described by Goodman and Kish (1950). It involved randomizing the population units and selecting systematically with a skip interval that was constant in terms of the size measures. After 1953, dozens of other methods followed in rapid succession. For descriptions of these early methods, see Brewer and Hanif (1982) and Chaudhury and Vos (1988). However, it seemed to be generally true that if the sample was easy to select, then the inclusion probabilities were difficult to evaluate, and the converse also holds.

Poisson sampling (Hájek, 1964) is one such method that deserves a special mention. Although in its original specification, it did not ensure samples of fixed size, it did have other interesting properties. To select a Poisson sample, each population in turn is subjected to a Bernoulli trial, with the probability of "success" (inclusion in sample) being $\pi_i$, and the selection procedure continues until the last population unit has been subjected to its trial. The achieved sample sizes are, however, highly variable, and consequently, Poisson sampling in its original form was not an immediately popular choice. However, several modified versions were later formulated; several of these and also the original version are still in current use.

One of the most important of these modified versions was Conditional Poisson Sampling or CPS, also found in Hájek (1964) and discussed in detail by Chen et al. (1994). For CPS, Poisson samples with a particular expected sample size are repeatedly selected, but only to be immediately rejected once it is certain that the eventual sample will not have exactly that expected sample size. One notable feature of CPS is that it has the maximum entropy attainable for any population of units having a given set of first-order inclusion probabilities $\pi_i$.[6] Several fast algorithms for using CPS are now available, in which the second-order inclusion probabilities are also computed exactly. See Tillé (2006).

In the meantime, however, another path of investigation had also been pioneered by Hájek (1964). He was concerned that the estimation of variance for the HTE was unduly complicated by the fact that both the Sen–Yates–Grundy formula for the randomization variance and their estimator of that variance required knowledge of the second-order inclusion probabilities. In this instance, Hájek (and eventually others) approximated the fixed sample size variance of the HTE by an expression that depended only on the first-order inclusion probabilities. However, initially these approximations were taken to be specific to particular selection procedures. For instance, Hájek's 1964 approximation was originally taken to be specific to CPS.

In time, however, it was noted that very different selection procedures could have almost identical values of the $\pi_{ij}$. The first two for which this was noticed were RANSYS, for which the $\pi_{ij}$ had been approximated by Hartley and Rao (1962), and the Rao–Sampford selection procedure (J.N.K. Rao, 1965; Sampford, 1967), for which

[6] Entropy is a measure of unpredictability or randomness. If a population is deliberately arranged in order of size and a sample is selected from it systematically, that sample will have low entropy. If however (as with RANSYS) the units are arranged in random order before selection, the sample will have high entropy, only a few percentage points smaller than that of CPS itself. While low entropy sample designs may have very high or very low randomization variances, high entropy designs with the same set of first-order inclusion probabilities all have more or less the same randomization variance. For a discussion of the role of entropy in survey sampling, see Chen et al. (1994).

they had been approximated by Asok and Sukhatme (1976). These were radically different selection procedures, but the two sets of approximations to the $\pi_{ij}$ were identical to order $n^3/N^3$. Although both procedures produced fixed size samples, and the population units had inclusion probabilities that were exactly proportional to their given measures of size, it appeared that the only other thing that the two selection procedures had in common was that they both involved a large measure of randomization. Entropy, defined as $\sum_{k=1}^{M}[P_k - log(P_k)]$, where $P_k$ is the probability of selecting the $k$th out of the $M$ possible samples, is a measure of the randomness of the selection. It therefore appeared plausible that all high-entropy sampling procedures would have much the same sets of $\pi_{ij}$, and hence much the same randomization variance. If so, it followed that approximate variance formulae produced on the basis of any of these methods would be valid approximations for them all, and that useful estimators of these approximate variances would be likely also to be useful estimators of the variances of the HTE for all high-entropy selection procedures.

Whether this is the case or not is currently a matter of some contention, but Preston and Henderson (2007) provide evidence to the effect that the several randomization variance estimators provided along these lines are all reasonably similar in precision and smallness of bias, all at least as efficient as the Sen–Yates–Grundy variance estimator (as measured by their randomization mean squared errors MSEs), and all a great deal less cumbersome to use.

In addition, they can be divided into two families, the members of each family having both a noticeable similarity in structure and a detectable difference in entropy level from the members of the other family. The first family includes those estimators provided by Hájek (1964), by Deville (1993, 1999, 2000; see also Chen et al., 1994) initially for CPS, and by Rosén for Pareto $\pi$ps (Rosén, 1997a, 1997b). The second family, described in Brewer and Donadio (2003), is based on the $\pi_{ij}$ values associated with RANSYS and with the Rao–Sampford selection procedure. These two procedures have slightly smaller entropies and slightly higher randomization variance than CPS, but both Preston and Henderson (2007) and Henderson (2006) indicate that the Hájek-Deville family of estimators should be used for CPS, Pareto $\pi$ps and similar selection procedures—thus probably including Tillé (1996)—while the Brewer-Donadio family estimators would be appropriate for use with RANSYS and Rao-Sampford.

It is also possible to use replication methods, such as the jackknife and the bootstrap, to estimate the HTE's randomization variance. The same Preston and Henderson paper provides evidence that a particular version of the bootstrap can provide adequate, though somewhat less accurate, estimates of that variance than can be obtained using the two families just described.

Finally, it is of interest that the "anticipated variance" of the HTE (that is to say the randomization expectation of its prediction variance, or equivalently the prediction expectation of its randomization variance; see Isaki and Fuller, 1982) is a simple function of the $\pi_i$ and independent of the $\pi_{ij}$. Hence, for any population that obeys the model of Eqns. (1), both the randomization variance and the anticipated variance of the HTE can be estimated without any reference to the $\pi_{ij}$.

### 3.8. *The generalized regression estimator*

Up to this point, it has been assumed that only a single auxiliary variable has been available for improving the estimation of the mean or total of a survey variable. It has also been assumed that the appropriate way to use that auxiliary variable was by using Eq. (1a), which implies a ratio relationship between those two variables. More generally, the survey variable could depend on a constant term as well, or on more than a single auxiliary variable, or both. However, that relationship is seldom likely to be well represented by a model that implies the relevance of ordinary least squares (OLS).

One case where OLS might be appropriate is where the survey variable is Expenditure and the auxiliary variable is Income. The relationship between Income and Expenditure (the Consumption Function) is well known to involve an approximately linear dependence with a large positive intercept on the Expenditure axis. But OLS assumes homoskedasticity (the variance of Expenditure remaining constant as Income increases) while it is more than likely that the variance of Expenditure increases with Income, and in fact the data from the majority of sample surveys do indicate the existence of a measure of heteroskedaticity. This in itself is enough to make the use of OLS questionable. Eq. (1c) allows for the variance of the survey variable to increase linearly with the auxiliary variable, and in fact it is common for this variance to increase somewhat faster than this, and occasionally as fast as the square of the auxiliary variable.

A commonly used estimator of total in these more general circumstances is the generalized regression estimator or GREG (Cassel et al., 1976), which may be written as follows:

$$\hat{Y}_{\mathrm{GREG}} = \hat{Y}_{\mathrm{HTE}} + \sum_{k=1}^{p} (X_k - \hat{X}_{\mathrm{HTE}k})\hat{\beta}_k, \tag{8}$$

or alternatively as

$$\hat{Y}_{\mathrm{GREG}} = \sum_{k=1}^{p} X_k\hat{\beta}_k + \left(\hat{Y}_{\mathrm{HTE}} - \sum_{k=1}^{p} \hat{X}_{\mathrm{HTE}k}\hat{\beta}_k\right). \tag{9}$$

In these two equations, $\hat{Y}_{\mathrm{HTE}}$ is the HTE of the survey variable, $\hat{X}_{\mathrm{HTE}k}$ is the HTE of the $k$th auxiliary variable and $\hat{\beta}_k$ is an estimator of the regression coefficient of the survey variable on the $k$th auxiliary variable, where the regression is on $p$ auxiliary variables simultaneously. One of those auxiliary variables may be a constant term, in which case there is an intercept estimated in the equation. (In that original paper, $\hat{\beta}_k$ was a generalized least squares estimator, but this was not a necessary choice. For instance, Brewer (1999b) defined $\hat{\beta}_k$ in such a way as to ensure that the GREG was simultaneously interpretable in the randomization and prediction approaches to sampling inference, and also showed that this could be achieved with only trivial increments to its randomization and prediction variances).

In the second of these two equations, the first term on the right-hand side is a prediction estimator of the survey variable total, but one that ignores the extent to which the HTE of the survey variable total differs from the sum of the $p$ products of the individual auxiliary variable HTEs with their corresponding regression estimates. Särndal et al. (1992) noted that the first term (the prediction estimator) had a randomization variance that was of a lower order of magnitude than the corresponding variance of the second term and therefore suggested that the randomization variance of the GREG estimator be estimated by estimating only that of the second term. It is true that as the sample size increases, the randomization variance of the prediction estimator becomes small with respect to that of the second term, but when the sample size is small, this can lead to a substantial underestimate of the GREG's randomization variance.

This is not an easy problem to solve wholly within the randomization approach, and in Chapter 8 of Brewer (2002, p. 136), there is a recommendation to estimate the anticipated variance as a substitute. (The anticipated variance is the randomization expectation of the prediction variance.). This is obviously not a fully satisfactory solution, except in the special case considered by Brewer, where the GREG had been devised to be simultaneously a randomization estimator and a prediction estimator, so more work on it seems to be called for. Another alternative would be to estimate the GREG's randomization variance using a replication method such as the jackknife or the bootstrap, but again this alternative appears to need further study. For more information regarding the GREG, see Särndal et al. (1992).

### *3.9. Permanent random number (PRN) sampling*

One of the important but less obvious objectives of survey sampling is to be able to control intelligently the manner in which the sample for a repeating survey is allowed to change over time. It is appropriate for a large sample unit that is contributing substantially to the estimate of total to remain in sample for fairly long periods, but it is not so appropriate for small population units to do the same, so it is sensible to rotate the sample around the population in such a way that the larger the unit is, the longer it remains in sample. One of the ways of doing this is to assign each unit a PRN, say between zero and unity, and define the sample as consisting of those population units that occupy certain regions of that PRN space. Units in a large-size stratum might initially be in sample if they had PRNs between zero and 0.2 for the initial survey, between 0.02 and 0.22 for the second, 0.04 and 0.24 for the third, and so on. In this way, each unit would remain in sample for up to 10 occasions but then be "rested" for the next 40. Those in a small-size stratum would remain occupy a smaller region of the PRN space, say initially between zero and 0.04, but the sample PRN space would be rotated just as fast so that units would remain in sample for no more than two occasions before being "rested."

From the data supplier's point of view, however, it is particularly inappropriate to be removed from the sample and then included again shortly afterwards. This can easily happen, however, if a population unit changes its size stratum, particularly if the change is upward. Consequently, it is inconvenient to use PRN sampling and size stratification together. Moreover, as has already been indicated in Section 3.5, stratification by size is a suboptimal way of satisfying the requirement that the larger the unit, the greater should be its probability of inclusion in sample.

Hence, when attempting to control and rotate samples using the PRN technique, it becomes highly desirable, if not indeed necessary, to find a better solution than stratification by size. Brewer (2002) (Chapter 13, pp. 260–265), provides a suggestion as to how this could be done. It involves the use of a selection procedure known as Pareto $\pi$ps sampling, which is due to Rosén (1997a, 1997b). This is a particular form of what is known as *order sampling*, and is very similar in its $\pi_{ij}$ values to CPS sampling, so it is a high-entropy sample selection procedure. It is, however, somewhat complicated to describe and therefore inappropriate to pursue further in this introductory chapter. Those who wish to pursue the possibility of using PRN sampling without stratification by size are referred to those two papers by Rosén and to Chapter 13 of Brewer (2002).

## 4. Conclusion

From the very early days of survey sampling, there have been sharp disagreements as to the relative importance of the randomization and prediction approaches to survey sampling inference. These disagreements are less severe now than they were in the 1970s and 1980s but to some extent they have persisted into the 21st century. What is incontrovertible, however, is that prediction inference is parametric and randomization nonparametric. Hence the prediction approach is appropriate to the extent that the prediction models are useful, whereas the randomization approach provides a robust alternative where they are not useful. It would therefore seem that ideally both should be used together, but there are many who sincerely believe the one or the other to be irrelevant. The dialogue therefore continues.

Both the randomization and the prediction approaches offer a wide range of manners in which the sample can or should be selected, and an equally wide range of manners in which the survey values (usually, but not exclusively consisting of population totals, population means, and ratios between them) can be estimated. The choices among them depend to a large extent on the natures of the populations (in particular, whether they consist of individuals and households, of establishments and enterprizes, or of some other units entirely) but also on the experience and the views of the survey investigators. However, there are some questions that frequently need to be asked, and these are the ones that have been focussed on in this chapter. They include, "What are the units that constitute the population?" "Into what groups or strata do they naturally fall?" "What characteristics of the population need to be estimated?" "How large a sample is appropriate?" (or alternatively, "How precise are the estimates required to be?") "How should the sample units be selected?" and "How should the population characteristics be estimated?"

In addition, there are many questions that need to be answered that fall outside the scope of the discipline of survey sampling. A few of them would be as follows: "What information are we seeking, and for what reasons?" "What authority, if any, do we have to ask for this information?" "In what format should it be collected?" "What organizational structure is required?" "What training needs to be given and to whom?" and not least, "How will it all be paid for?"

So those questions that specifically relate to survey sampling always need to be considered in this wider framework. The aim of this Chapter will have been achieved if the person who has read it has emerged with some feeling for the way in which the discipline of survey sampling can be used to fit within this wider framework.

*Sample Surveys: Design, Methods and Applications, Vol. 29A*
ISSN: 0169-7161
© 2009 Elsevier B.V. All rights reserved
DOI: 10.1016/S0169-7161(08)00002-3

2

# Sampling with Unequal Probabilities

*Yves G. Berger and Yves Tillé*

## 1. Introduction

Since the mid 1950s, there has been a well-developed theory of sample survey design inference embracing complex designs with stratification and unequal probabilities (Smith, 2001). Unequal probability sampling was first suggested by Hansen and Hurwitz (1943) in the context of sampling with replacement. Narain (1951), Horvitz and Thompson (1952) developed the corresponding theory for sampling without replacement. A large part of survey sampling literature is devoted to unequal probabilities sampling, and more than 50 sampling algorithms have been proposed. Two books (Brewer and Hanif, 1983; Tillé, 2006) provide a summary of these methods.

Consider a finite population $U$ of size $N$. Each unit of the population can be identified by a label $k = 1, \ldots, N$. A sample $s$ is a subset of $U$. A sampling design $p(.)$ is a probability measure on all the possible samples so that

$$p(s) \geq 0, \text{ for all } s \in U, \text{ and } \sum_{s \in U} p(s) = 1.$$

Let $n(s)$ denote the size of the sample $s$. When the sample size is not random, we denote the sample size by $n$. An unequal probability sampling design is often characterized by its first-order inclusion probabilities given by $\pi_k = p(k \in s)$. The joint inclusion probabilities of unit $k$ and $\ell$ are defined by $\pi_{k\ell} = p(k \in s \text{ and } \ell \in s)$.

Suppose we wish to estimate the population total

$$Y = \sum_{k \in U} y_k$$

of a characteristic of interest $y$, where $y_k$ is the value of a unit labeled $k$. An estimator of $Y$ is given by the $\pi$-estimator (Horvitz and Thompson, 1952; Narain, 1951) defined by

$$\widehat{Y}_\pi = \sum_{k \in s} \frac{y_k}{\pi_k}.$$

This estimator is design unbiased provided that all the $\pi_k > 0$.

Under unequal probability sampling, the variance of $\widehat{Y}_\pi$ may be considerably smaller than the variance under an equal probability sampling design (Cochran, 1963), when the correlation between the characteristic of interest and the first-order inclusion probabilities is strong. Alternative estimators when this correlation is weak are discussed in Section 3.

It is common practice to use inclusion probabilities that are proportional to a known positive size variable $x$. In this case, the inclusion probabilities are computed as follows

$$\pi_k = \frac{nx_k}{X}, \tag{1}$$

where $X = \sum_{k\in U} x_k$, assuming $nx_k \le X$ for all $k$. If $nx_k > X$, we set $\pi_k = 1$ and we recalculate the $\pi_k$ using (1) on the remaining units after substituting $n$ with $n$ subtracted by the number of $\pi_k$ equal to 1.

Another application of unequal probability sampling design is with multistage sampling, where the selection of primary units within strata may be done with unequal probability. For example, self-weighted two-stage sampling is often used to select primary sampling units with probabilities that are proportional to the number of secondary sampling units within the primary units. A simple random sample is selected within each primary unit.

The variance of the $\pi$-estimator plays an important role in variance estimation, as most estimators of interest can be linearized to involve $\pi$-estimators (see Section 5). The sampling variance of $\widehat{Y}_\pi$ is given by

$$\operatorname{var}(\widehat{Y}_\pi) = \sum_{k\in U}\sum_{\ell\in U}(\pi_{k\ell} - \pi_k\pi_\ell)\frac{y_k y_\ell}{\pi_k\pi_\ell}.$$

Horvitz and Thompson (1952) proposed an unbiased estimator of $\operatorname{var}(\widehat{Y}_\pi)$:

$$\operatorname{var}(\widehat{Y}_\pi) = \sum_{k\in s}\sum_{\ell\in s}\frac{\pi_{k\ell} - \pi_k\pi_\ell}{\pi_{k\ell}}\frac{y_k y_\ell}{\pi_k\pi_\ell}. \tag{2}$$

If the sample size is fixed, Sen (1953), Yates and Grundy (1953) proposed another estimator of $\operatorname{var}(\widehat{Y}_\pi)$:

$$\widehat{\operatorname{var}}(\widehat{Y}_\pi) = \frac{1}{2}\sum_{k\in s}\sum_{\ell\in s}\frac{\pi_k\pi_\ell - \pi_{k\ell}}{\pi_{k\ell}}\left(\frac{y_k}{\pi_k} - \frac{y_\ell}{\pi_\ell}\right)^2. \tag{3}$$

This estimator is design unbiased when $\pi_{k\ell} > 0$ for all $k, \ell \in U$. It can take negative values unless $\pi_k\pi_\ell - \pi_{k\ell} \ge 0, k \ne \ell \in U$. However, it is rarely used because the joint inclusion probabilities are sometimes difficult to compute and because the double sum makes (3) computationally intensive. In Section 4, we show that, in particular cases, the variance can be estimated without joint inclusion probabilities.

## 2. Some methods of unequal probability sampling

### 2.1. *Poisson sampling*

Poisson sampling was proposed by Hájek (1964) and discussed among others in Ogus and Clark (1971), Brewer et al. (1972, 1984), and Cassel et al. (1993a, p. 17). Each unit of the population is selected independently with a probability $\pi_k$. The sample size $n(s)$ is therefore random. All the samples $s \subset U$ have a positive probability of being selected and there is a non-null probability of selecting an empty sample. The sampling design is given by

$$P(s) = \left[\prod_{k\in s} \frac{\pi_k}{1-\pi_k}\right]\left[\prod_{k\in U}(1-\pi_k)\right], \text{ for all } s \subset U.$$

Since the units are selected independently, we have that $\pi_{k\ell} = \pi_k\pi_\ell$, for all $k \neq \ell$.

The variance of the $\pi$-estimator, given in (2), reduces to

$$\operatorname{var}\left(\widehat{Y}_\pi\right) = \sum_{k\in U} \frac{1}{\pi_k}(1-\pi_k)y_k^2,$$

which can be unbiasedly estimated by

$$\widehat{\operatorname{var}}\left(\widehat{Y}_\pi\right) = \sum_{k\in s}(1-\pi_k)\frac{y_k^2}{\pi_k^2}.$$

The estimator of variance is simple because it does not involve joint inclusion probabilities. Note that the Poisson sampling design maximizes the entropy (Hájek, 1981, p.29) given by

$$I(p) = -\sum_{s\subset U} p(s)\log p(s), \tag{4}$$

subject to given inclusion probabilities $\pi_k, k \in U$. Since the entropy is a measure of randomness, the Poisson sampling design can be viewed as the most random sampling design that satisfies given inclusion probabilities.

Poisson sampling is rarely applied in practice because its sample size is random implying a nonfixed cost of sampling. This design is, however, often used to model nonresponse. Moreover, Poisson sampling will be used in Section 2.7 to define the conditional Poisson sampling design which is also called the maximum entropy design with fixed sample size. The use of design that maximizes the entropy is useful because it allows a simple estimation for the variance.

### 2.2. *Sampling with replacement*

Unequal probability sampling with replacement is originally due to Hanssen and Hurwitz. Properties of this design are widely covered in the literature (Bol'shev, 1965; Brown and Bromberg, 1984; Dagpunar, 1988; Davis, 1993; Devroye, 1986; Ho et al., 1979; Johnson et al., 1997; Kemp and Kemp, 1987; Loukas and Kemp, 1983; Tillé, 2006).

Consider selection probabilities $p_k$ that are proportional to a positive size variable $x_k, k \in U$; that is,

$$p_k = \frac{x_k}{\sum_{\ell\in U} x_\ell}, k \in U.$$

A simple method to select a sample with unequal probabilities with replacement consists in generating a uniform random number $u$ in [0, 1) and selecting unit $k$ so that

$v_{k-1} \leq u < v_k$, where

$$v_k = \sum_{\ell=1}^{k} p_\ell, \text{ with } v_0 = 0.$$

This process is repeated independently $m$ times. Note that there are more efficient algorithms that may be used to select a sample with replacement with unequal probabilities (Tillé, 2006, p. 75).

Let $\tilde{y}_i$ denote the value of the characteristic of the $i$th selected unit and $\tilde{p}_i$, its associated selection probability. Note that, under sampling with replacement, the same unit can be selected several times. The ratios $\tilde{y}_i/\tilde{p}_i$ are $n$ independent random variables. The total $Y$ can be estimated by the Hansen–Hurwitz estimator

$$\widehat{Y}_{\text{HH}} = \frac{1}{m}\sum_{i=1}^{m}\frac{\tilde{y}_i}{\tilde{p}_i}.$$

This estimator is design unbiased as

$$\mathrm{E}\left(\widehat{Y}_{\text{HH}}\right) = \frac{1}{m}\sum_{i=1}^{m}\mathrm{E}\left(\frac{\tilde{y}_i}{\tilde{p}_i}\right) = \frac{1}{m}\sum_{i=1}^{m} Y = Y.$$

The variance of $\widehat{Y}_{\text{HH}}$ is given by

$$\operatorname{var}(\widehat{Y}_{\text{HH}}) = \frac{1}{m}\sum_{k\in U} p_k \left(\frac{y_k}{p_k} - Y\right)^2,$$

which can be unbiasedly estimated by

$$\widehat{\operatorname{var}}(\widehat{Y}_{\text{HH}}) = \frac{1}{m(m-1)}\sum_{i=1}^{m}\left(\frac{\tilde{y}_i}{\tilde{p}_i} - \widehat{Y}_{\text{HH}}\right)^2. \quad (5)$$

The Hansen–Hurwitz estimator is not the best estimator as it is not admissible because it depends on the multiplicity of the units (Basu, 1958, 1969; Basu and Ghosh, 1967). Nevertheless, the Hansen–Hurwitz variance estimator can be used to approximate the variance of the Horvitz–Thompson estimator under sampling without replacement when $m/N$ is small.

Sampling without replacement may lead to a reduction of the variance compared to sampling with replacement (Gabler, 1981, 1984). A design without replacement with inclusion probabilities $\pi_k$ is considered to be a good design if the Horvitz–Thompson estimator is always more accurate than the Hansen–Hurwitz estimator under sampling with replacement with probabilities $p_k = \pi_k/n$. Gabler (1981, 1984) gave a condition under which this condition holds. For example, this condition holds for the Rao–Sampford design given in Section 2.4 and for the maximum entropy design with fixed sample size (Qualité, 2008).

### 2.3. Systematic sampling

Systematic sampling is widely used by statistical offices due to its simplicity and efficiency (Bellhouse, 1988; Bellhouse and Rao, 1975; Berger, 2003; Iachan, 1982, 1983).

This sampling design has been studied since the early years of survey sampling (Cochran, 1946; Madow, 1949; Madow and Madow, 1944). There are two types of systematic design: a systematic sample can be selected from a deliberately ordered population or the population can be randomized before selecting a systematic sample. The latter is often called randomized systematic design.

A systematic sample is selected as follows. Let $u$ be a random number between 0 and 1 generated from a uniform distribution. A systematic sample is a set of $n$ units labeled $k_1, k_2, \ldots, k_n$ such that $\pi_{k_\ell - 1}^{(c)} < u + \ell - 1 \leq \pi_{k_\ell}^{(c)}$, where $\ell = 1, \ldots, n$ and

$$\pi_k^{(c)} = \sum_{\substack{j \in U \\ j \leq k}} \pi_j.$$

In the special case where $\pi_k = n/N$, this design reduces to the customary systematic sampling design, where every $a$th unit is selected and $a = \lfloor N/n \rfloor$.

In many practical situations, it is common practice to let the population frame have a predetermined order. For example, a population frame can be sorted by a size variable, by region, by socioeconomic group, by postal sector, or in some other way. In this case, systematic sampling is an efficient method of sampling (Iachan, 1982). Systematic sampling from a deliberately ordered population is generally more accurate than randomized systematic sampling (Särndal et al., 1992, p. 81), especially when there is a trend in the survey variable $y$ (Bellhouse and Rao, 1975).

The systematic design with a deliberately ordered population suffers from a serious flaw, namely, that it is impossible to unbiasedly estimate the sampling variance (Iachan, 1982), and customary variance estimators given in (3) are inadequate and can overestimate significantly the variance (Särndal et al., 1992, Chapter 3).

Systematic sampling from a randomly ordered population consists in randomly arranging the units, giving the same probability to each permutation, since random ordering is part of the sampling design. This design was first suggested by Madow (1949). Hartley and Rao (1962) developed the corresponding asymptotic theory for large $N$ and small sampling fraction. Under randomized systematic sampling, Hartley and Rao (1962) derived a design unbiased variance estimator (see Section 4).

For the randomized systematic design, the joint inclusion probabilities are typically positive and the variance can be unbiasedly estimated (Hájek, 1981; Hartley and Rao, 1962). With a deliberately ordered population, alternative estimators for the variance can be used (Bartolucci and Montanari, 2006; Berger, 2005a; Brewer, 2002, Chapter 9).

### 2.4. *Rao–Sampford sampling design*

The Rao–Sampford sampling design (Rao, 1965; Sampford, 1967) is a popular design used for unequal probability sampling without replacement. It is implemented by selecting the first unit with drawing probabilities $p_k = \pi_k/n$. The remaining $n - 1$ units are selected with replacement with drawing probabilities that are proportional to $\pi_k/(\pi_k - 1)$. The sample is accepted if the $n$ units drawn are all distinct, otherwise, it is rejected and the process is repeated. The first-order inclusion probabilities are exactly given by $\pi_k$. Sampford (1967) derived an exact expression for the joint inclusion probabilities $\pi_{k\ell}$.

The main advantage of this design is its simplicity. It also has a simple expression for the variance (see Section 4). However, this design is not suitable when the $\pi_k$ are large, as we would almost surely draw the units with large $\pi_k$ at least twice and it would not be possible to select one Rao–Sampford sample. For example, consider $N = 86$, $n = 36$, and $\pi_k$ proportional to $(k/100)^5 + 1/5$. The probability that all the units drawn from subsequent independent draws will be distinct is approximately $10^{-36}$ (Hájek, 1981, p. 70), which is negligible. Nevertheless, Tillé (2006, p. 136) and Bondesson et al. (2006) suggested several alternative algorithms to implement the Rao–Sampford design.

### 2.5. *Sampling by the splitting method*

The splitting method, proposed by Deville and Tillé (1998), is a general class of sampling designs without replacement with fixed sample size and unequal probabilities. First, each inclusion probability is split into two or more quantities. Secondly, one of these sets of quantities is randomly selected in such a way that the overall inclusion probabilities are equal to $\pi_k$. These steps are repeated until a sample is obtained.

This method can be implemented as follows. First, $\pi_k$ is split into two quantities $\pi_k^{(1)}$ and $\pi_k^{(2)}$, which satisfy the following constraint:

$$\pi_k = \lambda\pi_k^{(1)} + (1 - \lambda)\pi_k^{(2)},$$

with

$$0 \leq \pi_k^{(1)} \leq 1 \text{ and } 0 \leq \pi_k^{(2)} \leq 1,$$

$$\sum_{k\in U} \pi_k^{(1)} = \sum_{k\in U} \pi_k^{(2)} = n,$$

where $\lambda$ is any constant such that $0 < \lambda < 1$.

The method consists of choosing

$$\begin{cases} \pi_k^{(1)}, k \in U, & \text{with a probability } \lambda \text{ or} \\ \pi_k^{(2)}, k \in U, & \text{with a probability } 1 - \lambda. \end{cases}$$

After this first step, any design can be used to select a sample with inclusion probabilities $\pi_k^{(1)}$ or $\pi_k^{(2)}$. If some of the $\pi_k^{(1)}$ or $\pi_k^{(2)}$ are all equal to 0 or 1, we would sample from a smaller population. The splitting can in turn be used to select a sample with probabilities $\pi_k^{(1)}$ or $\pi_k^{(2)}$. We could also choose $\pi_k^{(1)}$ in such a way that the $\pi_k^{(1)}$ are all equal. In this case, simple random sampling without replacement can be used.

This approach can be generalised to a splitting method into $M$ sets of inclusion probabilities. First, we choose the $\pi_k^{(j)}$ and the $\lambda_j$ in such a way that

$$\sum_{j=1}^{M} \lambda_j = 1,$$

where

$$0 \leq \lambda_j \leq 1, \quad j = 1, \ldots, M,$$

$$\sum_{j=1}^{M} \lambda_j \pi_k^{(j)} = \pi_k,$$

$$0 \leq \pi_k^{(j)} \leq 1, \quad k \in U, j = 1, \ldots, M,$$

$$\sum_{k \in U} \pi_k^{(j)} = n, \quad j = 1, \ldots, M.$$

We then select one of the set of quantities of $\pi_k^{(j)}, k \in U$, with probabilities $\lambda_j$ , $j = 1, \ldots, M$. Secondly, any design can be used to select a sample with inclusion probabilities $\pi_k^{(j)}$ or the splitting step can be applied again.

Deville and Tillé (1998) showed that the splitting method defines new sampling designs such as the minimum support design, the splitting into simple random sampling, the pivotal method, and the eliminatory method (Tillé, 2006).

### 2.6. *Brewer sampling design*

Brewer (1963) proposed a design for selecting a sample of size $n = 2$. The properties of this design were studied by Rao and Bayless (1969), Rao and Singh (1973), Sadasivan and Sharma (1974), and Cassel et al. (1993a). Brewer (1975) generalised this design to any sample size (Brewer and Hanif, 1983, p. 26). This method is a draw by draw procedure, that is, a sample can be selected in $n$ steps. In this section, we show that this design is a particular case of the splitting method.

For simplicity, only the first step of the method is given. Consider

$$\lambda_j = \left\{ \sum_{k=1}^{N} \frac{\pi_k(n - \pi_k)}{1 - \pi_k} \right\}^{-1} \frac{\pi_j(n - \pi_j)}{1 - \pi_j}.$$

and

$$\pi_k^{(j)} = \begin{cases} \dfrac{\pi_k(n-1)}{n - \pi_j} & \text{if } k \neq j \\ 1 & \text{if } k = j. \end{cases}$$

The first-order inclusion probabilities are indeed given by $\pi_k$ because

$$\sum_{j=1}^{N} \lambda_j \pi_k^{(j)} = \pi_k.$$

At each step of the method, a unit is selected. Moreover, it is not necessary to compute all the $\pi_k^{(j)}$, as only the selected $\pi_k^{(j)}$, $k \in U$, need to be computed.

### 2.7. *Maximum entropy or conditional Poisson sampling design*

The maximum entropy design (Hájek, 1981) and the conditional Poisson design are the same design obtained from two different perspectives. The maximum entropy design is the design with fixed sample size that maximizes the entropy given in (4) for all the samples of fixed sample size $n$ subject to given inclusion probabilities $\pi_k$, $k \in U$. Hájek (1981) proposed to implement it by using a Poisson rejective procedure, that is, by reselecting Poisson samples until a fixed sample size is obtained. A rejective procedure consists in conditioning Poisson sampling design with respect to a fixed sample size.

Consider a Poisson sampling design with inclusion probabilities $\widetilde{\pi}_k$ and a random sample size $\tilde{n}$. This sampling design can be written as follows:

$$P(s) = \left[\prod_{k\in s} \frac{\widetilde{\pi}_k}{1-\widetilde{\pi}_k}\right]\left[\prod_{k\in U}(1-\widetilde{\pi}_k)\right].$$

The conditional Poisson sampling design is then given by

$$p(s) = P(s|\widetilde{n}_s = n) = \frac{P(s)}{\sum_{s\in\mathcal{S}_n} P(s)},\ s \in \mathcal{S}_n,$$

where $n$ is fixed and $\mathcal{S}_n$ is the set of all the samples of size $n$.

Conditional Poisson sampling can be implemented by using a rejective sampling procedure. Samples are selected with Poisson sampling and inclusion probability $\widetilde{\pi}_k$ until a fixed sample size $n$ is obtained. However, more efficient algorithms, such as a draw by draw procedure or a sequential procedure, are described for instance in Tillé (2006, pp. 90–95).

The main difficulty is that the inclusion probabilities $\pi_k$ of the design are different from the $\widetilde{\pi}_k$. Hájek (1964) proposed approximations for the inclusion probabilities (see also Brewer and Hanif, 1983, p. 40). Chen et al. (1994) proposed an algorithm that allows us to derive the inclusion probabilities of the conditional Poisson sampling $\pi_k$ from the inclusion probabilities of the Poisson sampling design $\widetilde{\pi}_k$. In an unpublished manuscript available from the author, Deville (2000) improved this algorithm and derived the following recursive formula:

$$\pi_k(\widetilde{\boldsymbol{\pi}}, n) = n\frac{\widetilde{\pi}_k(1-\widetilde{\pi}_k)^{-1}\left[1-\pi_k(\widetilde{\boldsymbol{\pi}}, n-1)\right]}{\sum_{\ell\in U}\widetilde{\pi}_\ell(1-\widetilde{\pi}_\ell)^{-1}\left[1-\pi_\ell(\widetilde{\boldsymbol{\pi}}, n-1)\right]},$$

where $\widetilde{\boldsymbol{\pi}}$ is the vector of inclusion probabilities $\widetilde{\pi}_k$.

This recursive equation allows us to compute $\pi_k$ from $\widetilde{\pi}_k$ easily. Deville (2000) also proposed that a modified Newton-Raphson method be used to compute the $\widetilde{\pi}_k$ from the given inclusion probability vector $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_N)$. The recursive equation is given by

$$\widetilde{\boldsymbol{\pi}}^{(i+1)} = \widetilde{\boldsymbol{\pi}}^{(i)} + \boldsymbol{\pi} - \boldsymbol{\pi}(\widetilde{\boldsymbol{\pi}}, n), \quad \text{for } i = 0, 1, 2, \ldots,$$

where $\widetilde{\boldsymbol{\pi}}^{(0)} = \boldsymbol{\pi}$.

Deville (2000) also proposed a recursive relation for computing the joint inclusion probabilities:

$$\begin{aligned}&\pi_{k\ell}(\widetilde{\boldsymbol{\pi}}, n)\\ &\quad= \frac{n(n-1)\exp\lambda_k \exp\lambda_\ell\left[1-\pi_k(\widetilde{\boldsymbol{\pi}}, n-2)-\pi_\ell(\widetilde{\boldsymbol{\pi}}, n-2)+\pi_{k\ell}(\widetilde{\boldsymbol{\pi}}, n-2)\right]}{\sum_{i\in U}\sum_{\substack{j\in U\\ i\neq j}}\exp\lambda_i\exp\lambda_j\left[1-\pi_i(\widetilde{\boldsymbol{\pi}}, n-2)-\pi_j(\boldsymbol{\lambda}, \mathcal{S}_{n-2})+\pi_{ij}(\widetilde{\boldsymbol{\pi}}, n-2)\right]},\end{aligned}$$

Additional developments on conditional Poisson sampling are given in Chen et al. (1994), Chen and Liu (1997), Chen (1998, 2000), Deville (2000), Jonasson and Nerman (1996), Aires (1999, 2000), Bondesson et al. (2004), Traat et al. (2004), and Tillé (2006).

### 2.8. *Order sampling*

Order sampling designs, developed by (Rosén 1997a, 1997b), are based upon an idea introduced by Ohlsson (1990a). The advantage of order sampling designs is their simplicity. Let $\pi_k$ be the target first inclusion probability of unit $k$. Consider a positive size variable $x_k > 0$ known for the whole population. The target inclusion probability $\pi_k$ is proportional to $x_k$ and computed as in (1). We generate $N$ uniform random numbers $\omega_k$ in [0,1] and the $n$ units that have the smallest values $\omega_k/\pi_k$ are selected. Other distributions for generating the random numbers can also be used, such as exponential distribution (Hájek, 1964) or Pareto (Rosén 1997a, 1997b) distribution. The main drawback of the method is that the inclusion probabilities are not exactly equal to $\pi_k$. Additional development on order sampling are given in Aires (1999, 2000), Ohlsson (1998), Rosén (2000), Matei and Tillé (2007), and Rosén (1995).

## 3. Point estimation in unequal probability sampling without replacement

We are often interested in estimating population totals of several characteristics of interest. It is therefore possible that some characteristics may not be related to the inclusion probabilities $\pi_k$. In this situation, Rao (1966) recommended the use of the following unweighted estimator

$$\widehat{Y}_{\mathrm{u}} = \frac{N}{n}\sum_{k\in s} y_k. \tag{6}$$

The design bias of this estimator is

$$\mathrm{bias}(\widehat{Y}_{\mathrm{u}}) = \frac{N}{n}\sum_{k\in U} y_k\pi_k - \sum_{k\in U} y_k = \frac{N^2}{n}\frac{1}{N}\sum_{k\in U}\left(y_k - \frac{\widehat{Y}_u}{N}\right)\left(\pi_k - \frac{n}{N}\right),$$

which is proportional to the covariance between $y_k$ and $\pi_k$. Thus, this bias is zero when $y_k$ and $\pi_k$ are uncorrelated. Rao (1966) showed that $\widehat{Y}_{\mathrm{u}}$ is on average more accurate than $\widehat{Y}_\pi$ because the average variance of $\widehat{Y}_{\mathrm{u}}$ is smaller under the following superpopulation model $\xi$,

$$y_k = \mu + \varepsilon_k, \tag{7}$$

with $\mathrm{E}_\xi(\varepsilon_k|\pi_k) = 0$, $\mathrm{E}_\xi(\varepsilon_k^2|\pi_k) = \sigma^2$, and $\mathrm{E}_\xi(\varepsilon_k\varepsilon_\ell|\pi_k) = 0$, where $\mathrm{E}_\xi(.)$ denotes the expectation under the superpopulation model $\xi$.

Amahia et al. (1989) considered the following linear combination of $\widehat{Y}_{\mathrm{u}}$ and $\widehat{Y}_\pi$

$$\widehat{Y}_a = (1-\rho)\widehat{Y}_{\mathrm{u}} + \rho\widehat{Y}_\pi,$$

where $\rho$ is the observed correlation between $y_k$ and $\pi_k$. This estimator gives more weights to $\widehat{Y}_\pi$ when $y_k$ and $\pi_k$ are highly correlated.

The Hájek (1971) estimator, given by

$$\widehat{Y}_{\mathrm{H}} = N\left(\sum_{k\in s}\frac{1}{\pi_k}\right)^{-1}\sum_{k\in s}\frac{y_k}{\pi_k}, \tag{8}$$

is an alternative estimator often used in unequal probability sampling. The estimator $\widehat{Y}_{\mathrm{H}}$ is approximately design unbiased. It should be used when $y_k$ and $\pi_k$ are uncorrelated because its variance may be small when $y_k$ follows model (7) (Särndal et al., 1992,

p. 258). This estimator is often used in practice because it is a weighted average with the sum of weights equal to $N$. This property is particularly useful for the estimation of counts that have to add up to a given constant. Note that with count estimation, the characteristic of interest might not be correlated with $\pi_k$.

When $y_k$ and $\pi_k$ are correlated, $\widehat{Y}_{\mathrm{u}}$ may not be efficient and therefore the $\pi$-estimator should be used instead. When $y_k$ and $\pi_k$ are uncorrelated, $\widehat{Y}_{\mathrm{u}}$ and $\widehat{Y}_{\mathrm{H}}$ should be used. Therefore, the choice of a point estimator should be driven by the correlation between $y_k$ and $\pi_k$ and the $\pi$-estimator should not be blindly used. Basu (1971) gave a famous example, where a circus owner wants to estimate the total weight of his 50 elephants. A sample of size one is selected with inclusion probabilities that are uncorrelated with the weight of each elephant: $\pi_1 = 99/100$ for Dumbo, the average elephant, and $\pi_k = 1/4900$ for the other elephants. Not surprisingly, Dumbo is selected. Let $y_1$ denote its weight. To estimate the total weight, a sensible estimator is $\widehat{Y}_{\mathrm{H}} = \widehat{Y}_{\mathrm{u}} = Ny_1$, which is different from the $\pi$-estimator $\widehat{Y}_\pi = y_1 100/99$.

Note that the variance estimator in (3) can be used to derive variance estimators for $\widehat{Y}_{\mathrm{u}}$, $\widehat{Y}_a$, and $\widehat{Y}_{\mathrm{H}}$. By substituting $y_k\pi_k N/n$ for $y_k$ in (3), we obtain a design unbiased estimator for the variance of $\widehat{Y}_{\mathrm{u}}$ when $\pi_{k\ell} > 0$. By substituting $y_k\pi_k/(n/N(1-\rho)+\rho\pi_k)$ for $y_k$ in (3), we obtain an approximately design unbiased estimator for the variance of $\widehat{Y}_a$ when $\pi_{k\ell} > 0$. By substituting $y_k - \widehat{Y}_{\mathrm{H}}$ for $y_k$ in (3), we obtain a approximately design unbiased estimator for the variance of $\widehat{Y}_{\mathrm{H}}$ when $\pi_{k\ell} > 0$.

The choice of the size variable should be driven by the correlation between the variable of interest and the size variable. Ideally, the size variable should be highly correlated with the variable of interest. However, in practice, we have several variables of interest and the size variable might be not correlated with all the variables of interest. In this situation, we recommend to use the simple mean (6) or the Hájek estimator (8) to estimate a total.

## 4. Variance estimators free of joint inclusion probabilities

Exact joint inclusion probabilities may be difficult or impossible to calculate. Futhermore, the double sum in (3) makes the Sen–Yates–Grundy estimator computationally intensive when the sample size is large. It is also inconceivable to provide these probabilities in released data sets, as the set of joint inclusion probabilities is a series of $n(n-1)/2$ values. Suppose that the sampling design uses single-stage stratified sampling with unequal probabilities within each stratum. Let $U_1, \ldots, U_H$ denote the strata. Suppose that a sample $s_h$ of size $n_h$ is selected without replacement within each stratum $U_h$ of size $N_h$. In this situation, we can estimate the variance of $\widehat{Y}_\pi$ approximately by

$$\widehat{\mathrm{var}}^*\left(\widehat{Y}_\pi\right) = \sum_{k\in s} \alpha_k \widehat{e}_k^2, \tag{9}$$

which is free of the $\pi_{l\ell}$. The $\widehat{e}_k$ are the residuals of weighted least squares given by

$$\widehat{e}_k = \frac{y_k}{\pi_k} - \sum_{h=1}^{H} \widehat{B}_h z_{kh},$$

and $\widehat{B}_h$ is the weighted least squares regression coefficient given by

$$\widehat{B}_h = \left( \sum_{k \in s} \lambda_k z_{kh}^2 \right)^{-1} \sum_{k \in s} \lambda_k z_{kh} \frac{y_k}{\pi_k},$$

where $z_{kh} = 1$ if $k \in U_h$ and otherwise $z_{kh} = 0$. The choice of $\alpha_k$ and $\lambda_k$ depends on the value of $n_h$ and on the sampling design implemented. Several choices are possible for the constants $\alpha_k$ and $\lambda_k$. A simple choice is $\alpha_k = \lambda_k = 1$, which gives the naive variance estimator under sampling with replacement given in (5). However, this approach usually leads to overestimation of the variance for large sampling fraction. When $\alpha_k = 1 - \pi_k(n_h - 1)/n_h$ for $k \in U_h$ and $\lambda_k = 1$, (9) reduces to the Hartley and Rao (1962) variance estimator. When $\alpha_k = \lambda_k = (1 - \pi_k)n_h/(n_h - 1)$, for $k \in U_h$, (9) reduces to the Hájek (1964) variance estimator.

For the randomized systematic sampling method, Hartley and Rao (1962) showed that var $(\widehat{Y}_\pi)$ reduces to

$$\operatorname{var}(\widehat{Y}_\pi) \approx \sum_{h=1}^{H} \sum_{k \in U_h} \pi_k \left( 1 - \frac{n_h - 1}{n_h} \pi_k \right) \left( \frac{y_k}{\pi_k} - \frac{Y}{n} \right)^2 \tag{10}$$

for fairly large $N_h$ and for small sampling fractions. Therefore, (9) will be a consistent estimator of (10) under the randomized systematic design, when $\alpha_k = 1 - \pi_k(n_h - 1)/n_h$ for $k \in U_h$, $\lambda_k = 1$. This choice is recommended when $n_h$ is small and $N_h$ is large, or when $n_h$ is large and $n_h/N_h$ is negligible.

Assuming $d_h = \sum_{\ell \in U_h} \pi_\ell (1 - \pi_\ell) \to \infty$, Hájek (1964) derived an approximation to $\pi_{k\ell}$ under maximum entropy sampling. By substituting this expression into (3), we have

$$\operatorname{var}(\widehat{Y}_\pi) = \sum_{k \in U} \pi_k (1 - \pi_k) e_k^2,$$

with

$$e_k = \frac{y_k}{\pi_k} - \sum_{h=1}^{H} B_h z_{kh},$$

where $B_h$ is the following population weighted least squares regression estimate

$$B_h = \left( \sum_{k \in U} (1 - \pi_k) z_{kh}^2 \pi_k \right)^{-1} \sum_{\ell \in U} (1 - \pi_\ell) z_{\ell h} y_\ell \pi_\ell.$$

Therefore, (9) will be a consistent estimator of (10) under maximum entropy sampling, when $\alpha_k = \lambda_k = 1 - \pi_k$ and $d_h \to \infty$. This choice is recommended when $n_h$ is large and the sampling fraction is not small. Berger (2007) showed that this choice gives a consistent estimator for the variance under the Rao–Sampford sampling design when $d_h \to \infty$, $H$ is bounded, and none of the $\pi_k$ less than 1 approach 1 asymptotically. Berger (2005a) showed that this choice is suitable for the Chao (1982) sampling design.

Other choices for $\alpha_k$ and $\lambda_k$ have been proposed in literature. When $\alpha_k = \lambda_k = (1 - \pi_k) \log(1 - \pi_k)/\pi_k$, (9) reduces to the Rosén (1991) estimator. When $\alpha_k = (1 - \pi_k) n_h (n_h - 1) \sum_{k \in s_h} (1 - \pi_k) \left( \sum_{k \in U_k} \pi_k (1 - \pi_k) \right)^{-1}$, (9) gives the Berger (1998) estimator. If $\alpha_k = \lambda_k = (1-\pi_k)^{-1}\left[1 - d_h^{-2}\sum_{\ell\in s_h}(1-\pi_\ell)\right]$ for $k \in U_h$, (9) gives the Deville (1999) variance estimator. Brewer (2002, Chapter 9) proposed two alternative choices for $\alpha_k$ and $\lambda_k$. Simulation studies by Brewer (2002), Haziza et al. (2004), Matei and Tillé (2005), and Henderson (2006) showed that (9) is an accurate estimator for various choices of $\alpha_k$ and $\lambda_k$. The variance estimator (9) may have a smaller mean square error than the exactly unbiased Sen–Yates–Grundy estimator in (3).

Berger (2005a) showed that (9) can be easily computed when $\alpha_k = \lambda_k$, as (9) reduces to $\widehat{\text{var}}^*(\widehat{Y}_\pi) = n\widehat{\sigma}_\varepsilon^2$, where $\widehat{\sigma}_\varepsilon^2$ is the observed residual variance of the regression

$$y_k^* = \sum_{h=1}^{H} \beta_h z_{\ell h}^* + \varepsilon_k$$

fitted with ordinary least squares, where the $\varepsilon_k$ are independent normal random variables with mean 0 and variances $\sigma_\varepsilon^2$, $y_k^* = y_k\pi_k^{-1}\alpha_k^{1/2}$ and $z_k^* = z_k\pi_k^{-1}\alpha_k^{1/2}$.

## 5. Variance estimation of a function of means

Assume that the parameter of interest $\theta$ can be expressed as a function of means of $Q$ survey variables, that is, $\theta = g(\mu_1, \ldots, \mu_Q)$, where $g(.)$ is a smooth differentiable function (Shao and Tu, 1995, Chapter 2), and $\mu_q$ is the finite population mean of the $q$th survey variables. This definition includes parameters of interest arising in common survey applications such as ratios, subpopulation means, and correlation and regression coefficients. It excludes parameters such as L-statistics (Shao, 1994) and coefficients of logistic regression, which cannot be expressed as function of means. The parameter $\widehat{\theta}$ can be estimated by the substitution estimator $\widehat{\theta} = g(\widehat{\mu}_{1H}, \ldots, \widehat{\mu}_{QH})$, in which $\widehat{\mu}_{qH}$ is the Hájek (1971) estimator of a the $q$th mean.

The variance of $\widehat{\theta}$ can be estimated by the linearized variance estimator (Robinson and Särndal, 1983) given by

$$\widehat{\text{var}}(\widehat{\theta})_{\text{L}} = \nabla(\widehat{\boldsymbol{\mu}})'\widehat{\Sigma}\nabla(\widehat{\boldsymbol{\mu}}),$$

where

$$\widehat{\Sigma} = \frac{1}{N^2}\sum_{k\in s}\sum_{\ell\in s}\left(\frac{\pi_{k\ell} - \pi_k\pi_\ell}{\pi_k\pi_\ell\pi_{k\ell}}\right)(\mathbf{y}_k - \widehat{\boldsymbol{\mu}})(\mathbf{y}_\ell - \widehat{\boldsymbol{\mu}})',$$

$$\nabla(\mathbf{x}) = \left(\frac{\partial g(\boldsymbol{\mu})}{\partial\mu_1}, \ldots, \frac{\partial g(\boldsymbol{\mu})}{\partial\mu_Q}\right)'_{\mu=\mathbf{x}},$$

$\mathbf{y}_k = (y_{1k}, \ldots, y_{Qk})'$, $\nabla(\mathbf{x})$ denotes the gradient of $g(\cdot)$ at $\mathbf{x} \in \mathbb{R}^Q$, $\widehat{\boldsymbol{\mu}} = (\widehat{\mu}_{1H}, \ldots, \widehat{\mu}_{QH})'$, and $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_Q)'$.

Customary jackknife variance estimators (Shao and Tu, 1995; Wolter, 1985) are not always consistent under unequal probability sampling without replacement (Demnati and Rao, 2004). Campbell (1980) proposed a generalised jackknife variance estimator that allows us to estimate the variance for unequal probability sampling and stratification. Campbell's generalised jackknife is given by

$$\widehat{\text{var}}(\widehat{\theta}) = \sum_{k\in s}\sum_{\ell\in s}\frac{\pi_{k\ell} - \pi_k\pi_\ell}{\pi_{k\ell}}u_k u_\ell,$$

where

$$
\begin{aligned}
u_j &= (1 - w_j)(\widehat{\theta} - \widehat{\theta}_{(j)}),\\
w_j &= \pi_j^{-1}\left(\sum_{k\in s}\pi_k^{-1}\right)^{-1},\\
\widehat{\theta}_{(j)} &= g(\widehat{\mu}_{1H(j)}, \ldots, \widehat{\mu}_{QH(j)}),\\
\widehat{\mu}_{qH(j)} &= N\left(\sum_{k\in s}\frac{\delta_{kj}}{\pi_k}\right)^{-1}\sum_{k\in s}\frac{\delta_{kj}y_k}{\pi_k},
\end{aligned}
$$

and $\delta_{kj} = 1$ if $k = j$ and $\delta_{kj} = 0$ otherwise. Berger and Skinner (2005) gave regularity conditions under which the generalised jackknife is consistent. They also showed that the generalised jackknife may be more accurate than the customary jackknife estimators. Berger (2007) proposed an alternative consistent jackknife estimator that is free of joint inclusion probabilities.

Many surveys use single imputation to handle item nonresponse. Treating the imputed values as if they were true values and then estimating the variance using standard methods may lead to serious underestimation of the variance when the proportion of missing values is large (Rao and Shao, 1992; Särndal, 1992). One can use the Rao–Shao method which consists of adjusting the imputed values whenever a responding unit is deleted. Berger and Rao (2006) showed that this method gives a consistent generalised jackknife variance estimator under uniform response.

## 6. Balanced sampling

### 6.1. Definition

A design is balanced if the $\pi$-estimators for a set of auxiliary variables are equal to the known population totals of auxiliary variables. Balanced sampling can be viewed as a calibration method embedded into the sampling design. Yates (1949) advocated the idea of respecting the means of known variables in probability samples. Yates (1946) and Neyman (1934) described methods of balanced sampling limited to one variable and to equal inclusion probabilities. The use of balanced sampling was recommend by Royall and Herson (1973) for protecting inference against misspecified models. More recently, several partial solutions were proposed by Deville et al. (1988), Deville (1992), Ardilly (1991), and Hedayat and Majumdar (1995). Valliant et al. (2000) surveyed some existing methods.

The cube method (Deville and Tillé, 2004) is a general method of balanced sampling with equal or unequal inclusion probabilities. Properties and application of this method were studied in Deville and Tillé (2004), Chauvet and Tillé (2006), Tillé and Favre (2004, 2005), Berger et al. (2003), and Nedyalkova and Tillé (2008). The cube method was used to select the rotation groups of the new French census (Bertrand et al., 2004; Dumais and Isnard, 2000; Durr and Dumais, 2002) and the selection of the French master sample (Christine, 2006; Christine and Wilms, 2003; Wilms, 2000). Deville and Tillé (2005) proposed a variance estimator for balanced sampling. Deville (2006) also proposed to use balanced sampling for the imputation of item nonresponse. The cube method can be implemented in practice by SAS® or R procedures (Chauvet and Tillé, 2005; Rousseau and Tardieu, 2004; Tardieu, 2001; Tillé and Matei, 2007).

Balancing is used when auxiliary information is available at the design stage. When balanced sampling is used, the Horvitz–Thompson weights are also calibration weights. Calibration after sampling is therefore not necessary. Balancing also provide more stable estimators as these weights do not depend on the sample.

### *6.2. Balanced sampling and the cube method*

Suppose that the values of $p$ auxiliary variables $x_1, \dots x_p$ are known for every unit of the population. Let $\mathbf{x}_k = (x_{k1} \cdots x_{kj} \cdots x_{kp})'$ be the vector of the $p$ auxiliary variables on unit $k$. For a set of given inclusion probabilities $\pi_k$, a design $p(.)$ is balanced with respect to the auxiliary variables $x_1, \dots, x_p$, if and only if it satisfies the balancing equations given by

$$\sum_{k \in s} \frac{\mathbf{x}_k}{\pi_k} = \sum_{k \in U} \mathbf{x}_k. \tag{11}$$

Balanced sampling generalises several well-known methods. For instance, if $\mathbf{x}_k = \pi_k$, then (11) is a fixed size constraint. It can be shown that, if the auxiliary variables are the indicator variables of strata, a stratified sampling design is balanced on these indicator variables.

However, it is often not possible to find a sample such that (11) holds, for example, when the right-hand side of (11) is an integer. Hence, an exactly balanced design often does not exist. For example, if $x_1 = 1, x_2 = 1, x_3 = 1, x_1 = 5$, and $\pi_k = 1/2$, for $i = 1, 2, 3, 4,$ the balancing equation becomes

$$\sum_{k \in s} 2x_k = \sum_{k \in U} x_k = 11, \tag{12}$$

which cannot hold. The aim is to select an exact balanced sample if possible, and an approximately balanced sample otherwise.

The name "cube method" comes from the geometrical representation of a sampling design. A sample can be written as a vector $\mathbf{s} = (s_1, \dots, s_N) \in \mathbb{R}^N$ of indicator variables $s_k$ that take the value 1 if the unit is selected and 0 otherwise. Geometrically, each vector $\mathbf{s}$ can be viewed as one of the $2^N$ vertices of a $N$-cube in $\mathbb{R}^N$. A design consists in allocating a probability $p(.)$ to each vertex of the $N$ cube in such a way that the expectation of $\boldsymbol{s}$ is equal to the inclusion probability vector $\boldsymbol{\pi}$, that is,

$$\mathrm{E}(\mathbf{s}) = \sum_{s \in \mathcal{S}} p(s)\mathbf{s} = \boldsymbol{\pi},$$

where $\boldsymbol{\pi} \in \mathbb{R}^N$ is the vector of inclusion probabilities. Thus, selecting a sample consists in choosing a vertex (a sample) of the $N$-cube that is balanced.

The balancing equations in (11) can also be written as

$$\sum_{k \in U} \mathbf{a}_k s_k = \sum_{k \in U} \mathbf{a}_k \pi_k \text{ with } s_k \in \{0, 1\}, k \in U,$$

where $\mathbf{a}_k = \mathbf{x}_k/\pi_k, k \in U$. The balancing equations define an affine subspace in $\mathbb{R}^N$ of dimension $N - p$ denoted $Q$. The subspace $Q$ can be written as $\boldsymbol{\pi}$ + Ker$\mathbf{A}$, where Ker$\mathbf{A} = \{\boldsymbol{u} \in \mathbb{R}|\mathbf{A}\boldsymbol{u}\}$ and $\mathbf{A} = (\mathbf{a}_1 \cdots \mathbf{a}_n \cdots \mathbf{a}_N)$.

It is possible to geometrically represent the situation when (12) does not hold. When the vertices of the intersection between the cube and $Q$ are also vertices of the cube, as in Fig. 1, a balanced sample can be selected. When the vertices of the intersection between the cube and $Q$ are not vertices of the cube, as in Fig. 2, it is not possible to select an exact balanced sample. In this situation, only an approximately balanced sample can be selected (see Section 6.4).

### 6.3. *The flight phase*

The cube method is made up of two parts: the flight phase and the landing phase. *The flight phase*, described in Algorithm 1 below, is a random walk which begins at the vector of inclusion probabilities and remains in the intersection of the cube and the constraint subspace $Q$. This random walk stops at a vertex of the intersection of the cube and the constraint subspace. There are several ways to implement this algorithm. Chauvet and

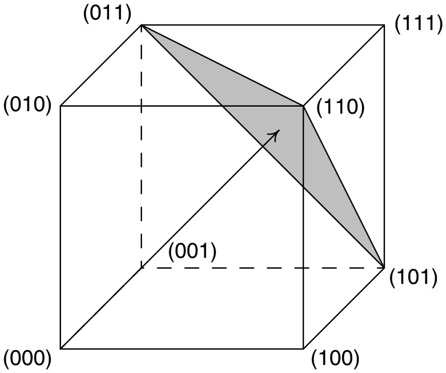


Fig. 1. Fixed size constraint of size 2: an exact balanced sample always exists.

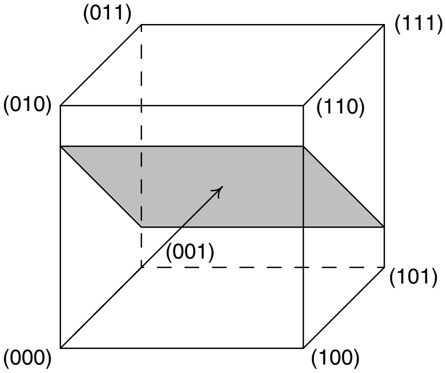


Fig. 2. The balanced constraints are such that an exact balanced sample does not exist.

Tillé (2006) proposed a fast algorithm whereby the calculation time increases linearly with the population size.

**Algorithm 1:** Flight phase of the cube method

```
First initialize with $\boldsymbol{\pi}(0) = \boldsymbol{\pi}$.
Next, at time $t = 1, \ldots., T$,
    1. Generate any vector $\boldsymbol{u}(t) = [u_k(t)] \neq 0$ so that
       (i) $\boldsymbol{u}(t)$ is in the kernel of matrix $\boldsymbol{A}$
       (ii) $u_k(t) = 0$ if $\pi_k(t)$ is an integer.
    2. Compute $\lambda_1^*(t)$ and $\lambda_2^*(t)$, the largest values so that
       $0 \leq \boldsymbol{\pi}(t) + \lambda_1(t)\boldsymbol{u}(t) \leq 1$,
       $0 \leq \boldsymbol{\pi}(t) - \lambda_2(t)\boldsymbol{u}(t) \leq 1$.
    3. Compute $\boldsymbol{\pi}(t+1) = \begin{cases} \boldsymbol{\pi}(t) + \lambda_1^*(t)\boldsymbol{u}(t) & \text{with a proba } q_1(t) \\ \boldsymbol{\pi}(t) - \lambda_2^*(t)\boldsymbol{u}(t) & \text{with a proba } q_2(t), \end{cases}$
       where $q_1(t) = \lambda_2^*(t)/\{\lambda_1^*(t) + \lambda_2^*(t)\}$ and $q_2(t) = 1 - q_1(t)\}$.
```

### *6.4. Landing phase*

*The landing phase* begins at the end of the flight phase. If a sample is not obtained at the end of the flight phase, a sample is selected as close as possible to the constraint subspace. At the end of the flight phase, Algorithm 1 stops on a vertex denoted $\boldsymbol{\pi}^*$ of the intersection between the cube and $Q$. It is possible to show that

$$\text{card } U^* = \text{card}\left\{k \in U | 0 < \pi_k^* < 1\right\} = q \leq p,$$

which means that the number of noninteger elements of $\boldsymbol{\pi}^*$ is smaller or equal to the number of balancing variables. The aim of the landing phase is to find a random sample $\boldsymbol{s}$ so that $\mathrm{E}\,(\boldsymbol{s}|\boldsymbol{\pi}^*) = \boldsymbol{\pi}^*$ and which is almost balanced.

Two solutions can be used to select the sample. The first solution consists of enumerating all the samples that are consistent with $\boldsymbol{\pi}^*$, a sample $\mathbf{s}$ being consistent if $s_k = \pi_k^*$ when $\pi_k^*$ is an integer. Then, a cost $C(\mathbf{s})$ is attached at each sample. This cost is equal to zero when the sample is balanced and which increases when the sample moves away from the subspace $Q$. Deville and Tillé (2004) proposed several $C(\mathbf{s})$. By a method of linear programming, it is possible to find a sampling design on the consistent samples that satisfies the inclusion probability $\boldsymbol{\pi}^*$ and which minimizes the average cost. Finally, a sample is selected at random, following this sampling design. This method can be used with a number of balancing variables that are less than 15 because it is necessary to enumerate the $2^{15}$ samples.

The second method can be used when the number of auxiliary variables is too large for the solution to be obtained by a simplex algorithm. At the end of the flight phase, an auxiliary variable can be dropped out. Next, one can return to the flight phase until it is no longer possible to "move" within the constraint subspace. Thus, the constraints are successively relaxed until the sample is selected.

*Sample Surveys: Design, Methods and Applications, Vol. 29A*
ISSN: 0169-7161

DOI: 10.1016/S0169-7161(08)00003-5

3

# Two-Phase Sampling

*Jason C. Legg and Wayne A. Fuller*

## 1. Introduction

Two-phase sampling is typically used when it is very expensive to collect data on the variables of interest, but it is relatively inexpensive to collect data on variables that are correlated with the variables of interest. For example, in forest surveys, it is very difficult and expensive to travel to remote areas to make on-ground determinations. However, aerial photographs are relatively inexpensive and determinations on, say, forest type are strongly correlated with ground determinations (see Breidt and Fuller, 1993; Schreuder et al., 1993).

Two-phase sampling was called double sampling by Neyman (1938) in the seminal article. Neyman states that the problem was posed to him at the U.S. Department of Agriculture. A survey was to be conducted to estimate the total of a characteristic $y$. The determinations were very costly, but another variable, say $x$, was known to be correlated with $y$ and was cheap to observe. Neyman's solution was to spend some of the available funds to make many cheap observations, divide the large sample into groups (second-phase strata) based on observed $x$'s, and select a sample from each of the groups.

To develop a formal description of two-phase sampling, let $U_N$ denote the set of indices for the finite population containing indices 1 to $N$. Let $\mathcal{F}_N$ be the finite population of all elements indexed in $U_N$. An initial, often large, sample of size $n_{1N}$ is selected from $\mathcal{F}_N$. Let $A_{1N}$ be the set of indices in the first-phase sample. A second-phase sample of size $n_{2N}$ is selected from $A_{1N}$. Let $A_{2N}$ be the set of indices in the second-phase sample. Often, only $\boldsymbol{x}$ is observed on elements in $A_1$ and both $\boldsymbol{x}$ and $y$ are observed for elements in $A_2$. For large sample approximations, we assume a sequence of finite populations and samples indexed by $N$, where each $\mathcal{F}_N$ is a sample from an infinite superpopulation with finite eighth moments. The subscript of $N$ will often be suppressed.

In Neyman's original formulation, the vector $\boldsymbol{x}_i = (x_{1i}, \ldots, x_{Gi})$ is the vector of indicators for $G$ groups, where the groups are also called second-phase strata. Let the first-phase sample be selected with inclusion probabilities $\pi_{1i} = \Pr(i \in A_1|\mathcal{F}_N)$ and let the first-phase weights be $w_{1i} = \pi_{1i}^{-1}$. Let

$$\begin{aligned}\bar{\boldsymbol{x}}_1 &= N^{-1} \sum_{i \in A_1} w_{1i} \boldsymbol{x}_i \\ &=: (\bar{x}_{11}, \ldots, \bar{x}_{1G})\end{aligned} \tag{1}$$

be the mean of $\boldsymbol{x}$ from the first phase. The mean $\bar{\boldsymbol{x}}_1$ estimates the fraction of the population in the groups. The second-phase sample in Neyman's problem is selected by taking samples within each group. Let the conditional second-phase inclusion probability for element $i$ be $\pi_{2i|1i} = \Pr(i \in A_2 | i \in A_1, \mathcal{F}_N)$ and the conditional second-phase weight be $w_{2i|1i} = \pi_{2i|1i}^{-1}$. An estimator of the mean for $y$ is

$$\bar{y}_{2,st} = N^{-1} \sum_{g=1}^{G} \sum_{i \in A_{2g}} w_{2i|1i} w_{1i} y_i, \tag{2}$$

where $A_{2g}$ is the set of indices for the second-phase sample in group $g$. Estimator (2) can be written as

$$\bar{y}_{2,st} = \sum_{g=1}^{G} \bar{x}_{1g} \bar{y}_{2g}, \tag{3}$$

where

$$\bar{y}_{2g} = \left( \sum_{i \in A_{1g}} w_{1i} \right)^{-1} \sum_{i \in A_{2g}} w_{2i|1i} w_{1i} y_i, \tag{4}$$

and $A_{1g}$ is the first-phase sample in group $g$. The form (3) is that of the usual stratified mean estimator with the population group sizes replaced by first-phase estimators. Estimator (2) was named the double expansion estimator (DEE) by Kott and Stukel (1997).

### *1.1. Double expansion estimator*

The DEE is the building block for more complicated two-phase estimators. The DEE for a total is

$$\hat{T}_{y,2} = \sum_{i \in A_2} w_{2i|1i} w_{1i} y_i. \tag{5}$$

The DEE is not the standard Horvitz–Thompson estimator because $\pi_{2i|1i}\pi_{1i}$ is not, in general, the same as $\pi_{2i} = \Pr(i \in A_2 | \mathcal{F}_N)$. Often, $\pi_{2i}$ is unknown. If $\pi_{2i|1i}$ depends on $A_1$, then the composition of all first-phase samples is needed to determine $\pi_{2i}$. However, conditional on $A_1$, The DEE is a Horvitz–Thompson estimator for the first-phase Horvitz–Thompson total estimator

$$\hat{T}_{y,1} = \sum_{i \in A_1} w_{1i} y_i. \tag{6}$$

The properties of the DEE are derived by applying the properties of Horvitz–Thompson estimators conditional on $A_1$. The DEE is unbiased because

$$E\{\hat{T}_{y,2}|\mathcal{F}_N\} = E\{E[\hat{T}_{y,2}|A_1, \mathcal{F}_N]|\mathcal{F}_N\} = E\{\hat{T}_{y,1}|\mathcal{F}_N\} = T_y, \tag{7}$$

where $T_y$ is the population total for $y$. The variance of the DEE is often expressed as

$$\begin{aligned} V\{\hat{T}_{y,2}|\mathcal{F}_N\} &= E\{V[\hat{T}_{y,2}|A_1, \mathcal{F}_N]|\mathcal{F}_N\} + V\{E[\hat{T}_{y,2}|A_1, \mathcal{F}_N]|\mathcal{F}_N\} \\ &= E\{V[\hat{T}_{y,2}|A_1, \mathcal{F}_N]|\mathcal{F}_N\} + V\{\hat{T}_{y,1}|\mathcal{F}_N\}. \end{aligned} \tag{8}$$

The first term on the right side of (8) can be estimated using the Horvitz–Thompson variance for a sample from $A_1$ of $w_{1i}y_i$ with inclusion probabilities of $\pi_{2i|1i}$. Denote this estimated conditional variance by $\hat{V}\{\hat{T}_{y,2}|A_1, \mathcal{F}_N\}$. Using the arguments used to construct the total estimator, an estimator of the Horvitz–Thompson first-phase variance estimator is

$$\hat{V}\{\hat{T}_{y,1}|\mathcal{F}_N\} = \sum_{i\in A_2}\sum_{j\in A_2} \pi_{2ij|1ij}^{-1}\pi_{1ij}^{-1}(\pi_{1ij} - \pi_{1i}\pi_{1j})\pi_{1i}^{-1}y_i\pi_{1j}^{-1}y_j, \tag{9}$$

where $\pi_{2ij|1ij} = \Pr\{(i, j) \in A_2|(i, j) \in A_1, \mathcal{F}_N\}$. See Särndal et al. (1992, Chapter 9). A second approach to variance estimation is to estimate each of the two terms of (8). Typically, $V\{\hat{T}_{y,1}|\mathcal{F}_N\}$ is more difficult to estimate than the conditional variance. We illustrate estimation of $V\{\hat{T}_{y,1}|\mathcal{F}_N\}$ using method of moments estimators.

**Example** Suppose the first-phase sample is a stratified random sample with $H$ strata, the first-phase sample is stratified into $G$ strata, and a second-phase sample is selected using a stratified random sample design. Then,

$$V\{\hat{T}_{y,1}|\mathcal{F}_N\} = \sum_{h=1}^{H} N_h(N_h - n_h)n_h^{-1}S_{yh}^2, \tag{10}$$

where

$$S_{yh}^2 = (N_h - 1)^{-1}\sum_{i\in U_h}(y_i - \overline{y}_{Nh})^2, \tag{11}$$

$\overline{y}_{Nh}$ is the population mean for stratum $h$, $U_h$ is the set of indices in stratum $h$, and $N_h$ is the number of elements in stratum $h$. An estimator of $S_{yh}^2$ is

$$\hat{S}_{yh}^2 = \left[\sum_{i\in B_{2h}} w_{B_{2hi}}(1 - w_{B_{2hi}})\right]^{-1} \sum_{i\in B_{2h}} w_{B_{2hi}}(y_i - \overline{y}_{B2,h})^2, \tag{12}$$

where

$$w_{B_{2hi}} = \left(\sum_{j\in B_{2h}} w_{2i|1i}w_{1i}\right)^{-1} w_{2i|1i}w_{1i}, \tag{13}$$

$$\bar{y}_{B2,h} = \left( \sum_{i \in B_{2h}} w_{2i|1i} w_{1i} \right)^{-1} \sum_{i \in B_{2h}} w_{2i|1i} w_{1i} y_i, \tag{14}$$

$B_{2h} = A_{1h} \bigcap A_2$, and $A_{1h}$ is the set of indices in the first-phase sample in stratum $h$. The estimator $\hat{V}\{\hat{T}_{y,1}|\mathcal{F}_N\}$ is formed by substituting (12) into (10). A conditional second-phase variance estimator is the variance estimator for the estimated total of the sample of $w_{1i}y_i$. The conditional variance is

$$V\{\hat{T}_{y,2}|A_1, \mathcal{F}_N\} = \sum_{g=1}^{G} n_{1g}(n_{1g} - n_{2g}) n_{2g}^{-1} S_{1wy,g}^2, \tag{15}$$

where

$$S_{1wy,g}^2 = \sum_{i \in A_{1g}} (n_{1g} - 1)^{-1} (w_{1i} y_i - \bar{y}_{1w,g})^2, \tag{16}$$

$$\bar{y}_{1w,g} = \sum_{i \in A_{1g}} n_{1g}^{-1} w_{1i} y_i, \tag{17}$$

$A_{1g}$ is the portion of the first-phase sample in second-phase stratum $g$, $n_{1g}$ is the number of elements in the first-phase sample in second-phase stratum $g$, and $n_{2g}$ is the sample size in second-phase stratum $g$. The conditional variance estimator $\hat{V}\{\hat{T}_{y,2}|A_1, \mathcal{F}_N\}$ is constructed by replacing the $S_{1wy,g}^2$ and $\bar{y}_{1w,g}$ in (15) and (16) with their sample estimators

$$\hat{S}_{1wy,g}^2 = (n_{2g} - 1)^{-1} \sum_{i \in A_{2g}} (w_{1i} y_i - \bar{y}_{2w,g})^2 \tag{18}$$

and

$$\bar{y}_{2w,g} = n_{2g}^{-1} \sum_{i \in A_{2g}} w_{1i} y_i. \tag{19}$$

The sum of $\hat{V}\{\hat{T}_{y,2}|A_1, \mathcal{F}_N\}$ and $\hat{V}\{\hat{T}_{y,1}|\mathcal{F}_N\}$ estimates $V\{\hat{T}_{y,2}|\mathcal{F}_N\}$. □

Variance estimation is considered further in Section 2.4. Also see Kott, 1990.

### *1.2. Costs for one- and two-phase designs*

Two-phase sampling attempts to reduce the variance of the estimated total by using the correlation between $\boldsymbol{x}$ and $y$ in constructing a total estimator. However, two-phase sampling is not always superior to one-phase designs. Given a fixed cost, selecting a first-phase sample reduces the number of observations on the response variable $y$. The relative cost between first- and second-phase observations, correlation between first- and second-phase observations, and the variance of the response variable determines whether a two-phase design is superior to a single-phase design.

For example, consider a first-phase sample selected using simple random sampling without replacement. Let the second-phase sample be a stratified random sample with proportional allocation. Let $\sigma_y^2$ be the population variance of $y$ and $\sigma_w^2$ be the within second-phase strata variance of $y$. Let $\sigma_b^2 = \sigma_y^2 - \sigma_w^2$. Assume each first-phase observation costs $c_1$ and each second-phase observation costs $c_2$. Ignoring the first-phase finite population correction,

$$V\{N^{-1}\hat{T}_{y,2}|\mathcal{F}_N\} = n_1^{-1}\sigma_y^2 + (n_2^{-1} - n_1^{-1})\sigma_w^2 = n_1^{-1}\sigma_b^2 + n_2^{-1}\sigma_w^2. \tag{20}$$

For a fixed total cost of $C$,

$$n_1 = \left(c_1 + c_2\sqrt{c_2^{-1}c_1\sigma_b^{-2}\sigma_w^2}\right)^{-1} C \tag{21}$$

and

$$n_2 = c_2^{-1}(C - c_1 n_1) \tag{22}$$

minimize (20).

Suppose we have 10,000 units to spend, the cost for a first-phase unit is one, the cost for observing $y$ is three, $\sigma_y^2 = 100$, and $\sigma_w^2 = 40$. The $n_1$ of (21) is 4141 and the $n_2$ of (22) is 1953. The two-phase variance in (20) is 0.035. As an alternative design, consider selecting a simple random sample of size 3333 and observing $y$ on the selected sample. The variance of the alternative design is 0.030. Therefore, the proposed two-phase design is less efficient than using only the first-phase design as a single phase. In order for two-phase sampling to be beneficial, the within strata variance must be smaller or (and) $c_2^{-1}c_1$ must be smaller than those of the example.

Suppose the cost of observing $y$ is increased from 3 to 100. Then the optimal sample sizes are $n_1 = 1100$ and $n_2 = 89$. The two-phase variance of the mean estimator is 0.504. If we use the simple random sample design, the variance of the mean is one. Therefore, the two-phase design is nearly twice as efficient as simple random sampling. If we keep the cost of observing $y$ at three and decrease $\sigma_w^2$ to 20, the optimal two-phase sample sizes are $n_1 = 5359$ and $n_2 = 1547$, the two-phase variance in (20) is 0.028, and the variance of the simple random sample mean is 0.030.

### *1.3. Uses for the two-phase sampling structure*

Two-phase designs and estimators were introduced as a way to reduce the variance of estimated parameters relative to single-phase sampling. The two-phase framework can be applied in missing data problems, sampling at multiple occasions, and situations without a good frame. For missing data in a survey, the first-phase sample is analogous to the target sample. The second-phase sample is the set of elements in the target sample that are observed, but the inclusion probabilities for the second-phase design are generally unknown and need to be estimated. A common assumption is that the second-phase sample is a stratified Bernoulli sample, where the strata are defined by known characteristics of the elements in the target sample.

For studies with multiple time observations, the first-phase sample can be considered to be the set of all units that will be observed at some time point. The sample observed at a particular time is a second-phase sample from that of first-phase sample. Estimators can be constructed that combine two-phase estimators across the two-phase samples. For example, see Breidt and Fuller (1999).

When studying rare, remote, or otherwise previously uncataloged populations, good quality sampling frames may not exist. A large sample can be used to identify a set of units in the population. Then a sample can be selected from the identified set. One approach is to use an area sample for the first phase, listing all units contained within the selected areas. A second-phase sample is then selected using the list.

## 2. Using auxiliary information in estimation

The DEE uses first-phase information in the conditional Horvitz–Thompson estimator. The observations on $\boldsymbol{x}$ can be used to construct other estimators. Ratio and regression estimators used in single-phase sampling have analogous two-phase forms. Typically, two-phase estimators are constructed by replacing known population quantities with their corresponding first-phase estimators. We shall focus on the two-phase regression estimator, which includes the two-phase version of the ratio estimator as a special case. Other forms of estimators have been suggested. See Shabbir and Gupta (2007) and Samiuddin and Hanif (2007).

### *2.1. Reweighted expansion estimator*

In stratified sampling at the second phase, an important estimator is

$$\hat{T}_{y,r,2} = \sum_{g=1}^{G} \hat{N}_{1g}\bar{y}_{2rg}, \tag{23}$$

where $g = 1, 2, \ldots, G$ are the second-phase strata,

$$\hat{N}_{1g} = \sum_{i \in A_{1g}} w_{1i} \tag{24}$$

and

$$\bar{y}_{2rg} = \left( \sum_{i \in A_{2g}} w_{1i} \right)^{-1} \sum_{i \in A_{2g}} w_{1i} y_i. \tag{25}$$

Estimator (23) is called the reweighted expansion estimator (REE) by Kott and Stukel (1997). Variance estimation for the REE has been studied by Rao and Shao (1992) and Kim et al. (2006). We call estimators of the form (25) Hájek mean estimators.

Under mild assumptions (Kim et al., 2006), the variance of the REE is

$$V\{\hat{T}_{y,r,2}|\mathcal{F}_N\} = V\{\hat{T}_{y,1}|\mathcal{F}_N\} + E\left\{ \sum_{g=1}^{G} n_{1g}^2 (r_g^{-1} - n_{1g}^{-1}) S_{1we,g}^2 | \mathcal{F}_N \right\} + o(n_1^{-1}N^2), \tag{26}$$

where $r_g = \pi_{2i|1i} n_{1g}$ is the second-phase sample size for stratum g,

$$S_{1we,g}^2 = (n_{1g} - 1)^{-1} \sum_{i \in A_{1g}} w_{1i}^2 e_{ig}^2, \tag{27}$$

$e_{ig} = y_i - \overline{y}_{Ng}$, and $\overline{y}_{Ng}$ is the population mean of $y$ in group $g$. In the proof of (26), it is assumed that $\pi_{2i|1i}$ is constant for all $i \in A_{1g}$. The REE is a weighted sum of ratio estimators, therefore the REE is subject to ratio bias. For equal probability first-phase samples, the REE is equal to the DEE and therefore the REE is unbiased (Cochran, 1977, Chapter 12). Under the assumptions used in computing the order of the variance approximation, the bias is negligible in large samples. The variance of the REE is small when the $y$'s within a group are homogeneous, whereas the variance of the DEE is small when $w_{1i}y_i$'s are homogeneous within a group.

### 2.2. *Two-phase regression estimator*

Let the vector $\boldsymbol{x}_i$ be observed for all elements in $A_1$. In the case of the stratified estimator for a stratified second-phase sample, $\boldsymbol{x}_i$ is a vector of stratum indicators. A two-phase regression estimator of the mean is

$$\overline{y}_{2,\text{reg}} = \overline{y}_{2\pi} + (\overline{\boldsymbol{x}}_{1\pi} - \overline{\boldsymbol{x}}_{2\pi})\hat{\beta}_{2\pi,y,x}, \tag{28}$$

where

$$(\overline{y}_{2\pi}, \overline{\boldsymbol{x}}_{2\pi}) = \left(\sum_{i\in A_2} w_{2i|1i}w_{1i}\right)^{-1} \sum_{i\in A_2} w_{2i|1i}w_{1i}(y_i, \boldsymbol{x}_i), \tag{29}$$

$$\overline{\boldsymbol{x}}_{1\pi} = \left(\sum_{i\in A_1} w_{1i}\right)^{-1} \sum_{i\in A_1} w_{1i}\boldsymbol{x}_i, \tag{30}$$

$$\hat{\beta}_{2\pi,y,x} = \left(\sum_{i\in A_2} w_{2i|1i}w_{1i}(\boldsymbol{x}_i - \overline{\boldsymbol{x}}_{2\pi})'(\boldsymbol{x}_i - \overline{\boldsymbol{x}}_{2\pi})\right)^{-1} \sum_{i\in a_2} w_{2i|1i}w_{1i}(\boldsymbol{x}_i - \overline{\boldsymbol{x}}_{2\pi})'y_i, \tag{31}$$

and $w_{2i|1i} = \pi_{2i|1i}^{-1}$. The two-phase regression estimator adjusts the direct estimator, $\overline{y}_{2\pi}$, by a multiple of the difference between first- and second-phase estimators for the mean of $\boldsymbol{x}$. An alternative to the regression coefficient $\hat{\beta}_{2\pi,y,x}$ is

$$\hat{\beta}_{2\pi,egls,x,y} = [\hat{V}(\overline{x}_{2\pi}|\mathcal{F}_N)]^{-1}\hat{C}(\overline{x}_{2\pi}, \overline{y}_{2\pi}|\mathcal{F}_N), \tag{32}$$

where the two estimators on the right side of (32) are design consistent estimators for $V(\overline{x}_{2\pi}|\mathcal{F}_N)$ and $C(\overline{x}_{2\pi}, \overline{y}_{2\pi}|\mathcal{F}_N)$, respectively. Often, $\hat{\beta}_{2\pi,egls,x,y}$ is difficult to evaluate.

### 2.3. *Approximations for two-phase regression estimators*

Since the two-phase regression estimator is a nonlinear function of first- and second-phase DEEs, we consider a Taylor linearization approximation to the error in $\overline{y}_{2,\text{reg}}$. To obtain the order of the approximation and a limiting distribution, we adopt the population and sample framework of Fuller (1975). Let $\{y_i, \boldsymbol{x}_i\}$ be a sequence of *iid* random variables with $4 + \delta$ moments. Let $\{\mathcal{F}_N, A_{1N}\}$ be a sequence of populations and first-phase samples. Suppose the first- and second-phase designs are such that

$$E\{|\overline{\boldsymbol{x}}_{2\pi} - \overline{\boldsymbol{x}}_N|^2|\mathcal{F}_N\} = O_p(n_1^{-1}), \tag{33}$$

$$E\{|\overline{\boldsymbol{x}}_{1\pi} - \overline{\boldsymbol{x}}_N|^2|\mathcal{F}_N\} = O_p(n_1^{-1}), \tag{34}$$

and

$$E\{|\hat{\beta}_{2\pi,y,x} - \beta_{y,x,N}|^2|\mathcal{F}_N\} = O_p(n_1^{-1}), \tag{35}$$

where $\bar{\boldsymbol{x}}_N$ is the finite population mean and

$$\beta_{y,x,N} = \left[\sum_{i=1}^{N}(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_N)'(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_N)\right]^{-1} \sum_{i=1}^{N}(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_N)'(\boldsymbol{y}_i - \bar{\boldsymbol{y}}_N). \tag{36}$$

Then,

$$\bar{y}_{2,\mathrm{reg}} - \bar{y}_N = \bar{e}_{2\pi} + (\bar{\boldsymbol{x}}_{1\pi} - \bar{\boldsymbol{x}}_N)\beta_{y,x,N} + O_p(n_1^{-1}), \tag{37}$$

where

$$\bar{e}_{2\pi} = \left(\sum_{i \in A_2} w_{2i|1i}w_{1i}\right)^{-1} \sum_{i \in A_2} w_{2i|1i}w_{1i}e_i \tag{38}$$

and $e_{Ni} = y_i - \bar{y}_N - (\boldsymbol{x}_i - \bar{\boldsymbol{x}}_N)\beta_{y,x,N}$.

The variance of the approximating variable in (37) is

$$\begin{aligned} V\{\bar{e}_{2\pi}+(\bar{\boldsymbol{x}}_{1\pi}-\bar{\boldsymbol{x}}_N)\beta_{y,x,N}|\mathcal{F}_N\} &= V\{\bar{e}_{2\pi}|\mathcal{F}_N\} + V\{(\bar{\boldsymbol{x}}_{1\pi} - \bar{\boldsymbol{x}}_N)\beta_{y,x,N}|\mathcal{F}_N\} \\ &\quad + 2C\{\bar{e}_{2\pi}, (\bar{\boldsymbol{x}}_{1\pi} - \bar{\boldsymbol{x}}_N)\beta_{y,x,N}|\mathcal{F}_N\}. \end{aligned} \tag{39}$$

If the regression coefficient $\hat{\beta}_{2\pi,y,x}$ is a consistent estimator of

$$\beta_{\mathrm{GLS}} = [V(\bar{\boldsymbol{x}}_{1\pi} - \bar{\boldsymbol{x}}_N)]^{-1}C(\bar{y}_{1\pi}, (\bar{\boldsymbol{x}}_{1\pi} - \bar{\boldsymbol{x}}_N)'), \tag{40}$$

the covariance term is zero. Also, the covariance is zero if both phases are stratified samples with the same sampling units and if indicators for the first-phase strata are elements of $\boldsymbol{x}$.

### 2.4. Variance estimation

We consider two approaches for estimating the variance of $\bar{y}_{2,\mathrm{reg}}$. In the first approach, the terms in the variance of the linearized variance are replaced with sample quantities. The second variance estimator is a replication variance estimator. A replication variance estimator is useful when the data set is released to practitioners and when different sampling units are used at different phases, whereas the linearization variance is often easier to compute when only a few quantities are of interest.

A variance estimator for $\bar{y}_{2,\mathrm{reg}} - \bar{y}_N$ when $C\{\bar{e}_{2\pi}, (\bar{\boldsymbol{x}}_{1\pi} - \bar{\boldsymbol{x}}_N)\beta_{y,x,N}|\mathcal{F}_N\}$ is approximately zero is

$$\hat{V}\{\bar{y}_{2,\mathrm{reg}}|\mathcal{F}_N\} = \hat{\beta}'_{2\pi,y,x}\hat{V}_1\{\bar{\boldsymbol{x}}_{1\pi}|\mathcal{F}_N\}\hat{\beta}_{2\pi,y,x} + \hat{V}\{\bar{e}_{2\pi}|\mathcal{F}_N\}, \tag{41}$$

where $\hat{V}_1\{\bar{\boldsymbol{x}}_{1\pi}|\mathcal{F}_N\}$ is an estimator of the variance of the first-phase mean of $\boldsymbol{x}$ and $\hat{V}\{\bar{e}_{2\pi}|\mathcal{F}_N\}$ is an estimator of the variance of the second-phase mean of $e_i$ constructed with the estimated errors $\hat{e}_i = y_i - \boldsymbol{x}_i\hat{\beta}_{2\pi,y,x}$. The construction of $\hat{V}\{\bar{e}_{2\pi}|\mathcal{F}_N\}$ requires the unconditional joint probabilities and can be difficult for some designs. In some cases, it is reasonable to treat the second-phase sampling as Poisson sampling.

Then, the unconditional joint probability, $\pi_{ij}$, is the first-phase joint probability, $\pi_{1ij}$ multiplied by the product $\pi_{2i|1i}\pi_{2j|1j}$. Alternatively, one can use the variance expression for the DEE to construct $\hat{V}\{\bar{e}_{2\pi}|\mathcal{F}_N\}$. For the variance of the DEE, one approximates $E[V\{\bar{e}_{2\pi}|A_1, \mathcal{F}_N\}|\mathcal{F}_N]$ with an estimator of the conditional second-phase variance $\hat{V}\{\bar{e}_{2\pi}|A_1, \mathcal{F}_N\}$ and adds an estimator of $V\{\bar{e}_{1\pi}|\mathcal{F}_N\}$. See Särndal et al. (1992, pp. 347–350).

There are as many consistent replication variance estimators for the variance of the regression estimator as there are replicate variance estimators for the first-phase estimated mean of $\boldsymbol{x}$. Let a replication variance estimator for the variance of the first-phase total estimator of $\boldsymbol{x}$ be

$$\hat{V}_1\{\hat{\boldsymbol{T}}_{x,1}\} = \sum_{k=1}^{L} c_k\left(\hat{\boldsymbol{T}}_{x,1}^{(k)} - \hat{\boldsymbol{T}}_{x,1}\right)'\left(\hat{\boldsymbol{T}}_{x,1}^{(k)} - \hat{\boldsymbol{T}}_{x,1}\right), \tag{42}$$

where $\hat{\boldsymbol{T}}_{x,1}^{(k)}$ is the $k$th replicate of the estimated total, $\hat{\boldsymbol{T}}_{x,1}$ is the Horvitz–Thompson first-phase total estimator, $L$ is the number of replicates, and $c_k$, $k = 1, 2, \ldots, L$, are constants determined by the replication method and sample design. Note that the jackknife, bootstrap, and balanced repeated replication procedures have replication variance estimators of the form (42). Assume the first-phase replication variance estimator is design consistent,

$$E\left\{\left[\left(V\{\hat{T}_{z,1}|\mathcal{F}_N\}\right)^{-1}\hat{V}_1\{\hat{T}_{z,1}\} - 1\right]^2 |\mathcal{F}_N\right\} = o_p(1), \tag{43}$$

where $z$ is any variable with bounded fourth moments. Let a replicate for the two-phase sample be created from a replicate of the first-phase estimator by applying the first-phase replicate creation rules to the second-phase elements. Thus, if a first-phase replicate is created by deleting a first-phase primary sampling unit, all second-phase units that were in the removed first-phase unit are removed from the second-phase sample. Then, a replicate of the two-phase regression estimator for the mean is

$$\bar{y}_{2,\text{reg}}^{(k)} = \bar{y}_{2\pi}^{(k)} + (\bar{\boldsymbol{x}}_{1\pi}^{(k)} - \bar{\boldsymbol{x}}_{2\pi}^{(k)})\hat{\beta}_{2\pi,y,x}^{(k)}, \tag{44}$$

where $(\bar{y}_{2\pi}^{(k)}, \bar{\boldsymbol{x}}_{2\pi}^{(k)}, \hat{\beta}_{2\pi,y,x}^{(k)})$ is computed using the second-phase sample of replicate $k$, and $\bar{\boldsymbol{x}}_{1\pi}^{(k)}$ is computed from the first-phase sample of replicate $k$. The replicate variance estimator is

$$\hat{V}_2\{\bar{y}_{2,\text{reg}}\} = \sum_{k=1}^{L} c_k\left(\bar{y}_{2,\text{reg}}^{(k)} - \bar{y}_{2,\text{reg}}\right)^2, \tag{45}$$

where the $c_k$ is that of (42).

Under mild assumptions on the first-phase design and given second-phase fixed-rate stratified sampling, the replication variance estimator satisfies

$$\hat{V}_2\{\bar{y}_{2,\text{reg}}\} = V\{\bar{y}_{2,\text{reg}}|\mathcal{F}_N\} - N^{-2}\sum_{i=1}^{N} w_{2i|1i}(1 - w_{2i|1i})e_i^2 + o_p(n_{1N}^{-1}), \tag{46}$$

where $e_i = y_i - \bar{y}_k - (x_i - \bar{x}_k)\beta_N$. The second term on the right side of (46) is small relative to $V\{\bar{y}_{2,\text{reg}}|\mathcal{F}_N\}$ if the first-phase sampling rate is small. The second term on the right side of (46) can be estimated using a DEE replacing $e_i$ with $\hat{e}_i$ or by creating additional replicates.

To prove (46), it is convenient to consider the second-phase sample to have been identified prior to selecting the first-phase sample. Fay (1996) describes this conceptual framework for the case of missing data. See Shao and Steel (1999) and Rao and Shao (1992) for applications of this framework. Let $\{a_{2i}\}$ be the second-phase sample indicators, where $a_{2i} = 1$ if element $i$ is to be in the second-phase sample and $a_{2i} = 0$ otherwise. Let the first-phase sample be selected from the population of $(a_{2i}, a_{2i}y_i, a_{2i}\boldsymbol{x}_i)$ vectors. Write the variance as

$$V_1\{\bar{y}_{2,\text{reg}}|\mathcal{F}_N\} = E\{V_1[\bar{y}_{2,\text{reg}}|(\boldsymbol{a}_2, \mathcal{F}_N)]|\mathcal{F}_N\} + V\{E[\bar{y}_{2,\text{reg}}|(\boldsymbol{a}_{2N}, \mathcal{F}_N)]|\mathcal{F}_N\}, \tag{47}$$

where $\boldsymbol{a}_{2N}$ is the $N$ dimensional vector of second-phase sample indicators. By (43), for second-phase Poisson sampling, and under regularity conditions, the first term on the right side of (47) is consistent for the variance of $\bar{y}_{2,\text{reg}}$. The second term on the right side of (47) is responsible for the second term on the right of the equality in (46). Using the arguments of Hájek (1960), result (46) for Poisson sampling can be extended to stratified second-phase sampling. Kim et al. (2006) provide details. for the DEE and REE.

## 2.5. A central limit theorem

Consistent variance estimation requires moments for the variables and some restrictions on the design. Additional assumptions are required to obtain a central limit theorem for two-phase estimators. Let $\{(y_i, \boldsymbol{x}_i)\}$ be a sequence of *iid* random variables with fifth moments, where the $\boldsymbol{x}_i$ contains a set of $G$ second-phase stratum indicators. Let $\{\mathcal{F}_k, A_{1k}\}$ be a sequence of populations and first-phase samples such that $\mathcal{F}_k \subset \mathcal{F}_{k+1}$, $A_{1k} \subset A_{1,k+1}$, and $\mathcal{F}_k$ contain the first $N_k$ elements of $\{y_i, \boldsymbol{x}_i\}$. Simple random sampling, Poisson sampling, and stratified random sampling can satisfy the postulated framework.

Assume

$$\left[V_1\{(\bar{y}_{1\pi}, \bar{\boldsymbol{x}}_{1\pi})|\mathcal{F}_k\}\right]^{-1/2}\{(\bar{y}_{1\pi}, \bar{\boldsymbol{x}}_{1\pi}) - (\bar{y}_k, \bar{\boldsymbol{x}}_k)\}|\mathcal{F}_k \xrightarrow{\mathcal{L}} N(\boldsymbol{0}, \boldsymbol{I})\ a.s., \tag{48}$$

where the almost sure convergence is with respect to the sequence of populations and samples. Assume the first-phase probabilities satisfy

$$K_L < n_{1k}^{-1} N_k \pi_{1ik} < K_M, \tag{49}$$

where $K_L$ and $K_M$ are positive constants. To ensure that the variance of the second-phase mean converges almost surely, assume the first-phase design is such that

$$\lim_{k\to\infty} N_k^{-1} \sum_{i\in A_{1k}} \pi_{1i}^{-1}(1, \boldsymbol{x}_i y_i, y_i^2)'(1, \boldsymbol{x}_i y_i, y_i^2) = \boldsymbol{H}\ a.s., \tag{50}$$

for some matrix of constants $\boldsymbol{H}$. Then,

$$\left[V_1\{\bar{y}_{2,st}|\mathcal{F}_k\}\right]^{-1/2}(\bar{y}_{2,st}-\bar{y}_k)\xrightarrow{\mathcal{L}}N(0,1), \tag{51}$$

where $\bar{y}_{2,st}$ is defined in (2) and $N_k^{-2}\left[V_1\{\bar{y}_{2,st}|\mathcal{F}_k\}\right]$ is the variance of the DEE in (5).

The variance in (51) can be replaced by a consistent estimator such as the replication variance estimator. The central limit theorem can be extended to the regression estimator since the regression estimator is a smooth function of first- and second-phase means. See Legg (2006) for an extension to a generalized least squares estimator. See Chen and Rao (2006) for a proof under alternative assumptions.

## 3. Three-phase sampling

The ideas and estimators for two-phase sampling can be extended to three or more phases. Let $A_3$ be the set of indices selected from $A_2$ with conditional inclusion probabilities $\pi_{3i|2i}=\Pr(\text{unit } i\in A_3|\text{unit } i\in A_2\bigcap A_1)$. Let $w_{3i|2i}=\pi_{3i|2i}^{-1}$. The DEE is generalized to a triple expansion estimator by multiplying the double expansion weight by $w_{3i|2i}$ to obtain a three-phase expansion estimator for a total,

$$\hat{T}_{y,3}=\sum_{i\in A_3}w_{3i|2i}w_{2i|1i}w_{1i}y_i. \tag{52}$$

The two-phase regression estimator can be extended to three or more phases by using a sequence of regression estimators. Suppose $(1,\boldsymbol{u}_1)$ is observed on the first-phase sample, $(1,\boldsymbol{u},\boldsymbol{x})$ is observed on the second-phase sample, and $(1,\boldsymbol{u},\boldsymbol{x},y)$ is observed on the third-phase sample. The first- and second-phase samples can be used to construct a regression estimator for the mean of $\boldsymbol{x}$. Let

$$\bar{\boldsymbol{x}}_{2,\text{reg}}=\bar{\boldsymbol{x}}_{2\pi}+(\bar{\boldsymbol{u}}_{1\pi}-\bar{\boldsymbol{u}}_{2\pi})\hat{\beta}_{2\pi,x,u}, \tag{53}$$

where

$$\hat{\beta}_{2\pi,x,u}=\left[\sum_{i\in A_2}(\boldsymbol{u}_i-\bar{\boldsymbol{u}}_{2\pi})'w_{2i|1i}w_{1i}(\boldsymbol{u}_i-\bar{\boldsymbol{u}}_{2\pi})\right]^{-1}\sum_{i\in A_2}(\boldsymbol{u}_i-\bar{\boldsymbol{u}}_{2\pi})'w_{2i|1i}w_{1i}(\boldsymbol{x}_i-\bar{\boldsymbol{x}}_{2\pi}), \tag{54}$$

$\bar{\boldsymbol{x}}_{2\pi}$ and $\bar{\boldsymbol{u}}_{2\pi}$ are second-phase Hájek means, and $\bar{\boldsymbol{u}}_{1\pi}$ is a first-phase Hájek mean. A regression coefficient for the regression of $(\boldsymbol{u},\boldsymbol{x})$ on $y$ using third-phase observations is then

$$\hat{\beta}_{3\pi,y,(u,x)}=\left[\sum_{i\in A_3}\boldsymbol{d}_{3i}'w_{3i|2i}w_{2i|1i}w_{1i}\boldsymbol{d}_{3i}\right]^{-1}\sum_{i\in A_3}\boldsymbol{d}_{3i}'w_{3i|2i}w_{2i|1i}w_{1i}(y_i-\bar{y}_{3\pi}), \tag{55}$$

where $\boldsymbol{d}_{3i}=(\boldsymbol{u}_i-\bar{\boldsymbol{u}}_{3\pi},\boldsymbol{x}_i-\bar{\boldsymbol{x}}_{3\pi})$ and $\bar{\boldsymbol{u}}_{3\pi}$, $\bar{\boldsymbol{x}}_{3\pi}$, and $\bar{y}_{3\pi}$ are three-phase Hájek means. The three-phase regression estimator for the mean of $y$ is

$$\bar{y}_{3,\text{reg}}=\bar{y}_{3\pi}+(\bar{\boldsymbol{u}}_{1\pi}-\bar{\boldsymbol{u}}_{3\pi},\bar{\boldsymbol{x}}_{2,\text{reg}}-\bar{\boldsymbol{x}}_{3\pi})\hat{\beta}_{3,y,(u,x)}. \tag{56}$$

The variance of the three-phase regression estimator can be expressed using conditional expectations and variances. The variance of $\bar{y}_{3,\mathrm{reg}}$ is

$$\begin{aligned} V\{\bar{y}_{3,\mathrm{reg}}|\mathcal{F}_N\} = {} & V\{E[E(\bar{y}_{3,\mathrm{reg}}|\mathcal{F}_N, A_1, A_2)|\mathcal{F}_N, A_1]|\mathcal{F}_N\} \\ & + E\{V[E(\bar{y}_{3,\mathrm{reg}}|\mathcal{F}_N, A_1, A_2)|\mathcal{F}_N, A_1]|\mathcal{F}_N\} \\ & + E\{E[V(\bar{y}_{3,\mathrm{reg}}|\mathcal{F}_N, A_1, A_2)|\mathcal{F}_N, A_1]|\mathcal{F}_N\}. \end{aligned} \tag{57}$$

The first two terms on the right side of the variance expression are difficult to estimate since they involve variances of unobserved means. As with the two-phase variance estimator, under certain designs, forms for the first two terms in (57) exist for which method of moments estimators from the third-phase sample can be used. The third term in the variance expression can be directly estimated using the conditional third-phase variance. Alternatively, variance estimators can be formed by extending the replication variance estimator to three phases or by estimating the terms in a Taylor linearization expression for (57).

For a linearization procedure, write the error in $\bar{y}_{3,\mathrm{reg}}$ as

$$\bar{y}_{3,\mathrm{reg}} - \bar{y}_N = \bar{\boldsymbol{e}}_{3\pi} + (\bar{\boldsymbol{u}}_{1\pi} - \bar{\boldsymbol{u}}_N, \bar{\boldsymbol{a}}_{2\pi})\beta_{3,y,(u,a),N} + O_p(n_3)^{-1}, \tag{58}$$

where $\bar{\boldsymbol{a}}_N = \mathbf{0}$,

$$e_{Ni} = y_i - \bar{y}_N - (\boldsymbol{u}_i - \bar{\boldsymbol{u}}_N, \boldsymbol{x}_i - \bar{\boldsymbol{x}}_N)\beta_{3,y,(u,a),N}, \tag{59}$$

$$a_{Ni} = \boldsymbol{x}_i - \bar{\boldsymbol{x}}_N - (\boldsymbol{u}_i - \bar{\boldsymbol{u}}_N)\beta_{2,x,u,N}, \tag{60}$$

and $\beta_{3,y,(u,a),N}$ and $\beta_{2,x,u,N}$ are population regression coefficients. If the three residual means, $\bar{e}_{3\pi}$, $\bar{\boldsymbol{u}}_{1\pi}$, and $\bar{\boldsymbol{a}}_{2\pi}$, are uncorrelated,

$$\begin{aligned} V\{\bar{y}_{3,\mathrm{reg}} - \bar{y}_N|\mathcal{F}_N\} = {} & V\{\bar{e}_{3\pi}|\mathcal{F}_N\} + \beta'_{y,u,N}V\{\bar{\boldsymbol{u}}_{1\pi}|\mathcal{F}_N\}\beta_{y,u,N} \\ & + \beta'_{y,a,N}V\{\bar{\boldsymbol{a}}_{2\pi}|\mathcal{F}_N\}\beta_{y,a,N}, \end{aligned} \tag{61}$$

where $\beta'_{y,(u,a),N} = (\beta'_{y,u,N}, \beta'_{y,a,N})$. For variance estimation, the regression coefficients and residuals are replaced by their sample estimators. As with two-phase variance estimation, Poisson sampling approximations or the conditional variance approximation from the DEE can be applied.

For more than three phases, one can create a regression estimator for each of the auxiliary variables using the largest data set for which the auxiliary variable is fully observed. The regression estimator for the mean of $y$ is then the Hájek mean estimator for $y$ adjusted by the differences between the regression estimators and the Hájek mean estimators of the auxiliary variables. An important special case of the three-phase estimator is that in which the first-phase sample is the finite population.

## 4. Two-phase estimation illustration

We use data from the U.S. National Resources Inventory (NRI) to illustrate some statistics associated with two-phase sampling. The NRI is conducted by the U.S. Natural Resources Conservation Service in cooperation with the Iowa State University Center for Survey Statistics and Methodology. The survey is a panel survey of land use and was conducted in 1982, 1987, 1992, 1997, and has been conducted yearly since 2000. Data are collected on soil characteristics, land use, land cover, wind erosion, water erosion, and conservation practices. The sample is a stratified area sample of the United States, where the primary sampling units are areas of land called segments. Data are collected for the entire segment on such items as urban lands, roads, and water. Detailed data on soil properties and land use are collected at a random sample of points within the segment. The sample for 1997 contained about 300,000 segments with about 800,000 points. The yearly samples are typically about 70,000 segments. See Nusser and Goebel (1997) for a more complete description of the survey.

The data in Table 1 are second-phase data for a county in Missouri for the year 2003. In Missouri, segments are defined by the Public Land Survey System. Therefore, most segments are close to 160 acres in size, but there is some variation in size due to variation in sections defined by the Public Land Survey System and due to truncation associated with county boundaries. The segment size in acres is given in column three of the table. Typically three points are observed in each segment. The points are classified using a system called broaduse, where example broad uses are urban land, cultivated cropland, pastureland, and forestland. Some of the broad uses are further subdivided into categories called coveruses, where corn, cotton, and soybeans are some of the coveruses within the cropland broaduse. In this example, we estimate acres of cultivated cropland. Segments were placed into first-phase strata based on location. We aggregated several first-phase strata to form the three first-phase strata identified

Table 1
Second-phase sample data

| $h$ | $g$ | Segment Size | $w_{1i}$ | $w_{2i\|1i}$ | 1997 Cultivated Cropland | 2003 Cultivated Cropland | 1997 Federal |
|---|---|---|---|---|---|---|---|
| 1 | 1 | 161 | 33 | 2.200 | 53.67 | 0.00 | 0.00 |
| 1 | 1 | 148 | 33 | 2.200 | 49.33 | 49.33 | 0.00 |
| 1 | 1 | 163 | 33 | 2.200 | 0.00 | 0.00 | 0.00 |
| 1 | 2 | 164 | 33 | 3.400 | 109.33 | 109.33 | 0.00 |
| 1 | 2 | 161 | 33 | 3.400 | 0.00 | 0.00 | 107.33 |
| 1 | 3 | 164 | 33 | 4.875 | 109.33 | 109.33 | 0.00 |
| 1 | 3 | 161 | 33 | 4.875 | 53.67 | 53.67 | 0.00 |
| 1 | 3 | 323 | 33 | 4.875 | 107.67 | 107.67 | 0.00 |
| 1 | 4 | 167 | 33 | 3.250 | 0.00 | 0.00 | 0.00 |
| 2 | 1 | 165 | 35 | 2.200 | 110.00 | 55.00 | 0.00 |
| 2 | 3 | 154 | 35 | 4.875 | 102.67 | 102.67 | 0.00 |
| 2 | 3 | 161 | 35 | 4.875 | 161.00 | 161.00 | 0.00 |
| 2 | 3 | 159 | 35 | 4.875 | 53.00 | 0.00 | 0.00 |
| 2 | 3 | 163 | 35 | 4.875 | 108.67 | 108.67 | 0.00 |
| 3 | 1 | 160 | 26 | 2.200 | 0.00 | 0.00 | 0.00 |
| 3 | 2 | 161 | 26 | 3.400 | 0.00 | 0.00 | 0.00 |
| 3 | 2 | 164 | 26 | 3.400 | 54.67 | 54.67 | 0.00 |
| 3 | 2 | 164 | 26 | 3.400 | 0.00 | 0.00 | 0.00 |
| 3 | 3 | 156 | 26 | 4.875 | 52.00 | 52.00 | 0.00 |
| 3 | 4 | 165 | 26 | 3.250 | 0.00 | 0.00 | 110.00 |
| 3 | 4 | 100 | 26 | 3.250 | 0.00 | 0.00 | 50.00 |
| 3 | 4 | 176 | 26 | 3.250 | 0.00 | 0.00 | 0.00 |

in column one of the table. The 1997 sampling weights are given in column four of the table. The inverses of the first-phase weights are small. Therefore, the first-phase finite population corrections can be ignored. The estimated segment acres of cropland for 1997 are given in column six. This number is the fraction of the points that are cultivated cropland multiplied by the segment size.

The 2003 NRI sample is a stratified second-phase sample selected from the 1997 sample. The segments in the 1997 sample were placed in strata for the second-phase (2003) sample on the basis of characteristics such as broaduses in the segment. For the purposes of this example, we use four second-phase strata. The second-phase stratum of segment $i$ is given in column two of Table 1. The population and sample segment counts for the first-phase sample are given in Table 2. The $(n_{1g}, n_{2g})$ for the second-phase strata are (11, 5), (17, 5), (39, 8), and (13, 4) for $g = 1, 2, 3$, and 4, respectively. The objective is to estimate the total acres of cultivated cropland for the county in 2003 and the change in acres of cultivated cropland from 1997 to 2003. Estimates of totals are given in thousands of acres.

The DEE estimate of 2003 cultivated cropland is 138.6 thousand acres. We compute an estimate of the variance of the DEE following the example in Section 1.1. The vector of first-phase estimated population variances computed by (12) is $(\hat{S}^2_{y1}, \hat{S}^2_{y2}, \hat{S}^2_{y3}) = (2.278 \times 10^{-3}, 3.190 \times 10^{-3}, 0.599 \times 10^{-3})$. The estimated variance of the first-phase total is 204.1. The estimated variances defined in (16) are $(\hat{S}^2_{1wy,1}, \hat{S}^2_{1wy,2}, \hat{S}^2_{1wy,3}, \hat{S}^2_{1wy,4}) = (0.957, 2.495, 3.112, 0.000)$. Then, $\hat{V}\{\hat{T}_{2,y}|A_1, \mathcal{F}_N\} = 584.7$, and the standard error of the DEE is 28.1. The DEE of change in acres of cultivated cropland from 1997 to 2003 is $-17.2$ with a standard error of 10.5.

The REE for 2003 cultivated cropland is computed as a two-phase regression estimator, where the vector of auxiliary variables is the vector of second-phase stratum indicators. That is,

$$x_{gi} = \begin{cases} 1 & \text{if segment } i \text{ is in second-phase stratum } g \\ 0 & \text{otherwise,} \end{cases} \tag{62}$$

for $g = 1, 2, 3, 4$. Let $\boldsymbol{x}_i$ be $(x_{1i}, x_{2i}, x_{3i}, x_{4i})$. The estimated regression coefficient is

$$\begin{aligned} \hat{\beta}_{2,y,x} &= \left(\textstyle\sum_{i \in A_2} w_{2i|1i} w_{1i} \boldsymbol{x}_i' \boldsymbol{x}_i\right)^{-1} \textstyle\sum_{i \in A_2} w_{2i|1i} w_{1i} \boldsymbol{x}_i' y_i \\ &= (22.2, 34.9, 88.0, 0.0)', \end{aligned} \tag{63}$$

where $\hat{\beta}_{2,y,x}$ is the vector of second-phase stratum weighted means. The first-phase total estimator for $\boldsymbol{x}$ is (366, 519, 1282, 420) segments. The REE for 2003 cultivated cropland is 139.0. To compute the variance of the REE, we treat the REE as an estimator of the

Table 2
First-phase counts

| Stratum | $N_h$ | $n_{1h}$ | $n_{2h}$ |
|---|---|---|---|
| 1 | 990 | 30 | 9 |
| 2 | 1155 | 33 | 5 |
| 3 | 442 | 17 | 8 |

first-phase Horvitz–Thompson total. Let

$$\hat{e}_{ri} = y_i - \boldsymbol{x}_i \hat{\beta}_{2,y,x} \tag{64}$$

be the residual from the two-phase regression. The variance estimator is

$$\hat{V}\{\hat{T}_{y,r,2}\} = \hat{V}\{\hat{T}_{y,1} | \mathcal{F}_N\} + \hat{V}\{\hat{T}_{e,r,2} | A_1, \mathcal{F}_N\}, \tag{65}$$

where $\hat{T}_{e,r,2}$ is the DEE of $\hat{e}_{ri}$. The $\hat{V}\{\hat{T}_{y,1}|\mathcal{F}_N\}$ component is the same as for the DEE. The $\hat{V}\{\hat{T}_{e,r,2}|A_1, \mathcal{F}_N\}$ component is estimated using the same procedure as for the conditional variance component in the DEE variance estimator replacing $w_{1i}y_i - \bar{y}_{2w,g}$ with $w_{1i}\hat{e}_{ri}$. The estimate $\hat{V}\{\hat{T}_{e,r,2}|A_1, \mathcal{F}_N\}$ is 739.4, which gives a standard error of the REE of 27.2. For change in cultivated cropland, the REE is $-17.4$ and the standard error of the REE is 10.4. For this very small illustration, there is a modest difference between the standard errors of the REE and the DEE.

The estimated segment acres of cultivated cropland in 1997 is highly correlated with the estimated segment acres of cultivated cropland in 2003. Therefore, we consider the regression estimator formed by adding 1997 cultivated cropland to the $\boldsymbol{x}$ vector used in the REE. The first-phase total estimate for cultivated cropland in 1997 is 153.7, and the regression estimate for cultivated cropland in 2003 is 136.2. The variance is computed using the same procedure used for the REE, where residuals for the expanded regression are used for the conditional variance component. The standard error for the 2003 cultivated cropland regression estimator is 16.7. For change in cultivated cropland, the regression estimate is $-17.5$ with a standard error of 10.3. Only three segments in the second-phase sample have a change in cultivated cropland. Therefore, including 1997 cultivated cropland in the regression model affects estimates of change much less than it affects estimates of level.

The acres of federal land and total acres in the county are known and used as controls in the NRI. For our county, the total acres of federal land is 27.2 and the total acres is 437.1. We incorporate the control totals into the regression estimator using the three-phase regression procedure. The first-phase is the population, but the calculations are a special case of three-phase estimation. We first create a regression estimate of the totals of the $\boldsymbol{x}$ vector composed of second-phase stratum indicators and 1997 cultivated cropland. Let $\boldsymbol{u}$ be the vector of first-phase stratum indicators, segment acres, and federal acres. Using first-phase data and weights $w_{1i}$, we construct the regression estimator of the total of $\boldsymbol{x}$ using $\boldsymbol{u}$ as the vector of auxiliary variables. The estimated totals are (392, 520, 1282, 393, 156.9), where 156.9 is the estimate for 1997 acres of cultivated cropland. Let $\boldsymbol{z}$ be the vector of second-phase stratum indicators, 1997 cultivated cropland, segment acres, and federal acres. Using the second-phase sample and weights $w_{2i|1i}w_{1i}$, we regress $y$ on $\boldsymbol{z}$ to obtain $\hat{\beta}_{3\pi,y,z}$. The regression estimate for 2003 cultivated cropland is $\hat{T}_{y,\text{reg},3} = (392, 520, 1282, 393, 156.9, 437.1, 27.2)(-29.98, -7.00, -15.59, -6.07, 1.02, 0.038, 0.0075)' = 138.8$.

We construct a variance estimator that is comparable to that used for the two-phase regression estimators. Because the regression estimator is used for the first-phase totals, the relevant variance for the complete first-phase sample is that of the deviation from the regression of $y$ on $\boldsymbol{u}$. Let $\hat{e}_{1\text{reg},i}$ be the residual from regressing $y$ on $\boldsymbol{u}$ using second-phase data and $w_{2i|1i}w_{1i}$ weights. The relevant deviation for the second-phase conditional variance is the residual from regressing $y$ on $\boldsymbol{z}$ using $w_{2i|1i}w_{1i}$ weights, denoted by $\hat{e}_{2\text{reg},i}$.

Then, an estimator for the variance of $\hat{T}_{y,\text{reg},3}$ is

$$\hat{V}\{\hat{T}_{y,\text{reg},3}|\mathcal{F}_N\} = \hat{V}\{\hat{T}_{e,1\text{reg},1}|\mathcal{F}_N\} + \hat{V}\{\hat{T}_{e,2\text{reg},2}|A_1, \mathcal{F}_N\}, \tag{66}$$

where $\hat{T}_{y,\text{reg},3}$ is the estimator of the total, $\hat{V}\{\hat{T}_{e,1\text{reg},1}|\mathcal{F}_N\}$ is the estimated variance of the first-phase total of $e$, and $\hat{V}\{\hat{T}_{e,2\text{reg},2}|A_1, \mathcal{F}_N\}$ is the estimated conditional variance of the second-phase total of $e$,

$$e_i = y_i - \boldsymbol{u}_i\beta_1, \tag{67}$$

and

$$\beta_1 = \left(\sum_{i\in U} \boldsymbol{u}_i'\boldsymbol{u}_i\right)^{-1} \sum_{i\in U} \boldsymbol{u}_i' y_i. \tag{68}$$

The first-phase component for 2003 cultivated cropland is

$$\begin{aligned}\hat{V}\{\hat{T}_{e,1\text{reg},1}|\mathcal{F}_N\} &= \sum\nolimits_{i\in A_2} \left[(n_{2gi})(n_{2gi}-1)^{-1}(n_2-3)(n_2-5)^{-1}\right] w_{2i|1i} w_{1i}^2 \hat{e}_{1\text{reg},i}^2 \\ &= 166.2, \end{aligned} \tag{69}$$

where $n_{2gi}$ is the second-phase stratum sample size for the stratum containing segment $i$ and $n_2$ is the sum of the second-phase stratum sample sizes. The estimator is nearly equal to the variance estimator obtained from using only first-phase strata. The $\hat{V}\{\hat{T}_{e,2\text{reg},2}|A_1, \mathcal{F}_N\}$ component is constructed in the same manner as the conditional variance for the REE with $\hat{e}_{2\text{reg},i}$ in place of $\hat{e}_{r,i}$. For 2003 cultivated cropland, $\hat{V}\{\hat{T}_{e,2\text{reg},2}|A_1, \mathcal{F}_N\} = 73.7$ and the standard error of $\hat{T}_{y,\text{reg},3}$ is 15.5. The inclusion of the control variables in the regression model reduces the variance only slightly. The segment acres and federal land provide little information on 2003 cultivated cropland beyond that in 1997 cultivated cropland. However, the regression weights from the three-phase regression give correct total "estimates" for total acres and federal acres. Also, the second-phase regression weights reproduce the first-phase estimate of 1997 cultivated cropland.

*Sample Surveys: Design, Methods and Applications, Vol. 29A*
ISSN: 0169-7161

DOI: 10.1016/S0169-7161(08)00004-7

# Multiple-Frame Surveys

*Sharon L. Lohr*

## 1. What are multiple-frame surveys, and why are they used?

Suppose you want to take a survey of statisticians in the United States. One approach might be to use the membership directory of the American Statistical Association (ASA) as a sampling frame and to take a probability sample of the persons on the membership list. Most of the persons contacted in this sample would likely be statisticians, so the frame would be cost-effective for contacting statisticians. However, many statisticians in the United States do not belong to the ASA, and you will have undercoverage of the population of statisticians if you sample only from the ASA membership directory.

You could improve coverage by taking an additional sample from a second sampling frame, such as the membership directory of the Institute of Mathematical Statistics (IMS). Thus, you could take a probability sample from frame A (the ASA directory), and independently take a probability sample from frame B (the IMS directory). The two frames overlap, since many statisticians belong to both organizations. As shown in Fig. 1, there are three population domains: domain *a* consists of the population units that are in frame A but not in frame B, domain *b* consists of the population units in frame B but not in frame A, and domain *ab* consists of the population units that are in both frames A and B.

Frames A (the ASA membership directory) and B (the IMS membership directory) have better coverage of the population but still do not include all statisticians. In this case, a three-frame survey could be used, where frame C might be a frame of the entire adult population. The population structure with three frames is shown in Fig. 2. This multiple-frame design has four domains: domain *c* is the part of the population in the complete frame C but not in either membership directory.

Since frame C contains the entire population, you might ask why you would go to the extra work of taking a multiple-frame survey: why wouldn't you just take a sample from frame C? The answer is because of cost: frame C, containing the entire population, is very expensive to sample from. Since most adults are not statisticians, many sampling designs from frame C will yield very few statisticians. Frames A and B, by contrast, are much less expensive to sample from. But they do not include the entire population of interest. By combining samples from frames A, B, and C, you can exploit the inexpensiveness of frames A and B while avoiding bias by including the missing part of the population—domain *c*—in frame C.

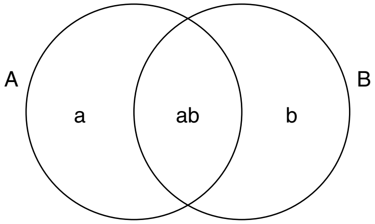


Fig. 1. Overlapping frames A and B and three domains.

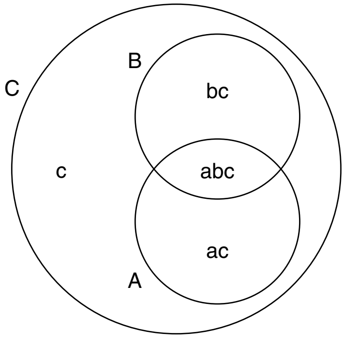


Fig. 2. Frame C contains the entire population; frames A and B overlap and are both contained in frame C.

There are several reasons to consider using a multiple-frame survey for data collection. Multiple-frame surveys can greatly improve efficiency of data collection when sampling a rare population. A rare population is a subgroup of interest that comprises only a small part, usually 10% or less, of the full population. In the example discussed above, statisticians are a small segment of the adult population; by supplementing the general population survey with samples from the membership directories, the number of statisticians in the data set is increased. A survey concerned with a specific population such as persons with a rare disease may obtain a larger sample size by sampling from additional frames that contain a high concentration of persons in the desired category. Other methods used for sampling rare populations are discussed in Chapter 6.

Multiple-frame surveys can be used when different modes of data collection must be used to reach segments of the population; for example, frame A in Fig. 1 might be a frame of landline telephones and frame B might consist of cellular telephone numbers. Hartley (1962), when introducing the theory of multiple-frame surveys, wrote that the main reason to consider a multiple-frame survey is to reduce data collection costs while still sampling from the entire target population:

> In sample survey methodology one often finds that a frame known to cover approximately *all* units in the population is one in which sampling is costly while other frames (e.g. special lists of units) are available for cheaper sampling methods. However, the latter usually only cover an unknown or only approximately known fraction of the population. This paper develops a general methodology of utilizing

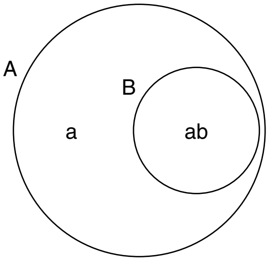


Fig. 3. Frame B is a subset of frame A.

> any number of such frames without requiring any prior knowledge as to the extent of their mutual overlap. (Hartley, 1962, p. 203)

The situation considered by Hartley (1962) is depicted in Fig. 3. This is the most common situation for which multiple-frame sampling is used. Often, frame B is a list frame; for example, in an agricultural survey, frame B might be a list of the large farms in a country. The list may be out of date, however, and not include farms that were recently established. In addition, the list will not include smaller agricultural holdings. Although efficient data collection methods can be used in frame B, the incomplete coverage of the population means that estimates of the total amount of land planted to soybeans using only the sample from frame B will likely be too small. Using an area frame for frame A gives complete coverage of farms in the country. In an area frame, the country is divided into geographical areas and area segments are sampled. Interviewers travel to each selected segment and collect data from every agricultural holding within the boundaries of the segment. The area frame thus includes all farms in the country, but data collection is much more expensive than in the list frame.

The U.S. Survey of Consumer Finances (Bucks et al., 2006) is one example of a dual-frame survey that accords with Fig. 3. One sample (the frame A sample) is selected from the U.S. population using an area frame with a stratified multistage sampling design. However, the U.S. Federal Reserve Board is also interested in characteristics of households who own investments such as tax-exempt bonds, and the sample from the complete frame will likely contain few households who own rarer asset types. Thus, a second sample is taken from a list frame (frame B) of records constructed from tax return information.

### *1.1. Screening multiple-frame surveys*

Consider the situation in Fig. 3, in which frame A contains the entire population and frame B contains only a subset of the population. In some instances, it is possible to determine which population units in frame A are in the overlap domain *ab* and remove them from frame A before sampling. In a dual-frame agricultural survey, a farm in the list frame would not be interviewed in the area frame sample. This situation, in which duplicated units in the sampling frames are removed before data collection so that the sampling frames are disjoint, is called a *screening* multiple-frame survey. González-Villalobos and

Wallace (1996) discuss examples and estimation methods in screening multiple-frame surveys.

Because the sampling frames have no overlap, population totals are easily estimated in screening multiple-frame surveys by summing the estimated population totals from each frame. Hansen et al. (1953) mentioned an early use of a screening dual-frame survey. They called it "joint sampling from a list and one- or two-stage area sampling" (p. 327), in which there is an incomplete, and possibly outdated, list of members of the target population. The list frame is considered to be stratum 1. The units in the list frame are then removed from the area frame so that the stratum 2 consists of population units that are not in stratum 1. A multistage sample is taken from stratum 2. As Hansen et al. (1953) pointed out, this sampling design is a special case of stratified sampling: every unit in the population is in exactly one stratum, and sampling is carried out independently in the two strata. Optimal allocation methods from stratified sampling can be used to determine the sample sizes in the two strata.

Screening can be done at any stage of data collection. One of the earliest dual-frame surveys was the Sample Survey of Retail Stores, conducted by the U.S. Census Bureau in 1949 (Hansen et al., 1953, p. 516). Primary sampling units (psus) were chosen using a stratified sample of groups of counties. Within each psu, it was not feasible to obtain a complete enumeration of current retail stores because of the high volatility of small businesses. Within each psu, a census of retail firms on a list compiled from the records of the Old Age and Survivors Insurance Bureau was taken; and an area sample was taken of firms not on the list. In this case, a dual-frame design was used within each selected psu. The sampling designs for the two frames were not independent, however, since the two samples shared the same psus. Thus, the estimator of total sales summed the two estimators within psus.

The National Survey of America's Families (Brick et al., 2002) has two components: a random-digit dialing survey of households with telephones and an area sample of households without telephones. Since a primary goal of the survey is to collect data on children in low-income households, it is important to include nontelephone households in the sample: low-income households are disproportionately likely to lack a telephone. After the area sample is selected, households are screened for the presence of a telephone; only households without a telephone are interviewed in frame A. As a result of the screening, frame A consists of households that are not found in frame B.

Even when the frames overlap, a multiple-frame survey can be analyzed as a screening multiple-frame survey by deleting records in the overlap domains. The U.S. National Science Foundation (2003) surveys used in the Scientists and Engineers Statistical Data System (SESTAT) are collected as overlapping three-frame surveys but analyzed as if they were screening three-frame surveys. The target population is noninstitutionalized U.S. residents under age 75 who (1) have at least a bachelor's degree and (2) either have a degree in science or engineering, or are working as a scientist or engineer. The SESTAT data are based on three-component surveys: (1) The National Survey of College Graduates (NSCG), with a sampling frame derived from the U.S. Census long form, which includes persons with at least a bachelor's degree; (2) the National Survey of Recent College Graduates (NSRCG), with a sampling frame of educational institutions, which samples persons who received a degree in science or engineering within the last two years; and (3) the Survey of Doctorate Recipients (SDR), which samples people who received doctoral degrees from a U.S. institution in science or engineering. The sampling frame for the SDR is constructed from the Survey of Earned Doctorates, which is a census of all doctoral degree recipients in the United States.

With the three sampling frames, a person can be included in more than one frame. For example, a person who obtained a bachelor's degree before the census year and then completed a doctoral degree would be included in all three sampling frames. The sampling frames for SESTAT are thus depicted in Fig. 2, with C the frame for the NSCG, B the frame for the NSRCG, and A the frame for the SDR. The multiplicity is not used in the analysis of the SESTAT data, however. Instead, a hierarchy is set up for inclusion in one of the frames. A person sampled in the NSCG is removed from the data analysis (assigned a weight of 0) if he or she could have been sampled in either the NSRCG or the SDR. Similarly, a respondent to the NSRCG is given weight 0 if the respondent is also in the sampling frame for the SDR. As a result, the combined SESTAT database is essentially a stratified sample with three strata: the SDR frame, the part of the NSRCG frame that is outside of the SDR frame, and the part of the NSCG frame that does not overlap with the other two frames.

By deleting records in a hierarchical manner, one can analyze data from overlapping multiple-frame surveys such as SESTAT as stratified samples. Such an approach, however, reduces the sample size and discards data. In the SESTAT surveys, relatively few units sampled in the NSCG are from the overlap domains, and so the loss in efficiency is small. In surveys with greater overlap, however, it is much more efficient to use all the information collected from the overlap domains when estimating population characteristics. Overlapping multiple-frame surveys are discussed in the next section.

### *1.2. Multiple-frame surveys with overlap*

In many multiple-frame surveys, the domain membership of a population unit is not known in advance. For example, if frame A consists of landline telephone numbers and frame B consists of cellular telephone numbers, it is unknown in advance whether a household member sampled using one frame also belongs to the other frame (Brick et al., 2006). The domain membership of a sampled person must be ascertained from the survey questions. In other instances, it may be difficult to match units in the frames and to ascertain whether a person in frame A is actually the same as a person with the same name in frame B.

Haines and Pollock (1998) described the use of multiple-frame surveys to estimate the size of a population and discussed the correspondence between multiple-frame surveys and capture–recapture methods for estimating population size. They combined information from incomplete list frames and an area frame to estimate the number of eagle nests in a region. Iachan and Dennis (1993) used a multiple-frame design to sample the homeless population, where frame A was homeless shelters, frame B was soup kitchens, and frame C consisted of street locations. This situation is depicted in Fig. 4. Although the union of the three frames still misses part of the homeless population, coverage is improved by having more than one frame.

Sometimes independent samples are taken from the same frame so that frames A and B coincide. This might occur if two sampling designs are desired. In a wildlife survey, one sample might be taken using a stratified multistage sample and the other using a sequential or adaptive sampling design.

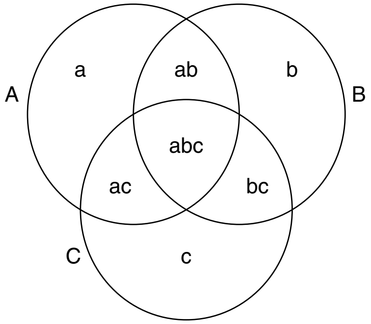


Fig. 4. Frames A, B, and C are all incomplete and overlap.

When the frames overlap, care must be taken when estimating quantities. Individuals in the overlap domains such as *ab* can be selected in either or both surveys, so estimators of population quantities must be adjusted to compensate for that multiplicity. In Section 2, we summarize point estimators that have been proposed for multiple-frame surveys.

## 2. Point estimation in multiple-frame surveys

Classical sampling theory considers a single frame. The universe $\mathcal{U}$ has $N$ units. The population total for a characteristic $y$ is $Y = \sum_{i=1}^{N} y_i$. A probability sampling design is used to select a sample $\mathcal{S}$ from the frame. We denote $\pi_i = P(i \in \mathcal{S})$ as the inclusion probability for the frame. The Horvitz–Thompson estimator of the population total is

$$\hat{Y}_{\text{HT}} = \sum_{i \in \mathcal{S}} w_i y_i,$$

where $w_i = 1/\pi_i$ is the sampling weight.

In this section, we discuss estimation methods for dual-frame surveys when independent samples are taken from the two frames. We consider here the situation of Fig. 1, the most general case. We still have $y_i$ as a characteristic of population unit $i$, and population total $Y = \sum_{i=1}^{N} y_i$. But now two samples are taken, one from the $N_A$ units in frame A and another sample from the $N_B$ units in frame B. Population units in the overlap domain *ab* can be sampled in either survey or both surveys.

Let

$$\delta_i(a) = \begin{cases} 1 & \text{if unit } i \text{ is in domain } a \\ 0 & \text{otherwise} \end{cases},$$

$$\delta_i(b) = \begin{cases} 1 & \text{if unit } i \text{ is in domain } b \\ 0 & \text{otherwise} \end{cases},$$

and

$$\delta_i(ab) = \begin{cases} 1 & \text{if unit } i \text{ is in domain } ab \\ 0 & \text{otherwise} \end{cases}.$$

For the general overlapping dual-frame survey depicted in Fig. 1, we can write the population total as the sum of the three domain totals:

$$\begin{aligned} Y &= \quad Y_a \quad + \quad Y_{ab} \quad + \quad Y_b \\ &= \sum_{i=1}^{N} \delta_i(a) y_i + \sum_{i=1}^{N} \delta_i(ab) y_i + \sum_{i=1}^{N} \delta_i(b) y_i . \end{aligned}$$

An estimator of $Y$ can be formed by summing estimators of $Y_a$, $Y_{ab}$, and $Y_b$. Let $\mathcal{S}_A$ denote the sample from frame A, with probability of inclusion in $\mathcal{S}_A$ denoted by $\pi_i^A = P\{i \in \mathcal{S}_A\}$; $\mathcal{S}_A$ contains $n_A$ observation units. Corresponding quantities for frame B are $\mathcal{S}_B$, $\pi_i^B = P\{i \in \mathcal{S}_B\}$, and $n_B$. Let $w_i^A$ be the sampling weight from frame A and $w_i^B$ be the sampling weight from frame B. These weights can be the inverses of the inclusion probabilities $\pi_i^A$ and $\pi_i^B$ or they can be the Hájek (1981) weights, with $w_i^A = N_A[\pi_i^A \sum_{j \in \mathcal{S}_A}(1/\pi_j^A)]^{-1}$ and $w_i^B = N_B[\pi_i^B \sum_{j \in \mathcal{S}_B}(1/\pi_j^B)]^{-1}$. Using the weights, we define domain estimators from the frame A sample:

$$\begin{aligned} \hat{Y}_a^A &= \sum_{i \in \mathcal{S}_A} w_i^A \delta_i(a) y_i \\ \hat{Y}_{ab}^A &= \sum_{i \in \mathcal{S}_A} w_i^A \delta_i(ab) y_i . \end{aligned}$$

The corresponding estimators from the frame B sample are

$$\begin{aligned} \hat{Y}_b^B &= \sum_{i \in \mathcal{S}_B} w_i^B \delta_i(b) y_i \\ \hat{Y}_{ab}^B &= \sum_{i \in \mathcal{S}_B} w_i^B \delta_i(ab) y_i . \end{aligned}$$

Under standard sampling theory results, each domain estimator is approximately unbiased for the corresponding population quantity. Population sizes are estimated by taking $y_i = 1$ for all units, resulting in estimators $\hat{N}_a^A$, $\hat{N}_{ab}^A$, $\hat{N}_{ab}^B$, and $\hat{N}_b^B$

In a screening dual-frame survey, domain *ab* is empty and domains *a* and *b* can be viewed as strata. The estimated total for the population in a screening dual-frame survey is

$$\hat{Y} = \hat{Y}_a^A + \hat{Y}_b^B .$$

If the frames overlap and observations in the overlap domain *ab* can be selected in either or both samples, we need to adjust for the multiplicity in *ab*. Combining both samples, without adjustment, would give a biased estimator:

$$E\left[\hat{Y}_a^A + \hat{Y}_b^B + \hat{Y}_{ab}^A + \hat{Y}_{ab}^B\right] \approx Y + Y_{ab} .$$

We thus need to combine the information from the independent estimators $\hat{Y}_{ab}^A$ and $\hat{Y}_{ab}^B$ to estimate $Y_{ab}$. The following sections describe some of the methods that can be used to estimate the population total $Y$.

### 2.1. Averaging the estimates from intersection domains

The simplest method to estimate the population total is to average the domain estimators for domains that are sampled in more than one frame. For the dual-frame survey in Fig. 1, $\hat{Y}^A_{ab}$ and $\hat{Y}^B_{ab}$ both estimate $Y_{ab}$. Thus, to avoid multiplicity problems, we can estimate $Y$ by

$$\hat{Y}_{\text{ave}} = \hat{Y}^A_a + \frac{1}{2}\left(\hat{Y}^A_{ab} + \hat{Y}^B_{ab}\right) + \hat{Y}^B_b.$$

More generally, a weighted average can be used, as proposed by Hartley (1962):

$$\hat{Y}(\theta) = \hat{Y}^A_a + \theta\hat{Y}^A_{ab} + (1-\theta)\hat{Y}^B_{ab} + \hat{Y}^B_b, \tag{1}$$

where $0 \leq \theta \leq 1$.

This estimator reduces the weight of each sampled unit in the intersection domain *ab* to compensate for the multiplicity. Define new weights

$$\tilde{w}^A_i = \delta_i(a)w^A_i + \theta\delta_i(ab)w^A_i$$

and

$$\tilde{w}^B_i = \delta_i(b)w^B_i + (1-\theta)\delta_i(ab)w^B_i.$$

Then,

$$\hat{Y}(\theta) = \sum_{i\in\mathcal{S}_A} \tilde{w}^A_i y_i + \sum_{i\in\mathcal{S}_B} \tilde{w}^B_i y_i.$$

Each domain estimator is approximately unbiased for estimating its population quantity, so $\hat{Y}(\theta)$ is an approximately unbiased estimator of the population total $Y$. Since frames A and B are sampled independently and $\theta$ is fixed, the variance of the estimator is

$$V[\hat{Y}(\theta)] = V\left(\hat{Y}^A_a + \theta\hat{Y}^A_{ab}\right) + V\left[(1-\theta)\hat{Y}^B_{ab} + \hat{Y}^B_b\right]. \tag{2}$$

Note that $\hat{Y}(0) = \hat{Y}_a + \hat{Y}^B_{ab} + \hat{Y}^B_b$; this corresponds to discarding the observations in domain *ab* from the sample in frame A. Estimation in screening multiple-frame surveys can thus be considered as a special case of the general situation.

### 2.2. Hartley's estimator

The estimator in Section 2.1 is simple to compute but may lose efficiency relative to other estimators. If the estimator $\hat{Y}^B_{ab}$ from frame B has much more precision than $\hat{Y}^A_{ab}$ for estimating the domain total $Y_{ab}$, it would make sense to rely more heavily on $\hat{Y}^B_{ab}$ for estimating $Y_{ab}$.

Hartley (1962, 1974) proposed choosing $\theta$ in (1) to minimize the variance of $\hat{Y}(\theta)$. Because the frames are sampled independently, the variance of $\hat{Y}(\theta)$ is given in (2). Thus, for general survey designs, the variance-minimizing value of $\theta$ is

$$\theta_{\text{opt}} = \frac{V(\hat{Y}^B_{ab}) + \text{Cov}\left(\hat{Y}^B_b, \hat{Y}^B_{ab}\right) - \text{Cov}\left(\hat{Y}^A_a, \hat{Y}^A_{ab}\right)}{V\left(\hat{Y}^A_{ab}\right) + V\left(\hat{Y}^B_{ab}\right)}. \tag{3}$$

Using $\theta_{\text{opt}}$ gives the minimum attainable variance:

$$V\Big[\hat{Y}(\theta_{\text{opt}})\Big] = V\Big(\hat{Y}_a^A + \hat{Y}_b^B + \hat{Y}_{ab}^B\Big) - \theta_{\text{opt}}^2\Big[V(\hat{Y}_{ab}^A) + V(\hat{Y}_{ab}^B)\Big].$$

It can be seen from (3) that the larger the $V(\hat{Y}_{ab}^B)$ is relative to $V(\hat{Y}_{ab}^A)$, the larger $\theta_{\text{opt}}$ is. Note that if either $\text{Cov}(\hat{Y}_b^B, \hat{Y}_{ab}^B)$ or $\text{Cov}(\hat{Y}_a^A, \hat{Y}_{ab}^A)$ is large in absolute value, it is possible for $\theta_{\text{opt}}$ to be smaller than 0 or greater than 1. When frame A and frame B are the same, that is, domains $a$ and $b$ are empty, however, $\theta_{\text{opt}}$ is between 0 and 1.

In practice, the variances and covariances in (3) are unknown, so the optimal value of $\theta$ must be estimated from the data. Let $\hat{\theta}_{\text{opt}}$ be the estimator of $\theta_{\text{opt}}$ that results when estimators of the variances and covariances are substituted into (3). The adjusted weights for Hartley's method become

$$\tilde{w}_{i,H}^A = \delta_i(a)w_i^A + \hat{\theta}_{\text{opt}}\delta_i(ab)w_i^A$$

and

$$\tilde{w}_{i,H}^B = \delta_i(b)w_i^B + (1 - \hat{\theta}_{\text{opt}})\delta_i(ab)w_i^B.$$

Since $\hat{\theta}_{\text{opt}}$ depends on the variances and covariances of the particular response studied, the weight adjustments may differ for each response. This can lead to inconsistencies among estimates. For example, suppose $\hat{Y}_1(\hat{\theta}_{\text{opt},1})$ estimates total medical expenses in the population over age 65, $\hat{Y}_2(\hat{\theta}_{\text{opt},2})$ estimates total medical expenses in the population aged 65 or less, and $\hat{Y}_3(\hat{\theta}_{\text{opt},3})$ estimates total medical expenses in the entire population. If the surveys have complex design, it is likely that $\hat{Y}_1(\hat{\theta}_{\text{opt},1}) + \hat{Y}_2(\hat{\theta}_{\text{opt},2}) \neq \hat{Y}_3(\hat{\theta}_{\text{opt},3})$.

### 2.3. The Fuller–Burmeister estimator

Fuller and Burmeister (1972) proposed modifying Hartley's estimator by incorporating additional information regarding the estimation of $N_{ab}$. The estimator is

$$\hat{Y}_{\text{FB}}(\beta) = \hat{Y}_a^A + \hat{Y}_b^B + \beta_1\hat{Y}_{ab}^A + (1 - \beta_1)\hat{Y}_{ab}^B + \beta_2\Big(\hat{N}_{ab}^A - \hat{N}_{ab}^B\Big). \tag{4}$$

As with Hartley's estimator, the parameters $\beta_1$ and $\beta_2$ are chosen to minimize the variance of $\hat{Y}_{\text{FB}}(\beta)$; the optimal values are

$$\begin{bmatrix}\beta_1\\ \beta_2\end{bmatrix} = -\begin{bmatrix} V(\hat{Y}_{ab}^A - \hat{Y}_{ab}^B) & \text{Cov}(\hat{Y}_{ab}^A - \hat{Y}_{ab}^B, \hat{N}_{ab}^A - \hat{N}_{ab}^B)\\ \text{Cov}(\hat{Y}_{ab}^A - \hat{Y}_{ab}^B, \hat{N}_{ab}^A - \hat{N}_{ab}^B) & V(\hat{N}_{ab}^A - \hat{N}_{ab}^B)\end{bmatrix}^{-1}$$
$$\times\begin{bmatrix}\text{Cov}(\hat{Y}_a^A + \hat{Y}_b^B + \hat{Y}_{ab}^B, \hat{Y}_{ab}^A - \hat{Y}_{ab}^B)\\ \text{Cov}(\hat{Y}_a^A + \hat{Y}_b^B + \hat{Y}_{ab}^B, \hat{N}_{ab}^A - \hat{N}_{ab}^B)\end{bmatrix}.$$

As with the optimal Hartley estimator, these variances and covariances must be estimated from the data; this results in a different set of weights being used for each response variable.

### 2.4. Single-frame estimator

Bankier (1986) and Kalton and Anderson (1986) proposed single-frame methods that combine the observations into a single data set and then adjust the weights in the intersection domain for multiplicity. An observation unit $i$ in domain $ab$ can be selected in $\mathcal{S}_A$ and in $\mathcal{S}_B$, so the expected number of times unit $i$ in $ab$ is selected is $\pi_i^A + \pi_i^B$. The Kalton and Anderson (1986) single-frame estimator uses $1/(\pi_i^A + \pi_i^B)$ as the weight for observation units in domain $ab$. Thus, if $w_i^A = 1/\pi_i^A$ and $w_i^B = 1/\pi_i^B$, the adjusted weight for sampled units in frame A is

$$\tilde{w}_{i,S}^A = \begin{cases} w_i^A & \text{if } i \in a \\ \left(1/w_i^A + 1/w_i^B\right)^{-1} & \text{if } i \in ab\,. \end{cases}$$

The adjusted weights in frame B are defined similarly, with

$$\tilde{w}_{i,S}^B = \begin{cases} w_i^B & \text{if } i \in b \\ \left(1/w_i^A + 1/w_i^B\right)^{-1} & \text{if } i \in ab\,. \end{cases}$$

Then, the single-frame estimator is

$$\hat{Y}_S = \sum_{i \in \mathcal{S}_A} \tilde{w}_{i,S}^A y_i + \sum_{i \in \mathcal{S}_B} \tilde{w}_{i,S}^B y_i. \tag{5}$$

If the frame population sizes $N_A$ and $N_B$ are known, the single-frame estimator may be calibrated to those sizes using either raking ratio estimation (Bankier, 1986; Rao and Skinner, 1996) or regression estimation (Lohr and Rao, 2000). The weights can be raked by replacing $\tilde{w}_{i,S}^A$ in frame A by $\tilde{w}_{i,S}^A(N^A/\sum_j \tilde{w}_{j,S}^A)$, then replacing the weight $\tilde{w}_{i,S}^B$ in frame B by $\tilde{w}_{i,S}^B(N^B/\sum_j \tilde{w}_{j,S}^B)$, and repeating for the weights in frames A and B until convergence. Rao and Skinner (1996) showed that the raking procedure converges and gave an expression for the final estimator.

Skinner et al. (1994) used single-frame estimation for a multiple-frame agricultural sample. Several responses were of interest; each was correlated with one of the possible stratification variables. Independent stratified random samples were selected from the sampling frame using one stratification variable in each sample. With four independent samples $\mathcal{S}_A$, $\mathcal{S}_B$, $\mathcal{S}_C$, and $\mathcal{S}_D$, the weight for unit $j$ selected to be in, say, the sample from frame A, was set to $\tilde{w}_j^A = (\pi_j^A + \pi_j^B + \pi_j^C + \pi_j^D)^{-1}$. The weights were then raked to the common population size $N$ as described in Skinner (1991).

If each sample is self-weighting, that is, all the weights for sampled units in $\mathcal{S}_A$ equal $w^A$ and all the weights for sampled units in $\mathcal{S}_B$ equal $w^B$, then $\hat{Y}_S$ is a special case of the estimator in (1). In that case, $\hat{Y}_S = \hat{Y}_a^A + \theta_S \hat{Y}_{ab}^A + (1 - \theta_S)\hat{Y}_{ab}^B + \hat{Y}_b^B$, with $\theta_S = \pi_i^A/(\pi_i^A + \pi_i^B) = [w^A(1/w^A + 1/w^B)]^{-1}$. The single-frame estimator gives higher weight to the estimator of $Y_{ab}$ from the frame with the higher inclusion probabilities. The sample from the frame with the higher inclusion probabilities does not always have the smaller variance, though, so $V(\hat{Y}_S) \geq V[\hat{Y}(\theta_{\text{opt}})]$, the Hartley estimator with optimal value of $\theta$.

Unlike the Hartley and Fuller–Burmeister estimators, the single-frame estimator does not depend on estimated covariances for the response variable. It uses the same set of weights for each response considered. Calculating the weights in domain $ab$, however, requires knowledge of the inclusion probabilities for both frames, not just the frame from which a unit was selected. If one of the samples, say the sample from frame A, is not self-weighting, then $\pi_i^A$ may be unknown for a unit selected in $\mathcal{S}_B$.

## 2.5. Pseudo maximum likelihood estimator

When $N_A$ and $N_B$ are known, Skinner and Rao (1996) proposed modifying an alternative estimator proposed by Fuller and Burmeister (1972) for simple random samples to obtain a pseudo maximum likelihood (PML) estimator that can be used with complex designs. The PML estimator uses the same set of weights for all response variables and has the form

$$\hat{Y}_{\text{PML}}(\theta) = \frac{N_A - \hat{N}_{ab}^{\text{PML}}(\theta)}{\hat{N}_a^A}\hat{Y}_a^A + \frac{N_B - \hat{N}_{ab}^{\text{PML}}(\theta)}{\hat{N}_b^B}\hat{Y}_b^B + \frac{\hat{N}_{ab}^{\text{PML}}(\theta)}{\theta\hat{N}_{ab}^A + (1-\theta)\hat{N}_{ab}^B}\left[\theta\hat{Y}_{ab}^A + (1-\theta)\hat{Y}_{ab}^B\right], \tag{6}$$

where $\hat{N}_{ab}^{\text{PML}}(\theta)$ is the smaller of the roots of the quadratic equation

$$[\theta/N_B + (1-\theta)/N_A]x^2 - [1 + \theta\hat{N}_{ab}^A/N_B + (1-\theta)\hat{N}_{ab}^B/N_A]x + \theta\hat{N}_{ab}^A + (1-\theta)\hat{N}_{ab}^B = 0.$$

Fuller and Burmeister (1972) and Skinner (1991) argued that when a simple random sample is taken in each frame, the estimator in (6) can be derived using maximum likelihood principles and thus is asymptotically efficient. The PML estimator substitutes design-consistent estimators of $N_{ab}$, $Y_{ab}$, $Y_a$, and $Y_b$ for the corresponding quantities in the maximum likelihood estimator derived using simple random sampling. The PML estimator is thus consistent under complex sampling designs; unlike the estimator in (4), it does not depend on the variances of $\hat{Y}_{ab}^A$ and $\hat{Y}_{ab}^B$. Although the PML estimator need not be optimal under complex sampling designs, it often performs well in practice.

Skinner and Rao (1996) suggested using the value of $\theta = \theta_P$ that minimizes the asymptotic variance of $\hat{N}_{ab}^{\text{PML}}(\theta)$:

$$\theta_P = \frac{N_a N_B V(\hat{N}_{ab}^B)}{N_a N_B V(\hat{N}_{ab}^B) + N_b N_A V(\hat{N}_{ab}^A)}. \tag{7}$$

The estimator in (6) adjusts the estimators of the three domain totals $Y_a$, $Y_{ab}$, and $Y_b$ by the optimal estimator of $N_{ab}$. Note that calculation of $\theta_P$ in (7) requires all three domains $a$, $b$, and $ab$ to be nonempty and requires the variances of $\hat{N}_{ab}^A$ and $\hat{N}_{ab}^B$ to be positive. Thus, a different method must be used to determine $\theta$ in (6) if, for example, the sample from frame $B$ is poststratified so that $\hat{N}_{ab}^B = N_{ab}$ and $V(\hat{N}_{ab}^B) = 0$. In such situations, a value of $\theta$ can be calculated using average design effects for a fixed subset of important variables (see Lohr and Rao (2006), for a discussion of this approach and references).

In practice, $N_a$, $N_b$, $V(\hat{N}_{ab}^A)$, and $V(\hat{N}_{ab}^B)$ are estimated from the data so that an estimator $\hat{\theta}_P$ of $\theta_P$ is substituted into (6). The adjusted weights are

$$\tilde{w}_{i,P}^A = \begin{cases} \dfrac{N_A - \hat{N}_{ab}^{\text{PML}}(\hat{\theta}_P)}{\hat{N}_a^A}\, w_i^A & \text{if } i \in a \\[2ex] \dfrac{\hat{N}_{ab}^{\text{PML}}(\hat{\theta}_P)}{\hat{\theta}_P\hat{N}_{ab}^A + (1-\hat{\theta}_P)\hat{N}_{ab}^B}\hat{\theta}_P\, w_i^A & \text{if } i \in ab \end{cases}$$

and

$$\tilde{w}^B_{i,P} = \begin{cases} \dfrac{N_B - \hat{N}^{\text{PML}}_{ab}(\hat{\theta}_P)}{\hat{N}^B_b}\, w^B_i & \text{if } i \in b \\ \dfrac{\hat{N}^{\text{PML}}_{ab}(\hat{\theta}_P)}{\hat{\theta}_P \hat{N}^A_{ab} + (1-\hat{\theta}_P)\hat{N}^B_{ab}}(1-\hat{\theta}_P)\, w^B_i & \text{if } i \in ab\,. \end{cases}$$

Although $\hat{\theta}_P$ depends on the estimated variances of the overlap domain size, it does not depend on covariances of other response variables. The PML estimator thus uses the same set of weights for each response variable. Skinner and Rao (1996) and Lohr and Rao (2006) found that the PML estimator has small mean squared error and works well with a wide variety of survey designs.

### 2.6. *Using adjusted weights to estimate population quantities*

Each of the methods for estimating the population total discussed in this section produces a set of adjusted weights for the observations from frame A and a set of adjusted weights for the observations from frame B. These weights may then be used to estimate population totals and other quantities. For example, to estimate a ratio $Y/X$ of two population totals $Y$ and $X$, adjusted weights are used to find $\hat{Y} = \sum_{i\in\mathcal{S}_A} \tilde{w}^A_{iy} y_i + \sum_{i\in\mathcal{S}_B} \tilde{w}^B_{iy} y_i$ and $\hat{X} = \sum_{i\in\mathcal{S}_A} \tilde{w}^A_{ix} x_i + \sum_{i\in\mathcal{S}_B} \tilde{w}^B_{ix} x_i$. Then, $Y/X$ is estimated by $\hat{Y}/\hat{X}$. The Hartley and Fuller–Burmeister methods, as noted above, allow the weight adjustments to depend on which response is considered. This can lead to anomalies: if $X$ is the total number of engineers in the population and $Y$ is the total number of male engineers, it is possible for the Hartley or Fuller–Burmeister weightings to result in $\hat{Y}/\hat{X} > 1$. The other estimators—averaging, single frame, or PML—will not have that problem.

Other quantities can also be estimated using the adjusted weights. For the Survey of Consumer Finances, discussed in Section 1, population medians are of interest as well as population means and totals. The adjusted weights can be used to calculate medians, percentiles, and other statistics by standard methods. To estimate the median value of equity holdings for households that own equities, one can use the value of $m$ solving

$$\frac{\displaystyle\sum_{i\in\mathcal{S}_A} \tilde{w}^A_i I(y_i \le m) + \sum_{i\in\mathcal{S}_B} \tilde{w}^B_i I(y_i \le m)}{\displaystyle\sum_{i\in\mathcal{S}_A} \tilde{w}^A_i + \sum_{i\in\mathcal{S}_B} \tilde{w}^B_i} = \frac{1}{2},$$

where $I(y_i \le m) = 1$ if $y_i \le m$ and 0 otherwise.

### 2.7. *Estimation with three or more frames*

Most of the estimators discussed so far have been for the two-frame situation. All these can be extended to the situation of $Q$ frames. The simplest estimator, again, averages the estimators for each domain. For the three-frame survey depicted in Fig. 4, this estimator is

$$\begin{aligned} \hat{Y}_{\text{ave}} = {} & \hat{Y}^A_a + \hat{Y}^B_b + \hat{Y}^C_c + \frac{1}{2}(\hat{Y}^A_{ab} + \hat{Y}^B_{ab}) + \frac{1}{2}(\hat{Y}^A_{ac} + \hat{Y}^C_{ac}) + \frac{1}{2}(\hat{Y}^B_{bc} + \hat{Y}^C_{bc}) \\ & + \frac{1}{3}(\hat{Y}^A_{abc} + \hat{Y}^B_{abc} + \hat{Y}^C_{abc}). \end{aligned} \tag{8}$$

Lohr and Rao (2006) developed and compared estimators for multiple-frame surveys when $Q \geq 3$. They found that with moderate sample sizes, the optimal Hartley and Fuller–Burmeister estimators become less stable because they require estimating a large covariance matrix. The PML estimator has good efficiency and performs well under a wide range of conditions. The single-frame estimator with raking ratio estimation and $\hat{Y}_{\text{ave}}$ perform well if the relative weightings of the domain estimators are close to the optimal values.

### *2.8. Choice of estimator*

Lohr and Rao (2000, 2006) compared the asymptotic efficiency of the estimators and showed that the Fuller–Burmeister estimator has the greatest asymptotic efficiency among the estimators discussed in this chapter. In practice, properties such as transparency, robustness to assumptions, and appropriateness for the survey should be considered in addition to efficiency. Although the Fuller–Burmeister estimator has the greatest asymptotic efficiency among all linear estimators, the Fuller–Burmeister and Hartley estimators both result in a different set of weights for each response variable considered. In addition, with more than two frames, these two estimators can be unstable because they depend on a high-dimensional matrix of estimated covariances.

In many situations, an investigator may want to use an estimator with fixed weightings for the different domains, as described in Section 2.1. In the SESTAT data, for example, persons in the Survey of Doctoral Recipients frame represent a small part of the frame for the National Survey of College Graduates. The National Science Foundation uses the estimator

$$\hat{Y} = \hat{Y}_c^C + \hat{Y}_{ac}^A + \hat{Y}_{bc}^B + \hat{Y}_{abc}^A.$$

Each domain total is estimated using exactly one of the surveys.

## 3. Variance estimation in multiple-frame surveys

Variance estimation can be more complicated for multiple-frame surveys than for a single-frame survey. For the simplest estimator, that in Section 2.1, variance estimation is straightforward. The estimator of the population total is

$$\hat{Y}(\theta) = \hat{Y}_a^A + \theta\hat{Y}_{ab}^A + (1-\theta)\hat{Y}_{ab}^B + \hat{Y}_b^B,$$

where $\theta$ is a fixed value between 0 and 1. Under the assumption that the samples are taken independently,

$$V[\hat{Y}(\theta)] = V\left(\hat{Y}_a^A + \theta\hat{Y}_{ab}^A\right) + V\left[(1-\theta)\hat{Y}_{ab}^B + \hat{Y}_b^B\right].$$

Each term in the expression for the variance can be estimated using the survey design, and the two pieces added to obtain an estimator $\hat{V}[\hat{Y}(\theta)]$. The same method can be used to estimate the variance of the estimated population total using the single-frame estimator.

Screening multiple-frame surveys were discussed above to be a special case of stratified sampling. Consequently, standard methods for stratified samples can be used to estimate variances.

Variance estimation can be more complicated for other estimators. The adjusted weights for the Hartley estimator of the population total depend on $\hat{\theta}_{\text{opt}}$, which is a function of the estimated covariances from both frames. Functions of totals, or other statistics such as percentiles, also rely in a more complex way on estimators from both samples.

Several methods can be used to estimate variances of estimated population quantities in general multiple-frame surveys. These methods include Taylor linearization techniques, jackknife, and bootstrap. The Taylor linearization and jackknife methods, discussed in Lohr and Rao (2000), assume that a population characteristic of interest $\tau$ can be expressed as a twice continuously differentiable function of population totals from the frames. For Taylor linearization, the partial derivatives of this function are used together with the estimated covariance matrix of the population totals estimated from frame A and the estimated covariance matrix of the population totals estimated from frame B, to give a linearized estimator of the variance of the estimator $\hat{\tau}$. For example, $\tau = Y/X$ might be a ratio of two population totals from a dual-frame survey, with

$$\hat{\tau} = \frac{\hat{Y}(\frac{1}{2})}{\hat{X}(\frac{1}{2})} = \frac{\hat{Y}_a^A + \frac{1}{2}\hat{Y}_{ab}^A + \frac{1}{2}\hat{Y}_{ab}^B + \hat{Y}_b^B}{\hat{X}_a^A + \frac{1}{2}\hat{X}_{ab}^A + \frac{1}{2}\hat{X}_{ab}^B + \hat{X}_b^B},$$

for $\hat{Y}(\frac{1}{2})$ defined in (1). The estimated totals from frame A are $(\hat{Y}_a^A, \hat{Y}_{ab}^A, \hat{X}_a^A, \hat{X}_{ab}^A)$ with estimated covariance matrix $S_A$, and the estimated totals from frame B are $(\hat{Y}_b^B, \hat{Y}_{ab}^B, \hat{X}_b^B, \hat{X}_{ab}^B)$ with estimated covariance matrix $S_B$. The linearization estimator of the variance is then

$$g_A^T S_A g_A + g_B^T S_B g_B,$$

where in this case $g_A = g_B = [\hat{X}(1/2)]^{-1}(1, 1/2, -\hat{\tau}, -\hat{\tau}/2)^T$ is the vector of derivatives used in the linearization. Under regularity conditions, the linearization estimator of the variance is consistent. It requires, however, that the derivatives be calculated separately for each estimator that is considered.

The jackknife estimator of the variance relies on the property that independent samples are taken from the two frames (Lohr and Rao, 2000). Suppose a stratified cluster sample is taken from frame A and an independent stratified cluster sample is taken from frame B. A jackknife variance estimator carries out the jackknife separately in frames A and B. Let $\hat{\tau}_{(hi)}^A$ be the estimator of the same form as $\hat{\tau}$ when the observations of sample psu $i$ of stratum $h$ from the frame-A sample are omitted from the data. Similarly, let $\hat{\tau}_{(lj)}^B$ be the estimator of the same form as $\hat{\tau}$ when the observations of sample psu $j$ of stratum $l$ from the frame-B sample are omitted. Then, if $\tilde{n}_h^A$ is the number of primary sampling units in stratum $h$ of the sample in frame A and $\tilde{n}_l^B$ is the number of primary sampling units in stratum $l$ of the sample in frame B, the jackknife estimator of the variance is

$$v_J(\hat{\tau}) = \sum_{h=1}^{H} \frac{\tilde{n}_h^A - 1}{\tilde{n}_h^A} \sum_{i=1}^{\tilde{n}_h^A} (\hat{\tau}_{(hi)}^A - \hat{\tau})^2 + \sum_{l=1}^{L} \frac{\tilde{n}_l^B - 1}{\tilde{n}_l^B} \sum_{j=1}^{\tilde{n}_l^B} (\hat{\tau}_{(lj)}^B - \hat{\tau})^2. \tag{9}$$

The jackknife estimator of the variance is consistent for smooth functions of population means. A bootstrap estimator of the variance can be constructed similarly, and properties of the bootstrap in multiple-frame surveys are a subject of current research.

## 4. Designing multiple-frame surveys

Multiple-frame designs are usually used because they result in better coverage and less cost than a single-frame survey. The survey needs to be carefully designed to realize those cost savings.

We first discuss design for screening multiple-frame surveys. Since screening multiple-frame surveys can be considered as a special case of stratified sampling, the same principles used in allocating sample sizes in stratified samples can be used to design the survey. For a simple example, suppose a simple random sample of size $n_A$ is to be taken from frame A, and an independent simple random sample of size $n_B$ is to be taken from frame B. Each unit from frame A has cost $c_A$ and each unit from frame B has cost $c_B$, so the total cost is $C = c_A n_A + c_B n_B$. Since overlapping units are removed from one of the frames in a screening survey, frames A and B are disjoint; consequently, the variance of the estimator $\hat{Y} = \hat{Y}_A + \hat{Y}_B$ is $V(\hat{Y}) = V(\hat{Y}_A) + V(\hat{Y}_B) = N_A^2 S_A^2 / n_A + N_B^2 S_B^2 / n_B$, where $S_A^2$ and $S_B^2$ are the variances in frames A and B. Minimizing the variance subject to fixed total cost gives sample sizes $n_A = kN_A S_A / \sqrt{c_A}$ and $n_B = kN_B S_B / \sqrt{c_B}$, where $k = C/(N_A S_A \sqrt{c_A} + N_B S_B \sqrt{c_B})$. These are the sample sizes given by Cochran (1977) for optimal allocation in stratified random sampling with two strata.

In an overlapping multiple-frame survey, the design and estimator need to be considered simultaneously. Hartley (1962, 1974) derived optimal designs for his estimator when a simple random sample is taken in each frame. For the dual-frame survey shown in Fig. 1, Hartley expressed the optimal sample sizes $n_A$ and $n_B$ as a function of the costs of data collection in each frame, the variances of the response variable within each domain, and the domain means.

Biemer (1984) and Lepkowski and Groves (1986) discussed designs for the situation in Fig. 3 when a stratified multistage sample is taken from each frame. Frame A is an area frame of dwellings, and frame B is a list of telephone numbers. They considered measurement error models, where the distribution of the measurement errors can differ for the two frames. Lepkowski and Groves (1986) argued that the survey designer needs to consider nonsampling bias as well as sampling variance when allocating resources to the two samples.

## 5. New applications and challenges for multiple-frame surveys

Historically, much of the development of multiple-frame survey theory has been motivated by the situation where one frame is an area frame and the other frame is a list frame. As populations become more demographically and technologically diverse, however, a number of researchers have explored the use of multiple-frame surveys to exploit alternative frames and modes of data collection that may provide cost savings.

The internet opens many possibilities for using multiple-frame surveys. It can provide an inexpensive method of data collection, but a frame of internet users rarely includes the entire population of interest. Blair and Blair (2006) considered using dual-frame surveys to sample rare populations where the web is used to sample members of the rare population who have internet access, and a general method such as a telephone survey is used for other persons. They considered using an online panel of persons who have agreed to participate in a study as frame B. Often, an online panel has hundreds of thousands of members, and demographic and other information may be known about the panel members. In some cases, then, panel members who belong to the rare population can be easily identified and contacted, giving a relatively large sample size with small cost. In this approach, it is important that a probability sample be taken from the sampling frame of the panel; a convenience sample of persons who happen to visit a Web site and take a survey cannot be used to make inference about a population (Couper, 2000).

Unfortunately, a panel of this sort often consists of volunteers in practice; estimators derived solely from the online panel sample thus do not reflect population quantities because the panel members are self-selected. A dual-frame approach, combining information from the online panel (frame B in Fig. 3) with that from a general population survey (frame A), can reduce the bias by covering the parts of the population not in the online panel. The panel often has only a small fraction of the population, though. Writing the population total as $Y = Y_a + Y_{ab}$, it will often be the case that $Y_{ab}$, the part of the total from units present in both frames, is negligible. Some researchers have suggested that the sample from the online panel can be taken as representative of the population with internet access, but this is a strong, and usually unjustified, assumption. The dual-frame design does, however, allow researchers to investigate differences in responses between respondents from the two frames. In some cases, it may be possible to use model-based estimators or poststratification to adjust for possible biases in the online panel sample, but more research needs to be done on this topic.

These new applications present challenges for using multiple-frame surveys. When the surveys from the different frames are taken using different modes—for example, when persons with internet access fill out an online questionnaire while persons in the telephone survey are asked questions by an interviewer—it is possible that differences in domain estimators are due to mode effects rather than sampling variability. In this case, $\hat{Y}^A_{ab}$ and $\hat{Y}^B_{ab}$ are not necessarily even estimating the same quantity.

It is also important to make sure that the same quantity is being estimated in the overlap domain when using multiple-frame methods to combine information collected from different surveys, as suggested by Elliott and Davis (2005). They combined information from the U.S. National Health Interview Survey (NHIS) and the U.S. Behavioral Risk Factors Surveillance System (BRFSS) to estimate prevalence of smoking for U.S. counties. They assumed that the estimates from the NHIS were less biased and adjusted the weights for observations in the BRFSS. This approach can give more precise estimates of smoking prevalence, but care must be taken to test whether the "smoking prevalence" measured in one survey is the same as the "smoking prevalence" measured in the other survey: question order, wording effects, or characteristics of the interviewers or survey administration may cause the underlying constructs to differ in the two surveys.

The estimators of population totals given in Section 2 assumed that domain estimators are approximately unbiased. This is not necessarily the case when there is nonresponse or measurement error. For the U.S. Survey of Consumer Finances discussed in Section 1 (Bucks et al., 2006), frame A is an area frame containing the entire U.S. population, whereas frame B is a list frame of wealthier households constructed from tax return information. The response rates differ greatly in the two samples. The response rate in the frame A sample using the area frame is about 70%, whereas the response rate in the list sample is closer to 30%. Households with high net worth are more likely to refuse to participate in the survey. The dual-frame estimators of population totals need to be adjusted for the nonresponse. This can be done, for example, by performing poststratification adjustments on the modified weights constructed using one of the methods in Section 2. Additional weighting adjustments can be done within each frame separately if auxiliary information is known. Calibration estimation (Deville and Särndal, 1992) can be used to give a final set of weights that satisfy the poststratification constraints for the separate frames and for their union.

Brick et al. (2006) experimented with using a dual-frame survey to reduce coverage bias from telephone surveys. In the setting of Fig. 1, they used a frame of landline telephone numbers for frame A and a frame of cellular telephone numbers for frame B. They found that the sample from frame B had substantially more households that only had a cell phone (domain *b*) than expected. Tucker et al. (2007) found that 31% of households in the overlap domain *ab*, having both landline and cell phones, reported that they rarely receive calls on the cell phone. Although those households are officially in the overlap domain, it is likely that they will be unreachable if selected in the cell phone sample. Consequently, a disproportionate number of the nonrespondents in the cell phone sample are from domain *ab*. The probabilities that a household will be a nonrespondent differ for each frame and also differ for the domains, which makes nonresponse adjustment difficult.

All the estimators for multiple-frame surveys require that the domain membership of sampled units is known. In a screening survey, you must be able to recognize whether a unit in the area frame is in the list frame, so each population unit belongs to exactly one frame. This is not always easy to do. In the National Survey of America's Families, described in Section 1.1, households in frame A who had a telephone were not to be interviewed. To avoid multiplicity, this screening needs to be accurate: if a substantial proportion of households interviewed from frame A actually have a telephone, they will be overrepresented in the survey. In multiple-frame surveys with overlap, misclassifying sampled units into the wrong domains can result in biased estimates of population quantities. If domain membership must be determined through the survey, for example, when a person reached through a landline telephone frame is asked about cellular telephone availability, careful pretesting is necessary to ensure that the domain assignment is accurate. Mecatti (2007) noted that specific estimators may reduce the effect of certain types of misclassification errors. If $\hat{Y}_{\text{ave}}$ in (8) is used to estimate a population total, the multiple-frame weight adjustment for a sampled individual depends only on the number of frames containing the individual; misclassifying a person from domain *ac* into domain *bc* does not change the estimator. In general, however, scenarios can be constructed in which each of the estimators is sensitive to misclassification error.

Multiple-frame surveys have great potential for improving the accuracy of estimates in regions where one survey has too small of a sample size to give sufficient precision for estimates. Such regions are called "small areas" (see Chapters 31 and 32). Rao (2003a, p. 23) suggested using dual-frame designs to improve small area estimates for subpopulations of interest. As an example, Rao discussed the Dutch Housing Demand Survey, in which the main personal-interview survey is supplemented by telephone surveys in some municipalities. Then dual-frame estimation can be used to find estimates of population quantities within those municipalities. There is also potential for using dual-frame surveys to supplement a general survey when more precision may be wanted for population subgroups in the future. The U.S. NHIS has been designed so that in the future, state data from the NHIS may be integrated with supplemental data from a random-digit dialing telephone survey (Botman et al., 2000, p. 4).

Multiple-frame surveys are becoming more widely used, and in numerous situations, they are the best method to obtain good coverage while containing costs for data collection. They must be designed and analyzed carefully, however, to account for multiplicity of population units among the frames. The design needs to take into account nonsampling error and possible domain misclassification as well as sampling errors.

## Acknowledgments

This research was partially supported by a grant from the U.S. National Science Foundation. The author is grateful to the reviewers for their helpful comments.

*Sample Surveys: Design, Methods and Applications, Vol. 29A*
ISSN: 0169-7161

DOI: 10.1016/S0169-7161(08)00005-9

5

# Designs for Surveys over Time

*Graham Kalton*

## 1. Introduction

Most of the literature on survey methodology focuses on surveys that are designed to produce a snapshot of the population at one point in time. However, in practice, researchers are often interested in obtaining a video of the changes that occur over time. The time dimension can be introduced by repeating the survey at different time points or by using some form of panel design. This chapter reviews the *design* choices available for surveys over time and provides a brief overview of many additional methodological complexities that arise. References are provided to direct readers to more detailed treatments of the various topics discussed. Methods for the *analysis* of surveys over time are addressed in Chapters 24, 33, 34, and 35.

Many surveys aim to estimate characteristics of a population at a specific point in time. For example, the U.S. Census of Population provides a snapshot of the population for April 1 at the beginning of each decade. In practice, of course, data collection for the majority of surveys cannot be conducted on a single day, but the goal is still to represent the population as of that date. Also, often some of the data collected will not relate to that specific date; some retrospective data are generally collected, such as employment status in a given earlier week, illnesses experienced in the past month, and expenditures over the past six months. Nevertheless, the objective of these *cross-sectional surveys* is to collect the data needed for describing and analyzing characteristics of the population as it exists at a point in time.

This focus of cross-sectional surveys on a particular point in time is important because both the characteristics and the composition of a population change over time. The survey estimates are therefore time specific, a feature that is particularly important in some contexts. For example, the unemployment rate is a key economic indicator that varies over time; the rate may change from one month to the next because of a change in the economy (with businesses laying off or recruiting new employees) and/or because of a change in the labor force (as occurs, e.g., at the end of the school year, when school leavers start to seek employment).

Changes in population characteristics over time raise many important issues for study. At one level, policymakers need to estimate population characteristics repeatedly over time to obtain estimates that are as current as possible. They are also interested in the change in the estimates across time: Has the unemployment rate increased or decreased since the previous survey? This change is termed the *net change* and reflects changes in both the characteristics and composition of the population. A more detailed analysis would involve understanding the components of change. To what extent is the change (or lack of change) due to population dynamics, with people entering the population through "births" (e.g., people reaching age 15, for surveys of adults; people who immigrate; or people who leave institutions, for surveys of the noninstitutional population) and leaving the population through "deaths" (e.g., people who die, emigrate, or enter an institution)? To what extent is the change due to changes in the statuses of the persons in the population? Furthermore, how does change operate in cases where there is a change in status? For example, assuming no population dynamics, if the unemployment rate undergoes a net increase of 1 percent, is that because 1 percent of previously employed persons lost their jobs or because, say, 10 percent lost their jobs and 9 percent of the previously unemployed found work? The decomposition of net change into its two components leads to a measure of *gross change*. While net change can be measured from separate samples for the two occasions, measuring gross change requires repeated measurements on the same sample, or at least a representative subsample.

There are two broad classes of objectives for surveys across time, and these give rise to different approaches to survey design. In many cases, the objectives are restricted to estimating population parameters at different time points and to estimating net changes and trends. This class of objectives also includes the estimation of average values of population parameters over a period of time. None of these objectives requires repeated measurements on the same sample. They can all be achieved by collecting the survey data from representative cross-sectional samples of the survey population at different time points. Satisfying these objectives imposes no restrictions on the relationships between the samples at different time points. In particular, these objectives can be met with samples selected entirely independently at each time point. They can also be met with samples that are constructed to minimize sample overlap across time to spread the respondent burden over different sample elements. Such *repeated surveys* are discussed in Section 2.

This first class of objectives can also be met with panel designs that include some or all of the sample members at different time points. In fact, the precision of cross-sectional and net change estimates can be improved using a *rotating panel* sample design that creates some degree of sample overlap over time. Rotating panel designs may also be used to eliminate telescoping effects that occur when respondents erroneously report an event as occurring in a given interval of time. Rotating panel designs are discussed in Section 3.

The second class of objectives focuses primarily on the estimation of gross change and other components of individual change, and on the aggregation of responses (e.g., expenditures) for individuals over time. These objectives can be satisfied only by some form of panel survey that collects data from the same individuals for the period of interest. Issues involved in conducting various types of *panel*—or *longitudinal—surveys* are discussed in Section 4.

Other objectives for surveys over time relate to the production of estimates for rare populations (i.e., a subset of the general population that has a rare characteristic). One such objective is to accumulate a sample of cases with the rare characteristic over time. If the characteristic is an event, such as getting divorced, then this objective can be satisfied by any of the designs. However, if the characteristic is stable, such as being a member of a rare racial group, accumulation works only when fresh samples are added over time. In either case, analysts need to recognize that the characteristics of members of the rare population may vary over time. For example, in a survey of recent divorcees, the economic consequences of the divorce may change over the period of sample aggregation.

A different objective with a rare population is to produce estimates for that population at various points in time. If the rare characteristic is a stable one, it may be economical to identify a sample of members of that population at an initial time point and then return to that sample repeatedly in a panel design. This approach has been used, for instance, in sampling graduate scientists and engineers in the U.S. Scientists and Engineers Statistical Data System (SESTAT) (Fecso et al., 2007). An initial sample was created based on data collected in the latest decennial Census of Population, and that sample was treated as a panel to be resurveyed at intervals during the next decade. While this scheme covered all those who were already scientists and engineers and living in the United States at the time of the census, there was a need to add supplemental samples of new U.S. graduates—"births"—as the decade progressed (with a remaining gap for scientists and engineers entering the United States after the census).

The final section of the chapter (Section 5) briefly summarizes the issues to be considered in making a choice of the type of design to adopt for surveying a population over time. It also summarizes the methodological challenges that are to be faced in producing valid findings from surveys over time.

## 2. Repeated surveys

This section discusses a range of issues and designs for surveys over time when the analytic focus is on the production of a series of cross-sectional estimates that can be used in analyses of net changes and trends at the aggregate level. The designs considered here are not structured to enable longitudinal analyses at the element level.

A common form of repeated survey is one in which separate samples of the ultimate sampling units are selected on each occasion. When the interval between rounds of a repeated survey is long (say, 5–10 years), the selection of entirely independent samples may well be an effective strategy. However, in repeated surveys with multistage sample designs and with shorter intervals between rounds, sizeable benefits may be achieved by retaining the same primary sampling units (PSUs), and perhaps also units at later stages (but not the ultimate units), at each round. Master samples of PSUs are widely used for national household survey programs because of the fieldwork and statistical efficiencies they provide, both for repeated surveys on a given topic and surveys that range over different topics (U.N. Department of Economic and Social Affairs Statistics Division, 2005, Chapter V). Overlapping higher level sampling units also leads to more precise estimates of net change over time. However, a master sample becomes increasingly less statistically efficient over time, as the population changes. When updated frame information becomes available and indicates that substantial population changes have occurred, the need arises to modify the PSUs' measures of size and revise the stratification. (In national household surveys, the availability of new population census results is usually the basis for an update). To address these issues, a variety of methods have been developed to retain as many sampled PSUs as possible in the new sample while updating the measures of size and strata (e.g., Ernst, 1999; Keyfitz, 1951; Kish and Scott, 1971).

With repeated surveys of businesses, a methodology based on some form of permanent random numbers (PRNs) is often used (see Ernst et al., 2000; Ohlsson, 1995; and Chapter 17 in this volume). In essence, the methodology consists of assigning a random number between 0 and 1 to each population element on a list frame. Then a disproportionate stratified sample can be readily selected by including all elements with random numbers less than the sampling fraction in each stratum. The random numbers assigned remain with the elements over time, with the result that an element selected in the sample at a given round of a repeated survey will also certainly be in the samples at all other rounds for which its selection probability is no lower than that of the given round. This flexible procedure automatically covers changes in overall and stratum sample sizes across rounds, elements that change strata, and births and deaths. It is primarily intended to improve the precision of estimates of change across rounds and sometimes to facilitate data collection, but it can also be used to generate a panel sample for a given period. The elements in the sample for all rounds of a given period constitute a probability sample of elements that exist throughout the period, with an element's probability of being in the panel given by the minimum of its selection probabilities across the rounds (Hughes and Hinkins, 1995). The Statistics of Income Division of the U.S. Internal Revenue Service uses a PRN methodology to sample both individual and corporate tax returns and has created panel files from the cross-sectional samples for longitudinal analysis (Dalton and Gangi, 2007). The PRN methodology can also be modified to provide sample rotation to limit respondent burden on sampled businesses, particularly small businesses for which selection probabilities are low. For example, the PRN for each business can be increased by, say, 0.1 on each round and taking the fractional part if the result exceeds 1.

A critical objective for repeated surveys is the production of valid estimates of trends throughout the period of interest, particularly change from one round to the next. However, changes in the survey design are often desirable, and unfortunately even small design changes may affect the estimates. Thus, changes in question wording or questionnaire content, mode of data collection, interviewer training, interviewer field force, sampling frame, coding procedures, and imputation and weighting procedures can all threaten the validity of trend estimates. It is, for example, well documented that changing the questionnaire content can lead to context effects that can distort trend estimates (see, e.g., Biemer et al., 1991; Tourangeau et al., 2000), and even the meanings of identical questions may change over time (see, e.g., Kulka, 1982). Even a major increase in sample size alone can affect the survey estimates because of the need to recruit new interviewers and perhaps because of a reduced level of effort to obtain responses. Those conducting repeated surveys—particularly a long-running series of repeated surveys—are frequently confronted with the dilemma of whether to improve the survey procedures based on experience gained in past rounds, general methodological research, and changes in the population and topics of current interest, or to stay with past methods to maintain valid trend estimates.

When a significant methodological change is found to be necessary, a common practice is to carry out a bridging survey for one or more time periods, that is, to conduct one part of the survey using the old methods and another part using the new methods simultaneously. For example, in preparation for a conversion of the monthly

U.S. Current Population Survey (CPS) from a combination of face-to-face paper-and-pencil interviewing (PAPI) and computer-assisted telephone interviewing (CATI) to a combination of computer-assisted personal interviewing (CAPI) and CATI, the U.S. Census Bureau and Bureau of Labor Statistics (2000) conducted CATI and CAPI overlap experiments in 1992, in which a sample of 12,000 households were interviewed with the CAPI/CATI combination using a revised CPS questionnaire, to provide estimates with the revised methodology that could be compared with those produced by the official CPS. The experiments found that the new questionnaire in combination with the computer-assisted data collection did not significantly affect the estimate of the overall unemployment rate but did affect a number of other estimates, such as the duration of unemployment for the unemployed and the proportion of employees working part-time.

As another example, the U.S. National Household Survey on Drug Abuse (NHSDA, now the National Survey on Drug Use and Health) underwent a major redesign in 1999 to adopt a new method of data collection (a change from PAPI to computer-assisted interviewing, including the use of an electronic screener for respondent selection within sampled households), a different sample design with a much larger sample (and hence the need for a larger interviewer pool), and some other changes. Interviews were conducted with a national sample of approximately 14,000 households using the old PAPI methodology to evaluate the effects of the redesign and to maintain comparable data for analyses of trends. Using this sample for comparison, Gfroerer et al. (2002) provide a detailed analysis of the various effects of the NHSDA redesign.

The data collected in a series of rounds of a repeated survey are sometimes combined to produce larger samples and hence reduce sampling errors, particularly for estimates pertaining to small population subgroups (Kish, 1999). Also, the data collection for a survey may be spread over time to facilitate the fieldwork. In this case, the survey's sample may be built up from a set of replicates, each of which can produce estimates—albeit less precise—for the entire survey population. The U.S. National Health and Nutrition Examination Survey (NHANES), for example, conducts medical examinations in mobile examination centers that travel around the country from one PSU to the next (Mohadjer and Curtin, 2008). The examination centers visit 15 PSUs per year. Each yearly sample is a replicate sample that can be used to produce national estimates, but the estimates generally have low precision. For most analyses, the samples are aggregated over three or six years.

Some repeated surveys are specifically designed to be combined to produce average estimates for a period of time for characteristics that change over time. The U.S. National Health Interview Survey (NHIS) is one such survey (Botman et al., 2000). The NHIS collects health-related data from weekly samples of persons, with the data being aggregated to produce annual estimates. The annual estimate of the two-week prevalence of a seasonal illness, such as influenza, is thus an average prevalence estimate across the year. The NHIS may also be aggregated over longer periods to produce estimates for rare subgroups or over shorter periods to produce estimates for a heavy outbreak of a disease such as measles.

The complexities involved in aggregating data from a series of rounds of a repeated survey are well illustrated by the U.S. Census Bureau's American Community Survey (ACS) (U.S. Census Bureau, 2006c; Citro and Kalton, 2007). A single-stage sample of households is selected each month for the ACS, accumulating to approximately two million responding households per year. The data are aggregated to produce

1-year period estimates for governmental units with populations of 65,000 or more, 3-year period estimates for governmental units with populations of 20,000 or more, and 5-year period estimates for all governmental units, including school districts and census tracts. A user needs to understand the difference between these period estimates and the usual point-in-time estimates, and to carefully assess their applicability for his or her needs.

An ACS period estimate reflects both any changes in the characteristic under study over the period and any changes in an area's population size and composition. Thus, period estimates will differ from point-in-time estimates for characteristics that can change markedly over a period (e.g., unemployment rates). A particular issue here relates to estimates that involve monetary amounts, such as income during the past 12 months, rent or value of the accommodation, or fuel costs last month. When aggregating data over, say, 5 years, should the effect of inflation on changes in such amounts be taken into account? If so, how should this be done? Point-in-time and period estimates will also differ for areas with highly seasonal populations and for areas of rapid growth or decline. The use of period estimates of totals (e.g., the number of persons in poverty) is particularly unclear for areas that experience substantial changes in population size during the period.

## 3. Rotating panel surveys

Repeated surveys of households may retain the same set of PSUs, second-stage sampling units, and units at other higher stages of sampling from one round to the next, but they select fresh samples of reporting units on each occasion.[1] In contrast, rotating panel surveys are designed to ensure some degree of overlap in the final sample units at specified rounds. However, unlike full panel surveys, not all the same final sample units are retained over all rounds. With a rotating design, each final sample unit remains in the sample for only a limited period of time.

Rotating panel designs are widely used for labor force surveys. In the monthly Canadian Labour Force Survey (LFS), for example, each sampled housing unit is included in the sample for six consecutive months. Each month, one-sixth of the housing units enter the sample and one-sixth of them rotate out of the sample (Statistics Canada, 1998). Table 1 illustrates this rotation scheme over a period of 12 months. The sample in month 1 is made up of six rotation groups, each comprising one-sixth of the overall sample. Rotation group A has been in sample for the previous five months and rotates out in month 2, rotation group B has been in sample for four previous months and will remain in sample in month 2 before rotating out in month 3, and so on.

Various rotating panel designs are used for labor force surveys in different countries. For example, the quarterly U.K. LFS employs a five-wave rotating panel design, with one-fifth of the sample entering each quarter and one-fifth rotating out (U.K. National Statistics, 2007). The U.S. monthly labor force survey—CPS—employs a more complex rotation scheme. Each sampled housing unit is in sample for eight months, but not

[1] Indeed, repeated surveys may be designed to minimize the chance that a reporting unit is selected for rounds that are close in time (as is the case with the U.S. American Community Survey, which ensures that housing units are not selected more than once in a five-year period).

Table 1
An illustration of the Canadian LFS six-month rotation scheme for a 12-month period

| Month | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| A | G | G | G | G | G | G | M | M | M | M | M |
| B | B | H | H | H | H | H | H | N | N | N | N |
| C | C | C | I | I | I | I | I | I | O | O | O |
| D | D | D | D | J | J | J | J | J | J | P | P |
| E | E | E | E | E | K | K | K | K | K | K | Q |
| F | F | F | F | F | F | L | L | L | L | L | L |

consecutive months (U.S. Census Bureau and Bureau of Labor Statistics, 2000). A housing unit enters the sample in a given month and stays in sample for the next three months, drops out of sample for 8 months, and then returns for four consecutive months. This scheme is termed a 4-8-4 rotation scheme.

The primary analytic objectives of these labor force surveys are the same as those of repeated surveys: to produce cross-sectional estimates at each time point and to measure net changes over time. As compared with independent samples for each round, a rotating design induces a correlation between estimates at rounds in which there is some sample overlap. Since this correlation is almost always positive, the overlap results in a reduction in the sampling error of estimates of net change. The 4-8-4 rotation pattern in the U.S. CPS, for example, is fashioned to provide substantial overlap from one month to the next and also from a given month in one year to the same month in the next year.

In fact, with a rotating panel design, the precision of estimates of current level and net change can "borrow strength" from the data collected during all previous rounds of the survey, using the technique of composite estimation. See U.S. Census Bureau and Bureau of Labor Statistics (2000) for the application of composite estimation in the CPS and Fuller and Rao (2001) for an investigation into the use of regression composite estimation in the Canadian LFS. See also Chapter 33 in this volume and Binder and Hidiroglou (1988) for more detailed discussions of composite estimation and optimal rotation designs.

As well as improvements in the precision of survey estimates, a rotating design can yield important cost savings because returning to the same housing units is often less expensive than starting afresh. In particular, while the initial interview may need to be conducted face-to-face, subsequent interviews may be conducted by telephone where possible. This procedure is used in the Canadian LFS and the U.K. LFS. In the U.S. CPS, face-to-face interviewing is required for the first and fifth interviews (when a sampled housing unit returns to the sample after a gap of eight months), but other interviews may be conducted by telephone. There is, however, a concern that responses obtained by telephone may not be comparable with those obtained by face-to-face interviewing. See, for example, de Leeuw (2005) on the effects of mixed-mode data collections in general and Dillman and Christian (2005) on the issues of mixing modes across waves of a panel survey.

The LFS conducted by the Australian Bureau of Statistics uses a rotating panel survey design, with sampled dwellings remaining in sample for eight consecutive months. Until

1996, face-to-face interviewing was used on each wave, but then telephone interviewing was introduced for all interviews except the first. This change was introduced over time by rotation group. Bell (1998) used this balanced feature to analyze the effects of the change while also taking account of rotation group bias (see below). He concluded that the change in data collection methods resulted in a transitory effect on estimates by labor force status, but the effect had almost disappeared by the end of the phase-in period.

Even apart from a change in mode, there is concern that the responses obtained in repeated interviews with the same respondents may not be comparable. This effect, which is termed *panel conditioning*, occurs when responses at later waves of interviewing are affected by the respondents' participation in earlier waves of the survey. For example, respondents may change their behavior as a result of being sensitized to the survey's subject-matter and perhaps learning something from the interview (such as the existence of a welfare program). Some respondents may change their response behaviors in later interviews, perhaps demonstrating better recall after learning more about the survey contents, being more motivated to give accurate responses, giving less-considered responses because they have lost interest, or responding to filter questions in a manner that will avoid lengthy sets of follow-up questions. Respondents may also seek to be overly consistent in responses to attitude items. See Waterton and Lievesley (1989) and Sturgis et al. (2009) for a discussion of possible reasons for panel conditioning effects and Cantor (2007) for an extensive review of the research on panel conditioning.

Yet another issue with panel designs, in general, is *panel attrition* (see also Section 4). While both cross-sectional and panel surveys are subject to total nonresponse at the initial wave of data collection, a panel survey also suffers losses at later waves. Although nonresponse weighting adjustments may help to compensate for panel attrition, the estimates derived from a rotation group that has been in sample for several waves, with associated panel attrition, may differ for this reason from those derived from rotation groups that have been in sample for shorter durations.

A well-documented finding in the U.S. CPS and the Canadian LFS is that the survey results for the same time point differ across rotation groups (Bailar, 1975, 1989; Ghangurde, 1982; U.S. Census Bureau and Bureau of Labor Statistics, 2000). This effect, which reflects some combination of panel conditioning and attrition effects, is variously known as *rotation group bias*, *time-in-sample bias*, and *month-in-sample bias*. The existence of this bias for a given monthly estimate almost certainly implies that the estimate is biased but, under an additive model of the bias by rotation group, estimates of month-to-month changes are unbiased provided that the rotation group pattern is balanced at each time point (Bailar, 1975). This balance is not always achieved; in any case, it does not hold during the start-up period for a rotating panel design. Also, as Solon (1986) shows, change estimates are biased under a model that includes a multiplicative bias term.

Rotating panel designs are not restricted to labor force surveys. They are also used for purposes other than efficiency of fieldwork and improved precision of estimates of change and level. An important reason for using a rotating panel design in the U.S. National Crime Survey (now the National Crime Victimization Survey) is for the purpose of *bounding*, to deal with telescoping effects (Cantor, 1989). These effects can occur when respondents are asked to report events that happened in a given period, such as victimizations that they have experienced in the past six months or any cars purchased in the past year. Telescoping occurs when a respondent reports an event as occurring within the reference period when, in fact, it fell outside the period. See Neter and Waksberg (1964a,b) for a classic study of both panel conditioning and telescoping effects. Telescoping effects can be addressed with a rotating panel survey—indeed, any panel survey—in which the sample is reinterviewed at intervals corresponding to the reference period: events reported at the current wave that were also reported at a previous wave can be discounted because they occurred prior to the current reference period. (See also Section 4.2 for a discussion of dependent interviewing, in which respondents are reminded of their answers to the previous wave.)

Another application of rotating panel designs is when respondents cannot be expected to recall accurately all the information required for a given reference period. They may then be reinterviewed at set intervals to report the information for shorter reference periods, with the information then being aggregated to provide the information required for the full period. The household component of the U.S. Medical Expenditure Panel Survey (MEPS), for example, uses a rotating panel approach to collect data on health status and health care access, utilization, and expenditures; data on panel members are collected over five waves covering 30 months, with a new panel being introduced every year (Ezzati-Rice and Cohen, 2004). The data from two consecutive rotation groups (panels) can be pooled for a given year, and the data aggregated to produce annual estimates. The Survey of Income and Program Participation, described in Section 4, has used a similar approach (Kalton et al., 1998).

A rotating panel design can sometimes provide the data needed for longitudinal analyses for a limited time span. For example, with the ultimate sampling units retained in the MEPS rotating panel for five waves, the data for one rotation group (panel) can be analyzed to study gross changes over a specific 30-month period (Ezzati-Rice and Cohen, 2004). However, in labor force surveys, the basis of the rotation scheme is usually housing units rather than household members. While this choice eliminates the need to trace and interview households or household members that move between waves, it fails to provide the data needed for longitudinal analyses of households or persons.

A general issue with cross-sectional estimation from panel surveys is that, unless special steps are taken, the sample at later waves does not represent elements that have entered the population after the sample was selected for the initial wave. This is seldom a very serious concern for rotating panel surveys because the duration of each rotation group in the panel is fairly short. Moreover, at any point in time, new entrants can appear in the sample through the more recent rotation groups provided that the sample coverage is updated for each group (such as by updating dwelling lists in sampled segments in a multistage design). They can therefore be properly represented in cross-sectional estimation by the use of an appropriate weighting scheme that reflects the fact that they had chances of selection in only some of the rotation groups on which the cross-sectional estimates are based.

## 4. Panel surveys

This section considers survey designs in which the same elements are followed over time. Such designs are commonly known as either panel surveys or longitudinal surveys. The term “panel survey” is used in this chapter for all such designs, with the term “longitudinal” being used to describe the data such designs produce.

The distinction between panel surveys and rotating panel surveys becomes blurred in the case of panel surveys of fixed duration, when fresh panels are introduced periodically; the new panel may overlap with the current panel or it may start only after the current panel has been terminated. The distinction made here is between surveys that focus primarily on cross-sectional estimates and estimates of net change, as with the rotating panels in labor force surveys, and panel surveys that are concerned primarily with longitudinal analyses at the individual unit level (e.g., gross change).

The first part of this section (Section 4.1) describes some types of panel surveys that are in widespread use, and the second part (Section 4.2) reviews a range of methodological issues that arise with panel surveys.

### *4.1. Types of panel survey*

The benefit of a full panel survey is that it produces the data needed for longitudinal analysis, thus greatly expanding on the analytic potential of a cross-sectional survey. Although panel surveys have a long history, interest in them has greatly increased in recent years, especially with developments in the computing power needed to handle complex longitudinal data files and the software needed for conducting longitudinal analyses. There are many panel surveys currently in operation and a sizeable literature on the conduct of such surveys (e.g., Duncan and Kalton, 1987; Kalton and Citro, 1993; Kasprzyk et al., 1989; Lynn, 2009; Trivellato, 1999).

Longitudinal data obtained from panel surveys provide the opportunity for a wide variety of analyses not possible with cross-sectional data. For example, longitudinal data are needed for analyses of gross change; durations of spells (e.g., of poverty); growth trajectories with growth curve modeling (e.g., children's physical and cognitive development); early indicators that predict later outcomes (e.g., environmental exposures in childhood and health outcomes in later years); and causal temporal pathways between "causes" and effects, using longitudinal structural equation modeling (e.g., self-efficacy as a mediator between stressful life events and depressive symptoms; see Maciejewski et al., 2000). Sometimes, the longitudinal data needed can be obtained by retrospective recall or from administrative records. Indeed, some panel surveys are based entirely on administrative records; for example, the U.K. Office for National Statistics Longitudinal Study links data across censuses together with vital event data on a sample basis (Blackwell et al., 2005). However, when the quality of retrospective recall is inadequate and administrative data are unavailable or insufficient, direct data collection in a panel survey is required.

Martin et al. (2006) provide brief descriptions of a sizeable number of large-scale panel surveys of social conditions conducted in several countries. These surveys cover topics such as physical and mental health and disability; physical, social, and educational development; employment history; family dynamics; effects of divorce; dynamics of income and assets; transitions to retirement and effects of aging; and social and cultural integration of immigrants. Most focus initially on specific areas but, over time, they are likely to cover a wide range of subject-matter. (Indeed, one of the advantages of the panel design is the opportunity for data collection on many topics at different waves of the survey.) A long-term panel survey may change its areas of inquiry over time in response to changing societal concerns and, in the case of panel studies of age cohorts, to examine changes in age-related topics of interest. Panel designs have also long been used in fields such as epidemiological research on specific illnesses (e.g., Doll and Hill, 1964), studies of voting behavior (e.g., American National Election Study, 2007), and studies to evaluate the effects of intervention programs (Piesse et al., 2009).

Although panel surveys are primarily concerned with producing the data needed for longitudinal analysis at the element level, in most cases they are also analyzed cross-sectionally for each wave of data collection. An important issue for cross-sectional analysis is the adequacy of the sample coverage at each wave: unless steps are regularly taken to "freshen" the cross-sectional samples by adding samples of new entrants to the population since the last sample update, the new entrants will not be represented in the cross-sectional analyses. New entrants are generally not included, and hence are not needed, for those longitudinal analyses that start with data from the first wave of the panel.

### *4.1.1. Cohort studies*

One class of panel survey is often known as a cohort study (Bynner, 2004). Many cohort studies take samples of persons of a particular age and follow them through important periods of transition in their lives. A birth cohort may be followed throughout the life course, and, indeed, the study can be extended to follow the offspring of the original cohort. Four British national birth cohorts well illustrate this type of design in a sequential form: the first was the 1946 National Birth Cohort, also known as the National Survey of Health and Development (Wadsworth et al., 2005); the subsequent studies are the 1958 National Child Development Study, the 1970 British Cohort Study, and the Millennium Cohort Study (Centre for Longitudinal Studies, 2007). A new U.S. birth cohort study, the National Children's Study, will enroll women early in their pregnancies, and even prior to pregnancy, to study the effects of environmental exposures and other factors in early life on the children's health and development till they are 21 years old (NCS, 2007).

Although birth cohorts provide extremely valuable longitudinal data for examining the effects of early childhood experiences on health and other factors much later in life, this great strength is accompanied by some limitations. The sample members of all age cohort studies are subject to the same historical events, or period effects (e.g., wars and environmental disasters) that affect the entire population at that time. They may also be affected by such events differentially because of their susceptibility at the ages at which the events were experienced (cohort effects). Data from a single cohort confound age, period, and cohort effects, and the results must be interpreted accordingly (Bynner, 2004; Yang, 2007). For instance, the effects of early life experiences on later outcomes must be viewed in the context of the period and cohort studied; the magnitude, or even the existence, of the effects may not apply to the current or later generations.

Another limitation of a birth cohort study is that its members would have to be followed for a very long time before it would be possible to investigate, say, the effects of retirement on health outcomes. For this reason, many age cohort studies start at later ages. For instance, the U.S. Health and Retirement Study (HRS) (Juster and Suzman, 1995), the English Longitudinal Study of Ageing (ELSA, 2007), and the Survey of Health, Ageing and Retirement in Europe (SHARE, 2007) follow cohort members from later middle age into later life. The U.S. Bureau of Labor Statistics' National Longitudinal Surveys (NLS) comprise seven cohorts defined by starting age group and sex (e.g., men and women aged 12–17 years, women aged 30–44, and men aged 45–59) (U.S. Bureau of Labor Statistics, 2005). The U. S. National Center for Education Statistics has conducted many cohort studies, including the Early Childhood Longitudinal Studies, Birth and Kindergarten Cohorts; the Education Longitudinal Study of 2002 (which is following a 10th grade cohort through high school to postsecondary education and/or work); the Beginning Postsecondary Students Longitudinal Study; the Baccalaureate and Beyond Longitudinal Study; High School and Beyond; and the National Longitudinal Study of the High School Class of 1972 (U.S. Institute of Education Sciences, National Center for Education Statistics, 2007).

### *4.1.2. Household panel surveys*

A second major class of panel survey is the household panel survey (Rose, 2000). The ongoing U.S. Panel Study of Income Dynamics (PSID) (Hill, 1992), started in 1968, provided a major impetus for this type of study. The PSID design has been adopted for studies in many countries, including the ongoing British Household Panel Survey (Taylor et al., 2007) (currently being expanded into the Understanding Society study), the ongoing German Socio-Economic Panel (German Institute for Economic Research, 2007), and the European Community Household Panel (ECHP, 2007). These surveys start with a sample of households and then follow household members for the duration of the panel. To reflect the economic and social conditions of the households of panel members at each wave, the surveys also collect data on persons with whom panel members are living at later waves, termed variously as *cohabitants*, *nonsample persons*, or *associated persons*. The surveys are thus, in reality, samples of persons rather than households. Households are constantly changing, with members joining and leaving, new households coming into existence, and others ceasing to exist. For this reason, the definition of a longitudinal household is extremely problematic unless the time period is very short (see, e.g., Duncan and Hill, 1985, and McMillen and Herriot, 1985). Person-level analyses with wave-specific household characteristics attributed to panel members are generally preferred for longitudinal analyses.

Other household panel surveys modeled along the lines of the PSID are the U.S. Survey of Income and Program Participation (SIPP) (Kalton et al., 1998), the Canadian Survey of Labour and Income Dynamics (SLID) (Statistics Canada, 2008), and the European surveys conducted under the European Union Statistics on Income and Living Conditions (EU-SILC) regulations (replacing the ECHP surveys) (Eurostat, 2005). A distinctive feature of all these surveys is that they are designed to last for a fixed duration. The SIPP panels, which collect data every four months, have varied between 2½ and 4 years in length; the SIPP design has varied between overlapping panels and abutting panels where a new panel starts only as the last panel ends. The SLID panels, which collect data twice a year, last for six years; the SLID uses a rotating design with a fresh panel starting every three years. The EU-SILC surveys follow a four-year rotation design.

An advantage of abutting panels is that the full sample size is available for longitudinal analyses for the period of the panel. However, abutting panels cannot handle longitudinal analyses for a period that spans two panels. Also, valid estimates of trends from cross-sectional estimates, such as the annual estimates that are important for these surveys, cannot be produced because of variable time-in-sample biases across years. Rotating designs with an annual rotation can produce acceptable trend estimates because of the constant balance of time-in-sample across waves.

The choice of the length of time for household panel surveys of limited duration depends on a combination of analytic objectives and practical data collection considerations, particularly respondent burden and its effects on response rates at later waves. In the SIPP, for example, the duration was extended from the original goal of eight waves (2⅔ years) to 12 waves (4 years) in 1996 to provide longer periods of observation for use in longitudinal analyses, such as durations of spells of poverty. (See Citro and Kalton, 1993, for a review of the SIPP program and a discussion of the SIPP panel length.) However, with three waves of data collection each year, the extension to four years was accompanied by lower response rates. Response rate and other practical implementation issues influenced the decision to move from the ECHP to the EU-SILC rotating design (Eurostat, 2005).

### *4.1.3. Cross-national panel surveys*

An important recent development in general survey research is the use of survey data for cross-national comparisons, enabling the investigation of the effects of differing societal conditions on the populations involved. This development applies equally with some panel surveys. Notable examples are the aging panel surveys (HRS, ELSA, and SHARE) that are coordinated across many countries and the household panel income surveys. While coordination is valuable, there remains a need for harmonization of the data collected in surveys in different countries. To facilitate cross-national research on economic and health issues in Australia, Canada, Germany, Great Britain, and the United States, Burkhauser and Lillard (2007) have created a Cross-National Equivalent File from the data collected in the household panel surveys in these countries.

## *4.2. Methodological issues in panel surveys*

Many of the same methodological issues apply to panel surveys and rotating panel surveys, although the importance of an issue may be different. For example, all forms of panel surveys must take into consideration the issues of sample attrition, panel conditioning, time-in-sample bias, and the need to cover new entrants to the population for cross-sectional estimation. However, concerns about sample attrition and covering new entrants increase with panels of longer duration, and conditioning effects are a serious concern for many forms of longitudinal analysis.

### *4.2.1. Maintaining panel participation*

Maintaining participation of panel members throughout the life of the study is a critical issue with a panel survey. As in a cross-sectional survey, a panel survey is subject to *total nonresponse*, which occurs when a sampled unit fails to participate in any wave of the panel. In addition, a panel survey is subject to *wave nonresponse*, which occurs when a sampled unit participates in some but not all waves of the survey.

Wave nonresponse may consist of *attrition nonresponse* (when the unit drops out of the panel at one wave and never returns to it) or *nonattrition nonresponse* (when a sampled unit misses a wave but responds at one or more later waves). The potential patterns of wave nonresponse depend on the following rules adopted for the panel. For instance, for practical reasons, many surveys make no attempt to convert initial nonrespondents into respondents at the next or later waves. Thus, initial nonrespondents are all total nonrespondents. Also, nonrespondents who adamantly refuse to participate or cannot be located at one wave may not be followed in subsequent waves and, hence, become attrition cases. Frequently, no attempt is made to follow up those who have missed two consecutive waves.

In general, panel surveys encounter the highest loss rate at the initial wave, after which high proportions of those responding at each successive wave also respond at the following wave. However, the accumulation of nonrespondents over time frequently results in a high overall nonresponse rate. With long-term panels, this situation raises the dilemma of whether to continue with the existing panel, with its increasing analytic potential, or terminate it and start afresh. Some panels—particularly those with high respondent burden—are designed to be of limited duration because of concerns about attrition. The possible biasing effects of accumulating nonresponse are a serious concern in nearly all panel surveys; see, for example, the special issue on attrition in longitudinal surveys in the *Journal of Human Resources* (Volume 33, Number 2, 1998).

The primary causes of wave nonresponse are loss to follow-up for panel members who move and refusals resulting from the repeated burden of being a panel member. A variety of methods are used to minimize loss to follow-up, particularly in panels with lengthy intervals between waves. One approach is to institute methods for tracking panel members so they can be located if they move. For instance, mailings such as birthday cards or newsletters with key findings from the last wave can be sent to panel members, using delivery methods that require the post office to forward mail and inform the sender of the new address. When the tracking methods fail, panel members must be traced to their new addresses. Collecting contact information (e.g., telephone numbers) for neighbors and relatives who are not likely to be mobile (e.g., parents of young people) can be helpful. Otherwise, a variety of web-based and other search procedures may be used. With sufficient effort, loss to follow-up can be limited, even for panels with long intervals between waves (Couper and Ofstedal, 2009).

Preventing the loss of panel members who are no longer willing to participate is a challenge. As with a cross-sectional survey, incentives may be offered to increase participation, but there are additional factors to be considered in a panel survey. Should incentives be offered at every wave? If not, will panel members paid an incentive at one wave refuse to participate when they are not offered an incentive at the next wave? The results of a 1996 U.S. SIPP incentive experiment, in which a monetary incentive was offered only at the first wave, do not support that concern. This experiment found that the sample loss at that and the following five waves was lower with an incentive of $20 than with an incentive of $10 and that both rates of loss were lower than the sample loss in the control group, which received no incentive (Kalton et al., 1998; Mack et al., 1999). With panel surveys, the opportunity exists to target incentives to respondents who demonstrated reluctance at the previous wave, such as those who failed to answer many questions. There are debatable equity issues with this procedure because it serves to reward behaviors that are undesirable from the survey organization's perspective. Another form of targeting is to offer increased incentives to those who facilitate the interview at a given wave. This form of targeting has been found to be cost-effective in the U.S. NLS, where larger incentives are offered to those who call in for a telephone interview (Olson, 2005). Incentives are used in one form or another in many panel surveys.

Minimizing respondent burden is another approach to limiting the loss to follow-up from refusals at later waves of a panel survey. One way to reduce respondent burden is via linkages to administrative data; such linkages can also provide data that panel members are unable to report accurately. In the Canadian SLID, income tax records are used to reduce the reporting burden on many respondents. The survey includes a January interview to collect data on labor market experiences, educational activity, and family relationships and a May interview to collect income data; more than 80 percent of the respondents grant Statistics Canada permission to collect income data from their tax files, and thereby avoid the burden of a May interview (Statistics Canada, 2008). Linkages to administrative data can also extend the period of observation of a panel survey without extending the period of the survey data collection, with its attendant nonresponse losses. A good example is provided by the U.S. HRS, a panel survey that starts with samples of persons aged 51–61. The analytic value of the survey data is greatly enhanced by linkages to lifetime earnings and benefits records in Social Security files and to health insurance and pension data from employers (Juster and Suzman, 1995). See Calderwood and Lessof (2009) for a discussion of linkages implemented in panel surveys in the U.K. For ethical reasons, panel members must be asked for their permission to make the linkages, and procedures must be put in place to ensure that the linkages cannot harm the panel members.

### *4.2.2. Measurement error*

Measurement errors are a concern in all surveys, but they are particularly problematic for longitudinal analyses of panel survey data. To illustrate this point, consider the estimation of the stability of an attitude score across two waves of a panel survey and assume that the scores are subject to random measurement errors that are independent between waves. While the cross-sectional and net change estimates are unbiased in this case, the stability of the estimates over time is underestimated. Another example is the important case of estimating gross change in employment status from a labor force survey. Even if the cross-sectional and net change estimates are considered acceptable, measurement errors can lead to serious overestimates of gross change (see, e.g., Chua and Fuller, 1987). Kalton et al. (1989) list a variety of sources of differential measurement errors that can distort the estimation of gross change: panel conditioning effects, change in mode of data collection, change in respondents between waves (including the possibility of proxy respondents when the sampled person cannot be contacted), changes in personnel (e.g., interviewers, coders of responses to open questions), changes in the questionnaire (with possible context effects even when the questions involved remain the same), changes in the questionnaire content, changes in interpretations of a question, imputation of missing responses, matching errors in linking the files for the two waves, and keying errors. The kinds of effects that measurement errors have on analyses of gross change also apply for many other forms of longitudinal analysis.

One way to try to reduce the overestimation of gross change is to use dependent interviewing, in which respondents are reminded of their responses on the previous wave of the panel. A risk with dependent interviewing is, of course, the generation of false consistency in the responses. Many studies have been conducted to evaluate the effect of dependent interviewing (e.g., Hill, 1994; Hoogendoorn, 2004; Jäckle, 2009; Lynn and Sala, 2006); see Mathiowetz and McGonagle (2000) and Jäckle (2009) for reviews of the technique. In general, dependent interviewing is thought to reduce response errors and the overestimation of gross change. Dependent interviewing may be applied in a proactive form by reminding respondents about their reported status at the previous wave (e.g., they were employed) and then asking for their current status, or in a reactive form by asking about their current status and any discrepancy from their previous response. Proactive-dependent interviewing can also be useful to reduce respondent burden, whereas reactive-dependent interviewing may be less susceptible to false consistency effects. In situations where household respondents may change between waves of a panel, dependent interviewing may result in the disclosure of one household respondent's responses to a different household respondent in the next wave. This feature raises a confidentiality issue that may need to be addressed with a consent form that permits sharing of responses across household members (Pascale and Mayer, 2004).

A particular aspect of excessive change between waves becomes evident in panels that ask respondents to report their statuses for subperiods within the interval between waves. Thus, for example, the U.S. SIPP collects data on a monthly basis within each four-month interval between waves. A common finding in this situation is that the amounts of gross change between adjacent months are much greater for pairs of months for which the data are collected in different waves than for pairs of months for which the data are collected in the same wave. This effect, which is termed the *seam effect*, has been investigated in many studies (see, e.g., Cotton and Giles, 1998; Kalton and Miller, 1991; Kalton et al., 1998; Moore and Kasprzyk, 1984; Rips et al., 2003). The effect is likely to be a combination of false consistency within a wave and overstatement of change across waves. Dependent interviewing can also be used to address this problem.

### *4.2.3. Weighting and imputation*

The standard weighting adjustment methods used to compensate for total nonresponse in cross-sectional surveys can be applied for total nonresponse in panel surveys (see Chapter 9). However, compensating for wave nonresponse (particularly, nonattrition nonresponse) and item nonresponse is much more challenging.

Apart from total nonrespondents, a good deal of information is known about the other cases with missing data based on the responses they provided in the wave(s) of data collection in which they have participated. One approach for handling missing data is to impute all the missing items, including all the items in waves in which the sampled unit is a nonrespondent. This approach has the advantage of retaining in the analysis file all the information that the sampled units have reported. However, the large amount of imputation involved raises concerns about distortions that the imputed values may introduce into an analysis. An alternative approach is to use nonresponse weighting adjustments to handle some or all of the missing waves. This approach limits the analysis file to sampled units that responded in all the relevant waves. It uses a limited number of the survey responses in making the adjustments, but the responses to other items are lost (Kalton, 1986; Lepkowski, 1989). Imputation is the natural solution when a survey unit fails to respond to just a few items, but the choice between imputation and weighting adjustments is less straightforward for wave nonresponse.

To retain the covariance structure in the dataset, imputation requires that all the other variables associated with the variable to be imputed should be used as auxiliary variables in the imputation model. Satisfying that requirement adequately is difficult enough with a cross-sectional survey, but it is much more challenging with a panel survey because the model has to incorporate variables from other waves of the survey and variables from the given wave. In particular, it is important to include the responses to the same variable at other waves, or gross change will be overestimated. A practical issue is that, in many cases, a completed data set is produced after each wave of a panel survey to enable analyses of all the data collected up to that point. Imputations for the current wave can use current and previous wave data as auxiliary variables but not data from later waves. One solution is to produce preliminary imputations for each wave and then produce final imputations when the next wave's data can be incorporated into the imputation scheme. Although somewhat laborious, this solution is used in the British Household Panel Survey (Taylor et al., 2007).

Concerns that mass imputation for wave nonresponse may introduce distortion into analyses have led to a general, but not universal, preference for weighting adjustments over imputation for handling this type of missing data. In the case of attrition nonresponse, a common practice is to develop weights for all those responding at each wave, based on data collected in previous waves (by the attrition definition, all those who responded at the current wave have responded at all previous waves). With the large amount of data available for respondents and attrition nonrespondents at a given wave, the development of the weighting adjustments is more complex than in most cross-sectional surveys, but the process is essentially the same. Procedures such as Chi-squared Automatic Interaction Detector (CHAID), propensity score weighting, and raking can be used in developing the weighting adjustments (see, e.g., Kalton and Brick, 2000; Rizzo et al., 1996). Analysts select the set of weights for the latest wave in which they are interested and they can then use that set of weights to conduct any longitudinal analyses they desire.

With many possible patterns of response/nonresponse for nonattrition cases across waves, the use of weighting adjustments in this case can result in the production of a multitude of sets of weights if all the data for responding waves are to be retained for analyses involving data from any given set of waves. Including attrition nonresponse, there is, in fact, a maximum of $2^H - 1$ patterns of response/nonresponse across $H$ waves, but this number is often reduced somewhat by the following rules that are used. Rather than compute sets of weights for each of the potential patterns, analysts often reduce the sets of weights by discarding data for some reported waves. For instance, discarding data from all waves after the first nonresponding wave converts all nonattrition cases to attrition cases and reduces the number of sets of weights to $H$. Starting from that basis, a restricted number of additional sets of weights may be added based on a review of the combinations of waves that are of analytic importance together with an assessment of the extent of data loss for these combinations resulting from a weighting scheme that treats all wave nonrespondents as attrition cases.

An alternative approach for handling certain nonattrition patterns is to use imputation rather than weighting. For example, in the 1991, 1992, and 1993 U.S. SIPP panels, a longitudinal imputation procedure was used for persons who missed one wave that fell between two waves in which they responded; in the 1996 panel, this procedure was extended to include two consecutive missing waves (see Kalton et al., 1998, for research on longitudinal imputation and the missing wave imputation procedures adopted for the SIPP).

A unique weighting issue arises with cross-sectional estimation in household panel surveys as a result of the collection of survey data for the cohabitants with whom panel members are living at each wave. To take advantage of all the data collected at a given wave in producing cross-sectional estimates for that wave requires a weighting scheme that takes account of the multiple routes of selection by which the household and household members can be selected for the sample at that wave. For example, if a member of an originally sampled household leaves to join another household, that person's new household could be selected either via that person's original household or via the original households of the other members of the new household. A weighting scheme that takes the multiple routes of selection into account and that depends only on the selection probabilities of originally sampled households is described by Ernst (1989), Kalton and Brick (1995), and Lavallée (1995, 2007b). Verma et al. (2007) discuss the application of this scheme to rotating panel designs, with particular reference to the EU-SILC surveys.

### *4.2.4. Sampling issues*

There are many special sampling considerations that arise with panel surveys. One concerns the degree of clustering to be used in selecting the sample for the first wave of a panel. The effectiveness of clustering in reducing interviewers' travel time and facilitating face-to-face callbacks dissipates over time as some panel members move to new addresses. Also, once a panel sample has been enrolled with face-to-face interviews, other methods of data collection that do not benefit from clustering (telephone, mail, and web) may be used in later waves. These considerations argue for less clustering in the first wave of a panel survey than in a cross-sectional survey, thereby cutting back the increases in the variances of survey estimates resulting from clustering.

Most surveys aim to produce estimates for certain subgroups of the population as well as for the total population. Smaller subgroups are often oversampled to generate sample sizes that produce adequate levels of precision for the subgroup estimates. The use of oversampling in a panel survey must be carefully assessed. With a long-term panel, survey designers should take into account that the survey objectives and subgroups of interest may change over time so that the initial oversampling may be detrimental later on. Also, the type of subgroup must be considered. When the defining characteristic of the subgroup is a static one, as with a racial/ethnic subgroup, oversampling can be particularly advantageous in a panel survey. In this case, the benefits of the oversampling apply throughout the life of the panel, whereas any additional costs associated with the oversampling are incurred only in the initial wave. However, when the defining characteristic is liable to change over time (e.g., being in poverty or living in a particular province), oversampling based on the initial state can be problematic, particularly when a high degree of oversampling is used. Over time, panel members will move into and out of the subgroup. As a result, subgroup members at later waves will have markedly different weights, leading to a serious loss in precision of subgroup estimates, even to the point that they may not be useful. An extreme example occurs with panels of businesses; often, highly disproportionate samples of businesses are used, but over time, some small businesses that were sampled at very low rates may grow substantially. These high-growth businesses retain the large weights associated with their initial selection probabilities, a feature that gives rise to a serious loss of precision in the survey estimates. If oversampling is to be used with nonpermanent subgroups in a panel survey, consideration should be given to keeping the variability in sampling rates within reasonable bounds to avoid the loss of precision associated with movement across subgroups.

In the types of panel survey described here, the primary focus is on providing the data needed for longitudinal analyses. However, panel survey data are also widely used for cross-sectional analyses of the data produced at each wave. An important issue for these analyses is representation of the full population at the time of the wave in question, that is, covering units that entered the population after the sample for the initial wave was selected. The same issue arises for longitudinal analyses, where the starting point for the analyses is later in the life of the panel. New samples may be added to give representation to new entrants at later waves. They may also be added to counteract the sample loss from initial wave and attrition nonresponse, in response to an expanded definition of the population of inference, to increase sample sizes for certain subgroups that have become of analytic interest, or just to expand overall sample size.

A number of panel surveys use methods to add sample at later waves as, for example was described in Section 1 for the SESTAT (Fecso et al., 2007). As another example, the U.S. National Education Longitudinal Study of 1988 (NELS:88) started with a sample of 8th grade students in 1988. At the first follow-up wave at 10th grade, the sample was freshened by adding a sample of 10th graders who were not in 8th grade in the 1987–1988 school year. The ongoing U.S. PSID, started in 1968, added a sample of post-1968 immigrants in 1997 (PSID, 2007). The German Socio-Economic Panel, started in West Germany in 1984, added a sample in East Germany in 1990, a sample of immigrants in 1994–1995, and further new samples since then to increase sample size and to provide an oversample of high-income households (German Institute for Economic Research, 2007). The weighting schemes for the various longitudinal data files can become complex when sample additions are introduced at later waves of a panel survey.

### *4.2.5. Ethical and data disclosure issues*

To conclude the discussion in this section, the special ethical issues and data disclosure risks associated with panel surveys deserve comment (Lessof, 2009). The requirement that sampled persons be informed about the purposes of the study at the outset can be difficult to satisfy in a long-term panel study, the purposes of which may change during the life of the panel. Also, those directing the study and conducting the data collections may change over time. Researchers should pay attention to these issues in designing consent forms.

Panel surveys are expensive to conduct, but they produce extremely rich data that can be valuable for analyses of many different subjects. To capitalize on the investment, the data should be made available to many researchers. However, the rich longitudinal data often pose a high disclosure risk (see, e.g., Béland, 1999). Although standard techniques such as data suppression (particularly of detailed geography), top coding, data swapping, and subsampling may provide adequate protection to enable the release of a public use data set for one wave, these techniques often do not afford sufficient protection for a public use panel file, which includes much more data. In this case, alternative methods may be needed to make the data available to researchers, such as restricted use files, secure data enclaves, and remote analyses. Another aspect of making panel data available for analysts is that full documentation must be maintained on an ongoing basis, both to advise analysts on the contents of the complex panel data and to record the survey details for use by those analyzing the data years later.

## 5. Conclusions

Those planning survey data collections to provide data across time have a choice between repeated cross-sectional, rotating panel designs, and full panel designs. If the data are to be used for longitudinal analyses, then only a panel design will serve the purpose. However, if the data are to be used only for overall trend analyses, any of the designs can be used, provided that the sample is freshened at each wave to give representation to new entrants to the population.

The design considerations for a series of repeated cross-sectional surveys would appear to be the same as those for a single cross-sectional survey, but in fact there are differences. Those planning a series of repeated cross-sectional surveys need to reflect on what the data needs might be in the future in order to cover them from the outset. During the course of the series, they will likely also face difficult decisions about changing aspects of the design to meet current conditions and conform to current best survey practice, or whether to stay with the existing design to maintain valid estimates of trends. Analysts of repeated surveys must be cognizant of any changes made to the design that may distort trend estimates.

Panel surveys are far more complex to design and analyze than cross-sectional surveys. In addition to general issues of survey design, designers of panel must pay a great deal of attention to such issues as maintaining the cooperation of panel members, tracking and tracing methods, introducing sample freshening to be able to provide valid cross-sectional estimates, the use of dependent interviewing, and the use of incentives. Analysts must be cognizant of the effects of measurement errors and panel conditioning on their analyses, as well as the likely deterioration in the representative nature of the sample over time.

*Sample Surveys: Design, Methods and Applications, Vol. 29A*
ISSN: 0169-7161

DOI: 10.1016/S0169-7161(08)00006-0

6

# Sampling of Rare Populations

*Mary C. Christman*

## 1. Introduction

In almost all research areas for which surveys are performed, whether it is social, environmental, biological, or other scientific fields, there are many situations in which it is desired to estimate the parameters of populations that are somehow rare. To define rare, we first need some common definitions. Let the population to be the set of elements that it is of interest to enumerate or characterize in some way. This is distinct from the sampling frame, which is the set of units that can be sampled in a probabilistic sampling design. For example, in a study of sexually transmitted diseases (STDs), it may be of interest to determine the health status of symptom-free individuals with STDs. So, the population of interest is the health status of all individuals who have an STD and are symptom-free. They cannot be enumerated for sampling purposes, and so the sampling frame could be the list of household addresses in the city where the study is to take place.

A "rare population" can have several meanings. First, a rare population is one in which the size of the population ($N$) is very small, for example, an endangered species. Here, even if the sampling frame and the population do not coincide, the number of sampling units containing the rare elements is small. For example, in a study of endangered species, the sampling frame might be circular plots within suitable habitat. The second definition relates to populations in which the presence of a particular trait is very low, such as genetic disorders that occur very infrequently in live births. Rare in this setting refers to the rarity of the sub-population ($M$ out of $N$) displaying the trait of interest. In this situation, it is often the case that the elements with the trait are not identifiable before sampling commences. Hence, screening methods which sample from both parts of the population and identify members of the subpopulation are required.

A third definition is where the elements are not necessarily rare but are cryptic or hidden. As a result, detectability is low and the population appears rare since so few are observed. In these cases, sampling designs that allow for estimation of both detectability and rarity are required to obtain sufficient information about the cryptic population.

The final definition of rare is the case where the population is not necessarily small or is the trait particularly rare but instead the proportion of *sampling units* containing elements from the population is very small. This occurs most frequently when the population or the trait within a population is highly clustered in space or time, and the sampling units are spatial regions (such as zip codes or circular plots) or time segments (such as a week or 12 hour part of the day). As a result, a large proportion of sampling units do not contain any elements of interest but those that do contain elements can contain high numbers of them. For example, when acquired immune deficiency syndrome (AIDS) was initially identified, it was shown that AIDS patients tended to congregate in areas with good health care facilities (Ellis, 1996). Any sampling design based on a sampling frame composed of large primary sampling units (PSUs), such as zip codes, would have had a high proportion of PSUs with no AIDS patients.

Rare populations require sampling designs that provide high observation rates while also controlling sample sizes. Hence, distinct sampling strategies have been developed to adjust for the infrequent observations of the elements of interest. The choice of sampling design is influenced by the objectives of the study. Common objectives that influence design are as follows: estimating population size ($N$) or density ($N/A$, where $A$ is area), developing probability maps for the presence of a rare trait, estimating the proportion of the population carrying a rare trait ($M/N$), comparing parameters among two or more populations, monitoring for temporal changes within the population, and detecting the impacts of interventions. A secondary objective may be the need for sufficient samples for other types of analyses unrelated to the population parameters, that is, when the object is to estimate other characteristics of the rare population. For example, in a study of a rare species, the researcher may be interested in obtaining sufficient numbers of observations for use in predictive models of habitat suitability or resource selection. Another example is estimating the median income of employed persons with a rare form of disability. In the following, we assume that the sampling frame is sufficiently large relative to the sample size, so that the finite population correction factor can be ignored.

Following is a review of methods that are used to estimate the parameters concerning rare populations. Previous reviews of sampling designs for rare populations and comparisons among designs include Kalton (1993, 2003), Kish (1965), Sudman and Kalton (1986), Sudman et al. (1988), Kalton and Anderson (1986), Thompson and Seber (1996), Christman (2000), Thompson and Collins (2002), and Turk and Borkowski (2005). More recently, Cervantes and Kalton (2007) described sampling rare populations using screening methods, such as telephone surveys.

## 2. Modifications to classical design-based sampling strategies

### 2.1. Random sampling

Simple random sampling is ill-suited for determining parameters of rare populations due to a high probability of a sample of fixed sample size containing many or all zeroes. This usually results in either a degenerate distribution around zero or a highly skewed empirical distribution with excess zeroes. In spite of inherent problems, research has been performed to determine the sample sizes needed to obtain sufficient observations to reduce the influence of the excess zeroes and increase the probability of observing individuals from a rare population. For example, Green and Young (1993) determined the sample size that ensures a specified probability of detection of the rare events assuming either a Poisson or a negative binomial distribution. They considered sampling of a spatial region that has been divided into non-overlapping quadrats. Smith (2006) modified their work to consider timed searches of a spatial region with the intention of increasing the probability of observing elements of a rare population. More generally, Venette et al. (2002) considered random sampling under three possible distributions: binomial, beta-binomial, and hypergeometric sampling strategies. Johnson et al. (2003) developed a Bayesian approach for estimating the sample size needed in surveys for determining the absence of a disease. Another approach is sequential sampling, according to a negative binomial distribution, that is random sampling until at least $r$ rare elements are observed in the sample. Liu (1999) described the confidence interval estimation when using such a sampling approach to estimate the prevalence of a rare disease. These approaches were preceded by the original work of Fisher (1934) and Rao (1965) on the effects of the methods of sampling, called ascertainment, on the distributions of the study variables. This approach is referred to as the method of weighted distributions. Fisher (1934), for example, compared sampling of families known to have albino children (the "Proband" method) with direct sampling of individuals with albinism ("Sib" method). Of interest was the estimation of the probability of having albinism given that a family had an albino child. In the Proband method, the distribution of number of albino children in a family of size $n$ is truncated due to the fact that families with no outward evidence of albinism would be excluded from the sample. Rao (1965) expanded on this work to allow for other reasons for nonsampling of individuals of interest. More recently, Chung and Kim (2004) described a Bayesian approach to the method of weighted distributions.

In social research, a common method to obtain sufficient sample size when the elements of interest are a subset of a larger population and the main focus is to study the characteristics of the rare subpopulation involves screening during random sampling. Screening is performed when the population is oversampled to ensure sufficient sample sizes of the subpopulation that is designated as rare. When screening is used with simple random sampling, such as when random digit dialing is used in telephone surveys, the total sample size needed to generate a sample size of $n_r$ elements in the rare subpopulation is $n_{\text{total}} = n_r/p_r$, where $p_r$ is the proportion of rare elements in the sampling frame. If in addition, nonresponse is observed, the estimated sample size needs to be modified to account for the degree of nonresponse (Cervantes and Kalton, 2007).

## *2.2. Stratified random sampling*

To control the costs of sampling for rare populations, several authors have suggested specialized methods for stratification and for allocation of samples among strata. Ericksen (1976), for example, recommended constructing strata such that rare elements are concentrated in one or a few strata and that those strata with rare elements be oversampled to obtain sufficient elements for precise estimation. In his study, the population of interest tended to be concentrated in specific parts of the study region and at least partial information on spatial distribution was available *a priori*. Kalton and Anderson (1986) described stratification and disproportionate samplings for estimating the prevalence of a rare trait in a population and estimating the mean of a rare population.

The efficiency of disproportionate sampling when the population is stratified depends on the effectiveness of the stratification in grouping the rare population into one of the strata. To compare disproportional to proportional allocation, for example, suppose the population is composed of $M$ rare sampling units and $N–M$ nonrare units, it is desired to estimate the mean, $\mu = M^{-1}\sum_{i=1}^{M} y_i$, of a certain variable for members of the rare subset of the population using stratified random sampling. Let the population be divided into two strata such that stratum 1 contains a proportion, $A \left(= \frac{M_1}{M}\right)$, of the rare population with the remainder $(1 - A)$ in stratum 2. Simple random samples of sizes $n_1$ and $n_2$ are taken from the strata where $m_1(\leq n_1)$ and $m_2(\leq n_2)$ are sampled members of the rare populations. The unbiased estimator of $\mu$ can be written in terms of $A$ as

$$\hat{\mu}_{st}^{A} = A\overline{y}_1 + (1 - A)\overline{y}_2, \tag{1}$$

where $\overline{y}_h = m_h^{-1}\sum_{j=1}^{m_h} y_{hj}$. If $m_1$ and $m_2$ are sufficiently large, the variance of (1) is approximately

$$v(\hat{\mu}_{st}^{A}) \cong A^2\frac{\sigma_1^2}{E[m_1]} + (1 - A)^2\frac{\sigma_2^2}{E[m_2]}, \tag{2}$$

where $\sigma_h^2 = (M_h - 1)^{-1}\sum_{j=1}^{M_h}(y_{hj} - \mu_h)^2$ with $\mu_h = M_h^{-1}\sum_{j=1}^{M_h} y_{hj}$, and $E[m_h]$ is the expected value of $m_h$ (Kalton and Anderson, 1986). To show the efficiency of disproportionate sampling relative to proportional sampling, let $W_h = \frac{N_h}{N}$ be the proportion of the total population in stratum $h$, $P = \frac{M}{N}$ be the proportion of the total population that is rare, $P_h = \frac{M_h}{N_h}$ be the proportion of units in the $h$th stratum that is rare, and $kf_2$ and $f_2$ be the sampling fractions in strata 1 and 2, respectively. Let $c$ be the ratio of the cost of sampling a member of the rare subset to that of a member in the nonrare subset of the population. Further, assume $\sigma_1^2 = \sigma_2^2$, which is not unreasonable since we are confining the estimation to the values of $y$ for the rare subset of the population only and are not stratifying based on the values of $y$ but instead on whether the unit is a member of the rare subset. Then, Kalton and Anderson (1986) showed that the ratio of the variance of $\hat{\mu}_{st}^{A}$ under disproportional allocation to proportional allocation is approximately

$$R \cong \frac{[kP - (k-1)W_1P_1][(c-1)\{P + (k-1)W_1P_1\} + (k-1)W_1 + 1]}{kP[(c-1)P + 1]}. \tag{3}$$

The optimal choice of $k$, the ratio of the sampling fractions in the two strata, is given by

$$\widetilde{k} = \sqrt{\frac{P_1[(c-1)(P - W_1P_1) + (1 - W_1)]}{(P - W_1P_1)[(c-1)P_1 + 1]}} \tag{4}$$

which reduces to $\widetilde{k} = \sqrt{\frac{P_1}{P_2}}$ when $c = 1$ (Kalton and Anderson, 1986). The degree to which the disproportional sampling outperforms the proportional sampling depends on both $A$ and $P_1$; the larger $A$ or $P_1$ is, the better disproportional sampling is relative to proportional sampling. Hence, the ideal condition to perform disproportional stratified random sampling is when most of the rare subset of the population is confined to a single stratum (large $A$) and that stratum has few nonrare elements (large $P_1$).

When the distribution of the rare elements among the sampling units is not known beforehand, a method for stratifying the population is desired. Here, we discuss two approaches: (1) two-phase sampling for stratification (Kalton and Anderson, 1986; Thompson, 2002; see also Chapter 3 of this volume) and (2) model-based approach. In two-phase sampling, at the first phase, units are screened using easily measured and inexpensive variables that are highly correlated with the rare trait. The sample is then divided into two or more strata according to the probability of being a member of the rare population estimated from the screening variables. A second sample is then taken from the now-stratified sampling units collected at the first phase. It is recommended that this phase of sampling use disproportional allocation. Fink et al. (2004), for example, used an initial screening questionnaire with eight questions to classify the incoming patients according to their risk for psychiatric disorders. The first-phase sample consisted of patients entering a hospital in Sweden and agreeing to answer the questionnaire. The answers were then used to stratify the sample into two groups, either at low risk or at high risk for psychiatric disorders. Second-phase sampling consisted of performing time-consuming detailed diagnostic interviews on all first-phase patients in the high-risk stratum and on a third of the individuals in the low-probability stratum. In this particular study, the ultimate goal was to determine the ability of the screening tool to identify severe disorders and so sampling the entire high-risk stratum was considered appropriate.

In two-phase sampling for stratification, a first-phase sample of size $n'$ is selected according to a probability-based design $D$ that yields an unbiased estimator of the population characteristic of interest. For example, the parameter might be the population mean $\mu_{\text{all}} = N^{-1}\sum_{i=1}^{N} y_i$ for a variable $Y$ which may take on nonzero values only for the population elements possessing the rare trait. Denote the estimator of the mean as $\hat{\mu}_{\text{all}}$. Since $\hat{\mu}_{\text{all}}$ is design-unbiased, $E_D[\hat{\mu}_{\text{all}}] = \mu_{\text{all}}$, with variance $V_D[\hat{\mu}_{\text{all}}]$ based on the design $D$. Before second-phase sampling, easily measured auxiliary information is used to classify the units in the first-phase sample into $H$ strata of sizes $n'_h$, $h = 1, \ldots, H$, $\sum_h n'_h = n$. The strata are constructed so that the rare units are congregated into one or few strata. Stratified random sampling, preferably with disproportional allocation, is now performed with $n_h$ units being sampled from $n'_h$, $h = 1, \ldots, H$. The estimator $\overline{y}_{\text{str}} = \sum_{h=1}^{H} \frac{n_h}{n'_h}\overline{y}_h = \sum_{h=1}^{H} \frac{1}{n'_h}\sum_{j=1}^{n_h} y_{hj}$ is conditionally unbiased for $\hat{\mu}_{\text{all}}$ and has conditional variance

$$\operatorname{var}(\overline{y}_{\text{str}}|\mathbf{y}_n) = \sum_{h=1}^{H}\left(\frac{n_h}{n'_h}\right)^2 \frac{\sigma^2_{h|\mathbf{y}_h}}{n_h},$$

where $\mathbf{y}_n$ is the vector of observations from the first phase, $\mathbf{y}_h$ is the subvector of $\mathbf{y}_n$ assigned to stratum $h$, and $\sigma^2_{h|\mathbf{y}_h}$ is the population variance of $\mathbf{y}_h$. Then, the unconditional mean and variance of $\overline{y}_{\text{str}}$ are $E_D[E_{\text{str}|D}[\overline{y}_{\text{str}}]] = E_D[\hat{\mu}_{\text{all}}] = \mu_{\text{all}}$ and

$$\begin{aligned}\operatorname{var}(\overline{y}_{\text{str}}) &= \operatorname{var}_D(E_{\text{str}|D}[\overline{y}_{\text{str}}]) + E_D[\operatorname{var}_{\text{str}|D}(\overline{y}_{\text{str}})]\\ &= \operatorname{var}_D(\hat{\mu}_{\text{all}}) + E_D\left[\sum_{h=1}^{H}\left(\frac{n_h}{n'_h}\right)^2 \frac{\sigma^2_{h|\mathbf{y}_h}}{n_h}\right].\end{aligned} \tag{5}$$

Hence, the variance is greater than would be obtained if the strata were constructed in advance of sampling, but it can be reduced by judicious choice of the sampling design for the first-phase sample and by performing disproportionate allocation for the second-phase stratified random sample.

Another approach to stratification uses model-based approaches to construct strata where at least one stratum is highly likely to contain the rare elements. Edwards et al. (2005), for example, hypothesized that the presence of four common species of lichens is correlated with the presence of several rare lichens. Using data collected at the nodes of a systematic grid placed over the study region, they used classification trees to predict the presence of the common species based on readily available data, such as topographic and bioclimatic variables. The resulting estimates were interpolated to produce probability maps which were then used to stratify the study area for subsequent sampling of rare species. They allocated stratum sample sizes proportional to the predicted probability of the common species. In a comparison of this approach to a systematic sampling design, detections of rare species increased from 1.2- to 5-fold for four of the five rare species.

Two-phase sampling has also been used for adaptive allocation of sample sizes to predefined strata when the population is rare. The first-phase sample is taken according to a stratified random design, possibly with Neyman allocation, and the observations are reviewed to determine which strata should receive additional sampling effort. The first description of the use of this approach was given by Francis (1984) who used adaptive allocation to estimate fish biomass from sampling at sea.

Consider sampling to estimate a population total $\tau = \sum_{h=1}^{H} \sum_{j=1}^{N_h} y_{hj}$ for some variable $Y$. For example, when sampling a spatially, highly clustered population, $y_{hj}$ might be the number of population elements in the $j$th sampling unit (e.g., PSU) within the $h$th stratum. Optimal allocation of the samples to strata often cannot be done *a priori* because of incomplete knowledge of the variances in the different strata. Instead, an initial stratified random sample with $n_{h1}$ samples in stratum $h$ is performed and the sample stratum variance $s_{h1}^2$, $h = 1, \ldots, H$, is calculated. Now, for highly clustered populations, $s_{h1}^2$ will be largest for those strata in which PSUs have large counts. Hence, additional samples should be assigned when possible to those strata. Francis (1984) recommended a sequential allocation of additional effort as follows. For each stratum, first calculate the stratum variance of the estimator from the first-phase sample. In the case of estimating the total, the estimator at the stratum level is $\hat{\tau}_{h1} = N_h \overline{y}_{h1}$, where $\overline{y}_{h1}$ is the sample mean per PSU, and $N_h$ is the total number of PSUs in stratum $h$ with estimated variance $\hat{v}(\hat{\tau}_{h1}) = \frac{N_h^2 s_{h1}^2}{n_{h1}}$. Then, the difference in variance if one additional sample is taken is approximated as follows:

$$G_h = N_h^2 s_{h1}^2 \left( \frac{1}{n_{h1}} - \frac{1}{n_{h1} + 1} \right) = \frac{N_h^2 s_{h1}^2}{n_{h1}(n_{h1} + 1)}.$$

One additional sample is taken in the stratum with the largest value of $G_h$ and a new $G_h$ is calculated. Another sample is taken in the stratum with the now largest $G_h$; this could be the same or a new stratum. The sequential sampling is repeated until the desired total sample size is reached. Since the initial estimates of the stratum variance are used throughout, the second-phase sampling allocation can be determined before the actual second-phase sampling effort is begun. Francis (1984) and Thompson and Seber (1996) described the use of this approach for a study of fish biomass based on stratified random tows of different lengths. Francis (1984) recommended combining the data from the two phases and using the usual estimator for stratified random sampling. In our description that would be

$$\hat{\tau}_h = N_h \overline{y}_h = N_h \frac{\sum_{j=1}^{n'_h} y_{hj}}{n'_h}, \tag{6}$$

where $n'_h$ is the final sample size $(n_{h1} + n_{h2})$ with variance

$$\hat{v}(\hat{\tau}_h) = N_h^2 \hat{v}(\overline{y}_h) = N_h^2 \frac{\sum_{j=1}^{n'_h} (y_{hj} - \overline{y}_h)^2}{n'_h(n'_h - 1)} = N_h^2 \frac{s_h'^2}{n'_h}.$$

For the population total,

$$\hat{\tau} = \sum_{h=1}^{H} \frac{N_h}{N} \hat{\tau}_h.$$

The variance of $\hat{\tau}$ is estimated by $\hat{v}(\hat{\tau}) = \sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 \hat{v}(\hat{\tau}_h)$.

Both $\hat{\tau}$ and $\hat{v}(\hat{\tau})$ are biased due to the adaptive allocation. Francis (1984) found that the bias decreased as the ratio $n_1/n$ approached 1 but that the efficiency of the adaptive allocation sampling also decreased as the ratio increased. Thompson and Seber (1996) used the Rao–Blackwell method to derive unbiased estimators of the total and variance using the complete two-phase samples. Within each stratum, the unbiased estimators are given by $\hat{\mu}_{\mathrm{RB}h} = E[\overline{y}_{h1}|\mathbf{y}_{hR}]$ and $s^2_{\mathrm{RB}h} = E[s^2_{h1}|\mathbf{y}_{hR}]$, where $\mathbf{y}_{hR}$ is the set of $y$-values corresponding to the distinct units for the entire two-phase sample of size $n'_h$ and is the complete sufficient statistic for $\mu_h$. The expectations are functions of the permutations of $\mathbf{y}_{hR}$. They further show that $\hat{\mu}_{\mathrm{RB}h} \geq \overline{y}_h$and hence that $\overline{y}_h$ is negatively biased.

## 3. Adaptive sampling designs

### *3.1. Multiplicity or network sampling designs*

In multiplicity sampling, the $N$ population elements are associated with sampling units, such as $N'$ households. When a sampling unit (household) is selected, information is obtained both on the individuals within the household as well as on individuals in other households who are linked to those in the sampled household (Sirken, 1970). This information is collected from the sampled unit; the linkages are not directly sampled as a result of being identified by the sampled unit. Hence, multiplicity sampling is distinct from the related link-tracing designs in which linked sampling units are then randomly selected for additional measurement.

The intent is usually estimation of the prevalence of a rare trait within a larger population. In conventional sampling, the likelihood of observing the rare subset of the population is generally so low that collecting additional information through the multiplicity design should decrease the sampling variance. For example, in a study on the prevalence of a rare genetically based cancer (such as the gene *BRCA-1* associated with a high probability of breast cancer), the usual, conventional sampling approach would be a random selection of households as PSUs. Because of the rarity of the cancer, the sampling errors of the estimators could be very large. To reduce the sampling error, a multiplicity design would include asking the sampled households to report cancer patients in other households. To control the amount of information generated, the reports for other households are limited by some counting rule. Examples of rules include reporting only siblings of the cancer patient who have cancer not living in the sampled household or reporting all children of the cancer patient not living in the household. In multiplicity sampling, every individual in the rare population is assumed to be linked with at least one sampling unit (e.g., household), such as the unit in which the individual is located or by being reported by another sampling unit under the imposed counting rule. The number of links leading to an individual is said to be its multiplicity.

In multiplicity sampling, each sampling unit reports several pieces of data: the individuals in the unit with the rare trait, its links to individuals with the rare trait in other sampling units, and the multiplicity of each link. Using this information, the estimator of the proportion of the population with the rare trait, $P = \frac{M}{N}$, is

$$\hat{P} = \frac{N'/n'}{N} \left\{ \sum_{i=1}^{N'} d_i \sum_{j=1}^{M} \frac{(a_{ij1} + a_{ij2})}{(S_{j1} + S_{j2})} \right\}, \tag{7}$$

where $N'$ is the number of sampling units in the sampling frame, $n'$ is the size of the random sample of units, $d_i$ is the indictor variable that the $i$th sampling unit is selected for the sample, $a_{ijk}$ is the indictor variable whether the $j$th event is reported for the $i$th sampling unit by the $k$th counting rule used, and $S_{jk}$ is the multiplicity of the $j$th event reported by sampling units for the $k$th counting rule (Czaja et al., 1986). The first counting rule is for the conventional sampling design of random selection of sampling units and so $S_{j1} = 1$. The other counting rule is for the multiplicity design. Assuming that no more than one event is reported in each sampling unit and that sampling is random and without replacement, the expected value of $\hat{P}$ is

$$E[\hat{P}] = P \left[ \theta_2^* + \left(\theta_1^* - \theta_2^*\right) \frac{1}{M} \sum_{j=1}^{M} \frac{1}{(1 + S_{j2})} \right],$$

where $\theta_k^*$ is the conditional probability of reporting an event under counting rule $k$, given that it is linked to a sampled unit under rule $k$. The variance of $\hat{P}$ is given by

$$\begin{aligned} \mathrm{Var}[\hat{P}] = P\frac{N'/N}{n'} & \left\{ \theta_2^* \left( \frac{1}{M} \sum_{j=1}^{M} \frac{1}{(1 + S_{j2})} \right) \right. \\ & \left. + \left(\theta_1^* - \theta_2^*\right) \left( \frac{1}{M} \sum_{j=1}^{M} \frac{1}{(1 + S_{j2})^2} \right) \right\} - \frac{E[\hat{P}]^2}{n'} \end{aligned}$$

(Czaja et al., 1986). Levy (1977) expanded the sampling design to include stratified random sampling of the sampling units. He derived the stratified estimators of the prevalence that allow linkages across strata boundaries.

Sirken (1970) derived relationships for the differences between multiplicity and conventional estimators of the prevalence of a rare trait within a population. Three multiplicity counting rules were considered. He compared the variances of the resulting estimators to that obtained under conventional random sampling without replacement and multiplicity. Overall, multiplicity sampling tends to reduce sampling error compared with the conventional sampling approach without multiplicity, but the degree of gain in efficiency depends on the particular counting rule used.

Sirken (1970) also points out that the gain in efficiency of multiplicity sampling is offset by the response errors associated with the additional information about linkages. Here, response error could be due to incorrect reporting of the links associated with a sampling unit or of the multiplicities of the links or both. Nathan (1976) expanded on this by performing a small experiment that allowed partitioning of the total mean squared error (MSE) of the multiplicity estimator into components for sampling variance, response error, and response bias. The sampling variance is the usual sampling error obtained under the design assuming perfect information. The response bias is due to counting bias, the loss of individuals who are not linked to any sampling unit in the sampling frame, and implementation bias, the bias due to reporting individuals who do not belong to the population of interest due to misinterpretation of the counting rule. An example of the latter is the case where the counting rule is to ask for events that occurred in the last year, such as births, but births in prior years are reported as well.

Nathan (1976) compared three sampling designs for estimation of births in 1973 in Israel: conventional random sampling of households, multiplicity sampling with a restricted counting rule, and a full multiplicity sampling design with an expansive ("full") counting rule. In his example, the restricted rule was to include births linked to the mother and maternal grandmother of the parent giving birth in the sampled household. The full multiplicity counting rule included the restricted rule and added reporting of the mother's sisters as well. Data were collected according to the study design except that a subsample of the individuals reported as linkages were surveyed using the same counting rules. This provided the additional information that allowed estimation of the components of the MSE. Nathan (1976) also found that the two multiplicity methods had high variance due to response error, but the full multiplicity had lower sampling variance and bias. Hence, full multiplicity was the most efficient, and conventional sampling performed better than the restricted multiplicity due to the increase in response error variance. Czaja et al. (1986) showed that multiplicity sampling in general had higher reporting biases but considerably lower MSE than conventional sampling designs. The gain in efficiency depended on the counting rule and the population sampled.

### *3.2. Link-tracing sampling designs*

These designs also go by the moniker of graph sampling designs (Thompson and Collins, 2002) and are a form of adaptive sampling since the final set of sampled units depends on the observations taken during sampling. Examples of link-tracing designs include multiplicity sampling (described above), random walk sampling (Klovdahl, 1989), snowball sampling (Kalton and Anderson, 1986), and respondent-driven sampling (Heckathorn, 1997). In all link-tracing designs, when a selected sampling unit is measured, the data collected include information about the elements in the sampled unit as well as data on linkages to other units. Except for multiplicity sampling, the linkages are used to select subsequent sampling units for inclusion. The selection in a random walk sampling design, for example, is based on randomly selecting one link to be measured at the next sampling step. In snowball sampling, $L \geq 1$ linkages are included in the next steps of sampling. The number of steps or stages of sampling is usually selected before sampling commences and generally depends on the rarity or obscurity of the subpopulation under study. The final sample consists of the original sampled units (stage 0) plus all units selected in the subsequent stages (stages 1 to $L$).

There are a number of problems associated with these approaches (Erickson, 1979). Foremost is that the linkages are nonrandom and possibly inaccurately reported. In addition, elements with large numbers of links tend to be oversampled relative to the elements with small multiplicities. Also, response error when sampling human populations, such as unresponsive answers or lies, can severely compromise the data collection. This is most likely to occur when the trait of interest is an illegal activity, such as drug use. Further, by their nature, the final sample is not a random selection and as such the estimators are biased and the variances are difficult to compute. Finally, the entire snowball sample is a function of the initial sample and so if the initial sample is biased then the entire data set is biased (Heckathorn, 1997). An attempt to correct for some of the flaws of snowball sampling is respondent-driven sampling where incentives are used at each selection stage to gather unbiased data from the respondent himself as well as complete sets of the respondent's linkages that can be used for subsequent sampling. Heckathorn (1997) argues that these dual incentives overcome the difficulties of snowball or random walk sampling, both the reporting errors and the method of choice of the initial sample. He states that the sampling is a Markovian process so that the final sample is independent of the initial sample, and unbiased estimators and standard errors can be calculated based on some additional assumptions (Heckathorn, 2002). He developed methods for determining the number of stages required to obtain this independence. Once respondent-driven sampling is completed, a common method for estimating the standard error of the estimators is to assume that the final sample is a random sample from the population since analytic solutions for the variance of respondent-driven sampling are not available. This is, of course, a biased estimator and so should be avoided. Salganik (2006) recommended the use of bootstrapping instead of estimation of variances and use percentile method (Efron and Tibshirani, 1993) for constructing confidence intervals.

### *3.3. Adaptive cluster sampling designs*

Adaptive cluster sampling was first described as a single-stage cluster sampling strategy in which the size and the distribution of the clusters are unknown before sampling (see, e.g., Francis, 1984; Thompson, 1990). Here, clusters refer to a group of secondary units, some of which display a characteristic which is of interest to the sample survey. In this sampling design, secondary units are selected and if a secondary unit demonstrates a particular attribute, then the primary unit of which it is a member is exhaustively sampled.

One of the earliest descriptions of a type of adaptive cluster sampling was given in Fisher (1934) in which it was of interest to estimate the proportion of children with albinism from parents who could produce children with albinism. This is distinct from the proportion of albinos in the general population; Fisher's interest was in comparing two sampling methods for determining if albinism is a Mendelian recessive gene. Fisher's "Sib Method" samples individuals; if the individual is an albino, they are asked about albinism of their siblings. Here, the secondary unit is the individual. If the individual had albinism, then the cluster (siblings) of which she/he is a member is then sampled. In Fisher's case, he did not explicitly indicate that individuals were sampled to determine if they were albinos; more likely a list frame from medical records was available so that the initial sample could be restricted to those of interest. Adaptive cluster sampling would be used if that list frame were unavailable as might be the case for more confidential information.

To conduct adaptive cluster sampling, a more formal definition of cluster is required. Clusters are constructed based on two pieces of information: the linkages among sampling units and the defining characteristic of a cluster. For example, a common approach in an ecological study of population size in a spatial setting is to construct a sampling frame composed of a set of contiguous, non-overlapping quadrats or grid cells in which $Y_i$ is the number of individuals within the $i$th cell. Linkages among the cells are based on attributes, such as common boundaries, distance between centroids, orientation relative to each other, or similar. For example, in the case of a regular grid, the linkages for the $i$th cell might be to cells with a border in common with the $i$th cell. The set of linkages for a cell is often referred to as the cell's neighborhood. For unbiased estimation, it is required that the linkages among units within a cluster be symmetric, that is, if the $i$th cell is linked to the $j$th cell, then the converse is also true. For the determination of what constitutes a cluster, that is, when the linkages are to be used in sampling, the researcher specifies a criterion for initiating the adaptive cluster sampling. This is similar to the counting rule used in multiplicity sampling. The choice for criterion is based on the $Y$-variable, such as $\{Y > c\}$ for $c$ a constant. A common choice, especially when the population is very rare, is $\{Y > 0\}$. The adaptive component of sampling is performed when a sampled unit, either in the initial sample or via a link to the sampled unit, meets the criterion $\{Y > c\}$. Occasionally, it is not possible to pick the value of $c$ before sampling, and so another approach is to base the criterion for adaptive cluster sampling on the order statistics observed in the initial sample (Thompson, 1996).

The set of units that would be sampled as a result of any single one of them being intersected in the initial sample is often referred to as a network; hence a requirement of a network is that every unit in the network is linked to every other unit in the network. A cluster then is composed of the network plus any additional units that would be sampled but which are not part of the network.

Any unit in the initial sample which does not satisfy the criterion to perform adaptive sampling is said to be a network of size 1. In adaptive cluster sampling, it is possible to sample units that are not part of the network. For example, consider a spatial region divided into square quadrats for sampling purposes; the linkages are defined to be the cells to the north or south. So, if a cell is selected in the initial sample and meets the criterion for adaptively sampling the links, then the adjacent cells to the north or south are sampled. If either of these cells meet the condition, the next cell(s) to the north or south is sampled. Sampling continues until either a boundary is encountered or a cell which does not meet the condition is measured. These last cells are not within the network of the cluster but are measured to determine the spatial extent of the cluster. These are often referred to as "edge units." The final sample consists of the units in the initial sample plus all units belonging to the networks intersected by the initial sample plus those edge units which are not part of any network.

Hence, the definition of a cluster is chosen before sampling commences, but a sampling frame listing the actual clusters in the population is not available. Instead, an initial sample of secondary units is selected using a probabilistic sampling design; if a sampled secondary unit has the particular attribute that is of interest, then sampling of the entire cluster in which it is a member is performed. The adaptive aspect of the design is the additional sampling for other members of the cluster when at least one member of the cluster is sampled. As a result, the initial sample size may be controlled, but the final sample size is random since the size of clusters that are sampled as a result of the initially sampled individuals' meeting the criterion for cluster sampling. To use Fisher's Sib Method as an example, the final sample size is the number of individuals sampled to identify those with albinism plus all the siblings of individuals with albinism were also sampled.

Besides not knowing the clusters *a priori*, another aspect that distinguishes adaptive cluster sampling from one-stage cluster sampling is that exhaustive sampling of a cluster includes secondary units that do not belong to the cluster. For example, in a study of the spatial distribution of weeds in an agricultural plot, it might be of interest to determine the number and spatial extent of weed clusters within the plot. Adaptive cluster sampling could be used first by taking an random sample of locations and by determining the presence of the weed at those sites. If the weed is present, then the adaptive sampling method could call for sampling for presence in one-meter quadrats surrounding the randomly sampled location. If those quadrats contain weeds, then the method calls for sampling quadrats around that site. This continues until the entire cluster of weeds has been delineated as a set of contiguous one-meter quadrats with weeds surrounded by a sample of quadrats without weeds. If the rule is to continue sampling in nearby 1 meter quadrats when weeds are present at a sampled location, then the final sample size is controlled only by the spatial distribution of the weeds and their proximity to each other. Hence, recent work has considered restricted or incomplete sampling of the cluster to control the final sample size (Lo et al., 1997; Salehi and Seber, 2002); in these designs, sampling stops once some predetermined number of members of the cluster has been sampled. Brown and Manly (1998) recommended an alternative approach for controlling the sample size which is based on sequential sampling of complete clusters and restricting the number of clusters sampled rather than the number of units within clusters.

The method of selecting of the initial sample in adaptive cluster sampling has been studied by Thompson (1990; simple random sampling), Salehi and Seber (1997a; simple random sampling without replacement of clusters), Thompson (1991a; one-stage cluster sampling, such as a systematic sample), Thompson (1991b; stratified random sampling), Borkowski (1999; Latin square sample), Christman and Lan (2001; sequential sampling with various stopping rules), Christman (2003; Markov chain one-per-stratum designs (see Breidt, 1995)), and Salehi and Seber (1997b; two-stage cluster sampling).

Unbiased estimation of population totals or means in unrestricted adaptive cluster sampling is accomplished using a modified version of either Horvitz and Thompson (1952) or Hansen and Hurwitz (1943) estimators. Both estimators weight each sampling unit's observation with the inverse of the probability of inclusion of that sampling unit. The distinction is that the Horvitz–Thompson estimator (HT) is used for sampling without replacement and includes each observation exactly once in the estimator while the Hansen–Hurwitz estimator (HH) for with replacement sampling includes observations as many times as they have been observed in the sample. As a result, the HH estimator is often easier to calculate than the HT estimator but generally has higher variance. The usual HH estimator of the population mean is

$$\hat{\mu}_{\mathrm{HH}} = \frac{1}{nN}\sum_{i=1}^{n}\frac{Y_i}{\alpha_i},$$

where $\alpha_i$ is the probability that the $i$th sampling unit is selected in a with-replacement sampling design. For adaptive cluster sampling, the estimator must be modified since $\alpha_i$ is calculable for units in the initial sample and units in the networks adaptively sampled from the initial sample but not for the edge units sampled as part of an adaptively sampled cluster. The modified HH estimator is given by

$$\hat{\mu}_{\mathrm{HH}} = \frac{1}{n'N}\sum_{i=1}^{n'}\frac{Y_i}{\alpha_i}, \tag{8}$$

where $n'$ equals the number of units selected in the initial sample plus the number of adaptively added units which met the criterion and so belonged to the networks, and $\alpha_i$ is now interpreted as the probability that the $i$th network is included in the sample. The variance of (8) is given by

$$\mathrm{var}(\hat{\mu}_{\mathrm{HH}}) = \frac{1}{N^2 n'}\sum_{i=1}^{N}\alpha_i\left(\frac{y_i}{\alpha_i} - N\mu\right)^2 \tag{9}$$

with unbiased estimator

$$\hat{v}(\hat{\mu}_{\mathrm{HH}}) = \frac{1}{N^2 n'(n'-1)}\sum_{i=1}^{n'}\left(\frac{y_i}{\alpha_i} - N\hat{\mu}_{\mathrm{HH}}\right)^2.$$

The HT estimator is similarly modified and is given by

$$\hat{\mu}_{\mathrm{HT}} = \frac{1}{N}\sum_{i=1}^{\nu}\frac{Y_i}{\pi_i},$$

where $\nu$ is the number of distinct units in the sample which belong to a network, and $\pi_i$ is the probability that the $i$th network is intersected by the initial sample. An alternative formulation uses the sums of the $y$-values of the sampled networks and is given by

$$\hat{\mu}_{\mathrm{HT}} = \frac{1}{N}\sum_{i=1}^{\kappa}\frac{Y_i^*}{\pi_i}, \tag{10}$$

where $\kappa$ is the number of sampled networks, and $Y_i^*$ is the sum of the $y$-values in the $i$th network. The variance of (10) is

$$\mathrm{var}(\hat{\mu}_{\mathrm{HT}}) = \frac{1}{N^2} \sum_{j=1}^{K} \sum_{k=1}^{K} Y_j^* Y_k^* \left( \frac{\pi_{jk} - \pi_j \pi_k}{\pi_j \pi_k} \right), \tag{11}$$

where $K$ is the number of networks in the population, and $\pi_{jj} = \pi_j$; an unbiased estimator of (11) is

$$\hat{v}(\hat{\mu}_{\mathrm{HT}}) = \frac{1}{N^2} \sum_{j=1}^{\kappa} \sum_{k=1}^{\kappa} \frac{Y_j^* Y_k^*}{\pi_{jk}} \left( \frac{\pi_{jk} - \pi_j \pi_k}{\pi_j \pi_k} \right).$$

Some examples of the intersection probabilities include

| Initial Sampling Design | $\pi_i^{\mathbf{a}}$ | $\pi_{ij}$ |
|---|---|---|
| Simple random sampling with replacement | $1 - \left(1 - \frac{x_i}{N}\right)^{n_1}$ | $1 - \left\{ \left(1 - \frac{x_i}{N}\right)^{n_1} + \left(1 - \frac{x_j}{N}\right)^{n_1} - \left(1 - \frac{x_i + x_j}{N}\right)^{n_1} \right\}$ |
| Simple random sampling without replacement | $1 - \left[ \binom{N - x_i}{n_1} \Big/ \binom{N}{n_1} \right]$ | $1 - \left[ \binom{N - x_i}{n_1} + \binom{N - x_j}{n_1} - \binom{N - x_i - x_k}{n_1} \right] \Big/ \binom{N}{n_1}$ |
| Stratified random sampling without replacement | $1 - \left[ \prod_{h=1}^{H} \binom{N_h - x_{hi}}{n_{h1}} \Big/ \binom{N_h}{n_{h1}} \right]$ | $1 - (1 - \pi_i) - (1 - \pi_j) + \left[ \prod_{h=1}^{H} \binom{N_h - x_{hi} - x_{hj}}{n_{h1}} \Big/ \binom{N_h}{n_{h1}} \right]$ |

[a] $N$ is the number of sampling units in the population, $n_1$ is the initial sample size, $x_i$ is the number of units in the $i$th network, and $x_{hi}$ is the number of units in the $i$th network of the $h$th stratum.

For unrestricted adaptive cluster sampling, it has been shown that adaptive cluster sampling outperforms nonadaptive sampling in most cases where the population is spatially rare, that is, when the individuals in the population tend to be highly clustered but are not necessarily rare in the sense of few individuals. Further, the efficiency of adaptive cluster sampling depends strongly on the within network variability, the definition of linkage (i.e., the neighborhood), the criterion for which adaptive sampling is initiated, and the size of the networks (Brown, 2003; Christman, 1997; Smith et al., 1995). Each of these attributes are confounded, and so it is difficult to choose a single measure by which to decide if adaptive cluster sampling is more efficient than simple random sampling. High within cluster variability usually leads to adaptive cluster sampling that is more efficient than an equivalently sized random sample. Conversely, small network sizes usually imply high efficiency since the final sample size is constrained. The choice of criterion $c$ can be used to control sample size, but it influences the within network variability as well as the networks sizes. Both within network variability and network sizes generally decrease as $c$ is increased, but low network variability leads to less efficiency in adaptive cluster sampling, whereas low network sizes increase efficiency.

A comparison of adaptive cluster sampling with stratified random sampling with disproportionate allocation has shown that the stratified sample with disproportionate allocation tends to outperform adaptive cluster sampling (Christman, 2000). In a study of adaptive cluster sampling, Brown (1999) compared an initial random sample to an initial stratified random sample and found that stratification did not increase the efficiency of the adaptive cluster sampling in the populations. Brown (1999) also compared adaptive allocation with stratified sampling to adaptive cluster sampling without stratification and found that although the estimator in (6) is negatively biased, adaptive allocation produced smaller MSE than the adaptive cluster sampling design.

The use of order statistics to determine the criterion that initiates adaptive sampling or the use of methods that restrict the final sample size (Brown and Manly, 1998; Salehi and Seber, 2002) lead to biased HH and HT estimators. For adaptive cluster sampling with order statistics, Thompson (1996) derived an unbiased estimator based on the HH estimator but the derived estimator is less efficient than if order statistics had not been used. He also provided an unbiased estimator using the Rao–Blackwell method; this new estimator has lower variance than either the derived estimator or the estimator based on simple random sampling. Su and Quinn (2003) used simulations to compare the efficiencies of the biased HT estimator and the unbiased HH estimator when order statistics are used. They found that the use of higher quantiles for the cutoff that initiates the adaptive sampling component decreases the efficiency of both estimators, but that the bias of the HT estimator also decreased.

## 4. Experimental design

When sampling is in a controlled environment, for example, to estimate the proportion of defects in widgets or the number of bugs in computer code, the approaches for estimating the prevalence or population sizes are different than for the methods used in survey sampling. For experiments to estimate rare events, the estimation method is based on probability distributions and the variances are model based rather than design based. Here, an example of the use of such approaches for planned experiments is given.

Hedayat and Sinha (2003) described a method for estimating the proportion of a certain brand of toys resulting in accidents; the toys may or may not exhibit the defect and so several trials must be run for each toy to estimate the probability that a toy will display the defect. The method is based on sampling $N$ toys and subjecting each of them to $k$ trials. Three outcomes are possible in the trials: the toy malfunctions, the toy is known not to malfunction, or the functioning is unknown after the $k$ trials. The estimator of the probability of observing the defect in the $k$ trials is a function of the ratio of the number of defects observed (at most one per toy selected) to the total number of trials ($kN$). The estimator is derived under the assumption of a multinomial distribution but is biased for small sample sizes. Since it is a maximum likelihood estimator, it is consistent and so the bias is smaller for larger sample sizes.

## 5. Confidence interval estimation

When the variables under study are characteristics, such as prevalence of a rare trait or the sum of a $Y$-variable for only those individuals in the population with the rare trait, the construction of confidence intervals using the assumption of asymptotic normality is sometimes appropriate. This is especially so when the $Y$-variable is well behaved, that is, not highly skewed, or when the sampling strategy is designed to obtain sufficient numbers of the rare elements. On the other hand, when the variables under study are attributes like counts, the empirical frequency distribution of the counts is usually highly skewed with an excess number of zeroes even when sampling is designed for rare populations (see, e.g., Christman and Pontius, 2000). As a result, the estimators of population characteristics are often highly skewed as well and use of symmetric confidence intervals is inappropriate. Alternatives that have been recommended include trimming and transformation (Keselman et al., 2002), bootstrapping (Brown and Manly, 1998; Christman and Pontius, 2000; Di Battista, 2003; Salganik, 2006), and jackknifing (Di Battista, 2003). Keselman et al. (2002) described various combinations of approaches to estimate when distributions are highly skewed and non-normal. They favored a combination of trimming and transformation of the data remaining after trimming and recommended bootstrapping to estimate variance. Christman and Pontius (2000) studied several bootstrapping approaches for adaptive cluster sampling of finite populations to develop confidence intervals around the HH estimator for estimating the means or abundances. Di Battista (2003) discussed jackknifing and bootstrapping for adaptive cluster sampling based on simple random sampling with and without replacement. Brown and Manly (1998) described bootstrapping for without replacement sampling using the method by Booth et al. (1994).

## 6. Summary

Several methods are available to study rare populations and the choice depends on the population under study, the objectives of the study, and the distribution of the rare elements to be sampled. None of the methods will perform well at small samples sizes when the elements are very rare and found in small groups of one or two. When the population has internal linkages such as are common when studying human populations where it is likely that individuals with the rare trait are familiar with others who exhibit the same rare trait, then the link-tracing designs are available. The advantages of these methods are that a large gain in information is available from a relatively small initial sample, but the dependence on accurate responses and identification of linkages is critical to the success of the methods. When the population is spatially distributed and the rare elements occur in spatially distinct groupings of reasonable sizes, then a method such as adaptive cluster sampling can be used to estimate population parameters. The availability of several different sampling strategies for the initial sampling allows for a variety of approaches to accurately and precisely estimate the parameters of interest. Overall, the final choice of design will depend on the cost and objectives of the study.

*Sample Surveys: Design, Methods and Applications, Vol. 29A*
ISSN: 0169-7161

DOI: 10.1016/S0169-7161(08)00007-2

7

# Design, Conduct, and Analysis of Random-Digit Dialing Surveys

*Kirk Wolter, Sadeq Chowdhury and Jenny Kelly*

## 1. Introduction

Random-digit dialing (RDD) is a method of probability sampling that provides a sample of households, families, or persons via a random selection of their telephone numbers. For simplicity of explication, we use the person as the final unit of analysis in this article; yet, virtually, all our comments and methods extend naturally to the household or family.

In this chapter, we discuss the design, conduct, and analysis of RDD surveys primarily in the context of large-scale work performed in the United States. We believe that the material generalizes to other countries with an established landline infrastructure. In the United States, there is generally no sampling frame that enables a direct sampling of persons. RDD changes the sampling unit from the person to the telephone number, for which sampling frames do exist. Then, people can be sampled indirectly through their telephone numbers, enabling valid inferences to populations of people.

In the modern era, the RDD survey has come to embody the following three elements: (1) random sampling of telephone numbers from a listing of all (or most) assigned telephone numbers; (2) dialing the selected numbers from a central call center(s); and (3) administering the survey questionnaire to residential respondents via a system of computer-assisted telephone interviewing (CATI). RDD surveys became an accepted form of survey research in the 1970s, and their prevalence increased considerably in the 1980s and 1990s.

Today, the RDD survey stands as one of the dominant survey forms for social-science and market research. It has attained this position because it offers important advantages in cost, timing, and accuracy. RDD surveys eliminate travel costs and, thus, are far less costly than surveys that use face-to-face interviewing methods. They may be more expensive, however, than mail or web surveys that shed labor costs by eliminating or reducing the use of human interviews. RDD surveys have the capacity to deliver survey information very quickly relative to surveys that use other modes of enumeration. They may be launched and their interviewing operations completed quickly in a matter of days or weeks. The survey questionnaire—possibly even with elaborate skip patterns, large lookup tables, and other complications—can be entered quickly into the CATI software. Since the interview data are already in machine-readable form, entered by the interviewer into the CATI system, there is no need for a data conversion from paper to computer format. Mail and web surveys may match RDD surveys in terms of start-up time, whereas only web surveys can match them in terms of data collection and delivery speed. Face-to-face surveys are much slower to launch, to complete data collection, and to deliver the survey data.

RDD surveys, and generally CATI surveys, also possess features that enhance data quality. Because interviewing operations are usually centralized in a small number of call centers, it is possible to achieve specified standards relating to the hiring, qualifications, and training of the interviewers and their supervisors. Supervisors can monitor interviews and the general performance of the interviewers; they can take corrective actions in real time through retraining or replacement of underperforming interviewers. A work force that undergoes continuous improvement has the capacity to produce better and better interview data. Computer edits can be built into the CATI instrument, thus limiting missing values and the possible entry of out-of-range values, erroneous skip patterns, and the like. Face-to-face surveys that use a system of computer-assisted personal interviewing (CAPI) can incorporate the advantages of online edits, yet such surveys, with or without CAPI, cannot match CATI surveys in terms of close, real-time monitoring of interviewers. Face-to-face surveys, however, can generally achieve higher response rates than CATI surveys; they may achieve greater data completeness relative to certain types of questions or subject matter (e.g., it is obviously not possible to collect biomarkers in a pure CATI survey) and they usually have the capacity to handle a longer interview. Mail surveys generally experience lower response rates than CATI surveys; they cannot offer the benefits of online edits and cannot accommodate complicated skip patterns. Web surveys may offer online edits, yet their response rates will likely be lower than those of RDD surveys. Depending on the target population, there may be no acceptable sampling frame to support the use of a web-interviewing approach, and thus, web surveys would tend to have lower coverage than RDD surveys.

Survey planning and the selection of a mode of interview require consideration of many complex trade-offs between cost, speed, and accuracy. The foregoing discussion reveals many circumstances in which the RDD survey will be preferred and demonstrates why RDD surveys have reached a dominant position in the survey research marketplace.

Before proceeding, it is also important to observe that modern surveys increasingly use multiple modes of enumeration to collect acceptable data in the fastest time feasible at an affordable price. While we do not explicitly treat mixed-mode surveys in this chapter, the methods we do present show how the telephone component of a mixed-mode survey may operate.

## 2. Design of RDD surveys

### *2.1. Structure of telephone numbers*

The telephone industry in the United States consists of many individual telephone companies. Each company manages a block of telephone numbers and assigns those numbers to their subscribers. Subscribers typically maintain their numbers through time, regardless of switching telephone companies. The telephone number itself consists of eleven digits as depicted in the following diagram:

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | - | N | P | A | - | N | X | X | - | H | I | J | K |

The first digit in the structure (i.e., 1) is a constant, which is the international county code for the United States, and is required to be dialed first for calling outside the local calling area and in some cases even within the local area. The three-digit area code is represented by the symbol NPA, the three-digit exchange code or prefix is represented by the symbol NXX, and the four-digit suffix is represented by the symbol HIJK. The symbol NPA-NXX-H represents a 1000-block of numbers.

Area codes represent compact geographic areas; they are generally non-overlapping and exhaust the land area of the United States. Not all three-digit combinations are in service at this time, and in some cases, area codes overlay one another. The area codes are generally nested within states, but they do not generally correspond to political, postal, or census geography. Exchange codes within area codes or 1000-blocks within exchange are designated for landline telephones, wireless (or cell) telephones, or some other type of use. Historically, the exchange codes for cell telephones have been excluded from sampling for RDD surveys. Exchange codes for landline telephones are not geographically compact nor are boundaries defined in a useful way. The area covered by an exchange can cross city or county boundaries, making it difficult to use the exchange directly for any geographic stratification. The four-digit suffix within an exchange also contains no useful geographic information. However, since the telephone companies activate banks of consecutive numbers and assign numbers to residential and nonresidential subscribers in such a way that the consecutive suffixes within an exchange may be clustered in terms of working or nonworking or residential or nonresidential status, banks of suffixes have sometimes been used as clusters in the sample selection process.

In what follows, we describe sampling frames and sampling designs for telephone numbers. We refer to 100 consecutive telephone numbers—from 1-NPA-NXX-HI00 to 1-NPA-NXX-HI99—as a *100-bank* or simply a *bank* of numbers. At any given point in time, some banks have not been assigned and placed into service, some are assigned to landline telephone subscribers, some to cell-telephone subscribers, some to other types of use, and a few to mixed use.

### 2.2. *Sampling frames*

The main premise of RDD is that each eligible person in the survey target population can be linked to—that is, reached and interviewed at—one or more residential telephone numbers in the population of landline telephone numbers. Sampling telephone numbers with known probabilities of selection means that people are selected with calculable probabilities, and thus, that valid inferences can be made to the population of eligible people. Another key premise of RDD is that each telephone number is linked to a specific, identifiable (approximate) geographic location. This feature makes it possible to select representative samples of people in defined geographic areas such as states, counties, or cities.

At the outset, it is clear that RDD surveys provide no coverage of people who do not have ready access to a landline telephone in their home, including those who have access to a cell telephone only and those who have no access to any telephone. We discuss this undercoverage and approximate ways of adjusting for it later.

A major challenge for RDD surveys is the development of a complete list of telephone numbers that covers all the remaining persons in the target population, that is, all persons in the population with access to a landline telephone. A natural but naive approach would be to sample telephone numbers at random from residential telephone directories. This method is nearly always unsatisfactory because not all residential numbers are listed in a directory and not all listed numbers actually connect to an occupied residence. The former problem is the more acute since it would result in an additional bias of undercoverage to the extent that people with an unlisted telephone are different from people with a listed telephone with respect to the issues under study in the survey. Various authors have shown that the sociodemographic characteristics of the persons in households with unlisted telephone numbers are different from those of persons in households with listed telephone numbers (see, e.g., Brunner and Brunner, 1971; Fletcher and Thompson, 1974; Glasser and Metzger, 1975; Leuthold and Scheele, 1971; Roslow and Roslow, 1972; Shih, 1980). The latter problem leads to inefficiency but not bias because some telephone numbers will screen out as nonresidential.

As an alternative to the directory-based sampling frame, one could consider the use of the conceptual list of all telephone numbers in assigned landline telephone banks. Currently, there are over 718 million numbers in such banks in the United States. Although this frame covers both listed and unlisted numbers, it can be highly inefficient to work with. The major problem is that this frame includes a large percentage of nonworking and nonresidential telephone numbers. Dialing purely at random from such a frame would result in relatively few residential calls, many unproductive calls, and the consumption of excessive cost and time.

To make the conceptual list viable as a sampling frame, some method of sampling is needed to diminish the rate of out-of-scope calls and increase the rate of productive residential calls. In fact, such methods have been developed and the conceptual list of telephone numbers does provide the sampling frame for most RDD surveys conducted today. We provide a brief account of useful sampling methods in the next section.

### 2.3. *Sampling procedures*

Various sampling procedures have been developed and used over the years for random digit dialing (Glasser and Metzger, 1972; Hauck and Cox, 1976; Sudman, 1973; Tucker et al., 1992; Waksberg, 1978). Lepkowski (1988) provides an extensive discussion of telephone sampling methods used in the United States until the late 1980s. We have already noted that pure random sampling of telephone numbers is unworkable on grounds of high cost and time. We proceed to describe three of many methods of sample selection that have found favor in RDD surveys over the years through increases in cost-effectiveness.

Sudman (1973) described a procedure in which directory-based sampling is combined with RDD sampling. The procedure considers each block of 1000 consecutive telephone numbers as a cluster, and the clusters are selected by obtaining a simple random (or systematic) sample of numbers from a telephone directory and the corresponding clusters of all selected numbers are included in the sample. Then, calls within each selected cluster are made using RDD sampling to reach a predetermined number of households with listed telephone numbers. Because the sampling of clusters is based on the directory-listed numbers, the clusters are selected with probability proportional to the (unknown) count of listed telephone numbers in the cluster. The procedure improves over the unrestricted method by concentrating on the banks with one or more listed telephone numbers, but it still requires making a large number of calls to reach the predetermined number of listed telephone numbers. Also, since the initial sample is selected based on a telephone directory, the procedure introduces a small bias of undercoverage by omitting the clusters with recently activated numbers or with no listed numbers.

During the 1970s and 1980s, the *Mitofsky–Waksberg method* (Waksberg, 1978) was used widely in RDD surveys. It involves a two-stage sampling procedure, consisting of (i) a selection of banks at the first stage, using Lahiri's (1951) rejective method for probability proportional to size (PPS) sampling, and (ii) a random selection of telephone numbers within selected banks at the second stage. The idea is to select banks with probability proportional to the number of working residential numbers (WRNs) they contain. Then, telephone numbers are sampled in the selected banks, which tend to be rich in WRNs. In this manner, dialing of unproductive numbers (nonworking or nonresidential) is greatly reduced. Each bank is considered as a cluster or primary sampling unit (PSU). A random bank is first selected and then a random telephone number is dialed within the selected bank. If the dialed number is determined to be nonworking or nonresidential, the bank is rejected and excluded from the sample. However, if the number dialed is determined to be a WRN, then the bank is accepted and additional telephone numbers are selected from the bank at random and dialed, until a fixed number, $k$, of residential numbers is reached. The two-stage process is continued until a predetermined number of banks, $m$, is selected. The total sample size of WRNs is, therefore, $n = m\,(k+1)$. If $M$ denotes the number of banks in the sampling frame and $K_i$ is the number of WRNs in the population within the $i$th bank, then the probability that a given bank $i$ is selected and accepted at a given draw is

$$\pi_i = p_i \frac{1}{1-\ddot{p}} = \frac{K_i}{M\overline{K}}, \tag{1}$$

where

$$p_i = \frac{1}{M}\frac{K_i}{100}, \tag{2}$$

$$\ddot{p} = \frac{1}{M}\sum_{i'=1}^{M}\left(1 - \frac{K_{i'}}{100}\right), \tag{3}$$

and $\overline{K} = \sum_{i'=1}^{M} K_{i'}/M$. Thus, the rejective selection method is a PPS sampling method, despite the fact that the measure of size $K_i$ is unknown at the time of sampling! Only banks with one or more WRNs have a nonzero probability of selection, thus reducing the number of unproductive calls. The conditional probability that a given WRN $j$ is selected, given that its bank $i$ is selected and accepted on the given draw, is simply $\pi_{j|i} = \min\{k+1, K_i\}/K_i$. Thus, the unconditional probability of selecting a given WRN on the given draw is

$$\pi_{ij} = \pi_i \pi_{j|i} = \frac{\min\{k+1, K_i\}}{M\overline{K}} = \frac{k+1}{M\overline{K}}. \tag{4}$$

Typically, $k$ is specified to be a small number, usually no more than 4 so that it is usually less than the number of WRNs in the bank. As long as $k < K_i$ for all $i$, all WRNs in the population have an equal probability of selection. The values of $m$ and $k$ can be optimized for a given ratio of the costs of an unproductive call to the costs of a productive call (including the cost of calling, interviewing, and processing).

While the Mitofsky–Waksberg sampling scheme makes considerable improvements over unrestricted random sampling, it runs into severe operational problems. Because of potential nonresponse – or at least a lag in response between the initial release of the number and its resolution – for the first selected telephone number, the critical determination of whether a given bank is accepted or rejected may be unacceptably delayed. In addition, because of the risk of subsequent nonresponse within the bank or because of small $K_i$, it is sometimes difficult to locate $k$ residential numbers. Due to these operational problems and since another method of sampling has proved to be successful, rejective selection has now sharply declined in use for modern, large-scale RDD surveys.

*List-assisted sampling* is a term used to describe a class of methods that select the sample from the conceptual list of all telephone numbers in assigned landline banks while exploiting information in residential telephone directories to improve the efficiency of the selection. In 1+*sampling*, one restricts the sampling frame to all assigned landline banks in which one or more telephone numbers are listed in the residential telephone directory. Banks containing zero listed numbers are dropped. Then, a probability sample of telephone numbers is selected from the remaining banks. The method provides complete coverage of all listed and unlisted numbers in banks with at least one listed number, omitting only unlisted numbers in banks with no listed number. The scheme covers around 98% of the universe of landline telephone households and excludes only the approximately 2% of households with unlisted numbers in zero banks (Giesbrecht et al., 1996; Fahimi et al., 2008; Boyle et al., 2009). The bias due to the noncoverage of these unlisted telephone households in zero banks is thought to be small in many surveys (Brick et al., 1995). The method easily extends to $p$+ sampling, where $p$ is a small integer, say between 1 and 5. The larger the $p$ is, the greater the rate of productive calls, saving time and money. The downside of a larger $p$ is a reduced population coverage rate and an increased risk of bias. Ultimately, $p$+ sampling seems to have the following three virtues:

- Easy to implement
- Acceptably small undercoverage bias, especially for 1+ designs
- Yields an unclustered sample with a smaller design effect and larger effective sample size than the aforementioned clustered sampling designs

As a result, $p$+ sampling has emerged as the dominant form of RDD sampling today.

Table 1 describes our recent experience at NORC in calling 1+ samples. Approximately 24% of the telephone numbers may be classified as WRNs and 59% as something else, such as disconnected lines or businesses. Almost 17% of telephone numbers cannot be classified either way, meaning that information is incomplete and it is not possible to resolve whether the numbers are residential or not, despite repeated callbacks.

Table 1
Directory-listed status and working residential number status for U.S. telephone numbers in 1 + banks: 2005

| Directory-Listed Telephone Status | Working Residential Number Status | | | |
|---|---|---|---|---|
| | No | Yes | Not Resolved | Total |
| No | 49.1% | 4.4% | 8.0% | 61.5% |
| Yes | 10.0% | 19.9% | 8.6% | 38.5% |
| Total | 59.0% | 24.3% | 16.7% | 100.0% |

*Note*: The proportion of unresolved numbers depends in part on the calling protocol and on the length of the data-collection period. The proportions we cite are in the context of social-science surveys conducted by NORC with ample periods for data collection and relatively high response rates.

Almost 39% of telephone numbers are listed in a directory, whereas 61% are not. In all, 10% of numbers are directory-listed but turn out not to be WRNs and, on the other hand, 4% of numbers are not directory listed yet turn out to be WRNs. Lags in the development of directories and the timing of the field period explain the apparent misclassifications. From these data, one can calculate that 82% of resolved WRNs are listed in a directory. The remaining 18% of resolved WRNs are the unlisted. This unlisted percentage varies considerably from one part of the country to another. To restrict sampling to directory-listed numbers only would risk a bias in survey statistics, and the bias could be differential from one area to another, clouding comparisons.

### 2.4. *Stratification*

Little auxiliary information is typically available on the RDD sampling frame and opportunities for stratification of the sample are limited. Stratification is possible by directory-listed status or by whether a mailing address can be linked to the telephone number. Some broad geographic information is embedded within the structure of telephone numbers, which can be used for coarse geographic stratification. Since area codes are nested within states, stratification by state is feasible. Finer geographic stratification is difficult because exchanges may cross area boundaries.

It is possible to make an approximate assignment of telephone exchanges to finer census-defined geographic areas by geocoding the addresses of the listed telephone numbers within exchanges. Given a census-defined area of interest, one calculates the *hit rate* and the *coverage rate* for each exchange, where the hit rate is the proportion of listed telephone numbers in the exchange that belongs to the designated area, and the coverage rate is the proportion of listed telephone numbers in the designated area that is covered by the exchange. The sampling statistician may implement a rule involving these two factors to stratify exchanges by finer geographic areas, but this is an imperfect process. For example, each exchange may be classified to one of the set of geographically defined strata, spanning the target population, according to a majority rule; that is, one may assign the exchange to the stratum for which its hit rate is the maximum over the set of strata. As a second example, one may classify an exchange to a designated area if the hit rate exceeds a threshold, such as 0.05, and the cumulative coverage rate over all such exchanges exceeds another threshold, such as 0.95.

Stratification by socioeconomic status becomes possible by mapping census tracts onto telephone exchanges based upon the geocoded addresses of listed telephone numbers. Again, such mapping is necessarily approximate. Census variables at the tract level—such as race/ethnicity, age, sex, income, poverty, education, and housing tenure variables—can then be donated to the exchange and, in turn, exchanges can be assigned to socioeconomic strata. Among other things, such stratification enables one to oversample subpopulations of interest (Mohadjer, 1988; Wolter and Porras, 2002). Variables of this kind are sometimes called contextual or environmental variables.

### 2.5. *Determination of sample size*

The sample of telephone numbers selected for an RDD survey will need to be many times larger than the required number of completed interviews. During data-collection operations, there will be a number of losses to the sample, and the initial sample of telephone numbers must be large enough to offset the losses.

As usual, the statistician must begin by determining the effective number of complete interviews needed to achieve the survey's resource constraints and goals for statistical precision and power. Multiplying by the anticipated design effect—due to any clustering, differential sampling rates, and differential weighting effects—gives the target number of completed interviews.

The next step is to inflate the sample size to account for sample attrition due to nonresolution of telephone numbers, resolution of nonresidential telephone numbers, failure to complete the survey screening interview for some WRNs, the survey eligibility rate, and failure to complete the main interview among some eligible respondents.

The attrition starts with the nonresolution of many numbers regarding their WRN status. Despite repeated callbacks, it will be impossible to resolve many telephone numbers as to whether they are WRNs or something else. Then, among the resolved numbers, a large percentage will be non-WRNs, such as business numbers, computer modems, or disconnected lines. Once a WRN is identified, the next step is to conduct a brief screening interview to determine eligibility for the main survey. Screening for the eligible population is usually not possible beforehand because eligibility is not known at the time of sampling. (A special case is where all persons are eligible for the survey. In this case, there is in effect no screening interview.) Some screening interviews will not be completed because the respondent is never home or refuses to cooperate. Among completed screeners, households containing no eligible people are omitted (or "screened out") from the main interview. Finally, some main interviews will be missing because the eligible respondent is not at home or refuses to participate.

To determine the appropriate inflation of the sample size, the statistician must make assumptions about the foregoing factors based on general experience or information available from prior surveys. Ultimately, the target sample size in terms of telephone numbers in the $h$th sampling stratum is given by

$$n''_h = \frac{n'_h}{\pi_{h1}\,\pi_{h2}\,\pi_{h3}\,\pi_{h4}\,\pi_{h5}} = \frac{n_h D_h}{\pi_{h1}\,\pi_{h2}\,\pi_{h3}\,\pi_{h4}\,\pi_{h5}}, \tag{5}$$

where $n'_h$ is the target number of completed interviews, $n_h$ is the effective sample size required from stratum $h$, $D_h$ is the design effect assumed for stratum $h$, $\pi_{h1}$ is the resolution completion rate assumed in stratum $h$, $\pi_{h2}$ is the WRN rate among resolved numbers assumed in stratum $h$, $\pi_{h3}$ is the screener completion rate assumed in stratum

$h$, $\pi_{h4}$ is the eligibility rate among screened households assumed in stratum $h$, and $\pi_{h5}$ is the interview completion rate among eligible people assumed in stratum $h$.

## 2.6. *Emerging problems and solutions*

The RDD sampling frames and sampling designs discussed so far cover the population of people who reside in households with at least one landline telephone and fail to cover people who live in cell-phone-only households and nontelephone households. This undercoverage was of little concern in the early cell-telephone era. Yet at this writing, concern among survey researchers is growing. Using data from the National Health Interview Survey, Blumberg et al. (2006) show that 9.6% of adults (in the period January to June 2006) live in households with only cell-telephone service. Three years earlier (in the period January to June 2003), this population stood at only 2.8% of adults. Throughout this three-year period, adults in nontelephone households remained steady at 1.8–2.0% of the adult population. Some characteristics of the population with no telephone or only a cell telephone tend to be different from those of the population with a landline telephone. Thornberry and Massey (1988) discuss the patterns of landline telephone coverage across time and subgroups in the United States for the period of 1963–1986. Adults living with unrelated roommates tend to be cell-telephone-only at a higher rate than other adults. Other domains displaying a higher rate of cell-phone-only status include renters, young adults (age 18–24), males, adults in poverty, and residents of the South and Midwest regions. The population with no telephone service is likely to be unemployed, less well educated, below the poverty line, or in older age groups (Blumberg et al., 2006, 2005; Khare and Chowdhury, 2006; Tucker et al., 2007). Tucker et al. (2007) also show the rapid growth of the cell-telephone population using data from the Consumer Expenditure Survey and the Current Population Survey. Because of the current size of the populations without a landline telephone and the likelihood of their continued growth, there is increasing concern about potential bias in standard RDD surveys that omit these populations. What is not known for sure is whether cell-only and nontelephone populations differ from landline populations with respect to the main characteristics under study in surveys.

Until now, the main approach used to compensate for the undercoverage in RDD survey statistics has been the use of various calibration adjustments in the weighting process (Brick et al., 1996; Frankel et al., 2003; Keeter, 1995; Khare and Chowdhury, 2006). With the rapidly increasing cell-phone-only population, consideration must now be given to direct interviewing of this population. A dual-frame approach with supplementation of the RDD frame by an area-probability frame has received consideration in the past, but the approach is too expensive for general use. Dual frame designs using the traditional RDD frame of landline telephone numbers and a supplementary frame of numbers in cell-telephone banks must be considered.

Studies are being conducted to investigate the viability of interviewing respondents via cell telephones (Brick et al., 2007; Wolter, 2007). Sampling frames for cell-telephone numbers can be constructed from the Telcordia® TPM™ Data Source or Telemarketing Data Source. Both can be used to identify prefixes or 1000-blocks that are likely to be used for cell telephones. In addition, many cell-telephone numbers in use today are the result of "porting" the numbers from landline use to cell-telephone use. Such numbers are, by definition, located in landline blocks and are not covered by the type of sampling frame just described.

A number of additional problems for RDD surveys are emerging due to the recent rapid advancement in telephone technology. Voice over internet protocol (VoIP), which offers routing of telephone conversations over the internet or through any other IP-based network, is penetrating the mass market of telephony. Under the VoIP technology, a user may have a telephone number assigned under an exchange code in one city but can make and receive calls from another city or indeed anywhere in the world, just like a local call. The possibility of a universal telephone numbering system is also on the horizon, whereby a subscriber may have a number under any exchange code but can live anywhere in the country or even anywhere in the world. Call forwarding is another problem where a landline WRN may be forwarded to a business number, a cell-telephone number, a VoIP number, or a landline number out of area. Since geographic stratification for an RDD survey is usually based on the telephone area and exchange codes, extra screening efforts to identify the location of a number will be required with the increasing use of these systems. Also, increasing use of these systems will increase the rate of out-of-scope residential numbers or differences between frame and actual locations of in-scope sampling units, in turn increasing costs.

Sometimes, two telephone numbers are linked to the same landline telephone without the knowledge of a subscriber. The hidden number is called a *ghost number*, and calls made to this number will reach the subscriber just as will calls made to the real telephone number. The respondent's selection probability increases to an unknown extent and the statistician is left with no information to allow adjustment in the survey weighting.

A final set of problems has arisen from the rapid deployment and now widespread use of caller id and voice mail/answering machines. More and more people are using these technologies to screen their calls, and it is becoming more and more difficult to reach them or even resolve whether the telephone number is a WRN or not. In our recent experience with large studies in the United States, approximately 17% of all unresolved telephone numbers are due to answering machine systems and the lack of sufficiently detailed recorded messages or other markers that would allow us to otherwise classify the cases as WRNs, businesses, or other nonresidential. We also find that approximately 45% of our unresolved telephone numbers are due to repeated noncontact on all call attempts. Some unknown portion of this percentage is assumed to be due to people using caller-id to screen their calls. While caller-id and answering machines do not alter sampling units or their probabilities of selection, they do substantially increase survey costs by inflating the sample size necessary to achieve a specified number of completed interviews. They correspondingly depress response rates and thus heighten concern among statisticians about potential bias in RDD survey statistics.

## 3. Conduct of RDD surveys

The advent of CATI systems and automated dialers has made RDD surveys, and telephone surveys generally, a workhorse of social-science and market research. Scheduling, working, and manipulating the sample of telephone numbers to achieve a targeted number of completed interviews by stratum, on a timely basis, at low cost, and with a high response rate, is the task of the sample-management function.

### *3.1. Technology of data collection*

CATI is a term used to cover all computer-aided aspects of telephone interviewing. It covers both hardware requirements (including telephony systems) and software. Some CATI systems use a single integrated piece of software that controls the sample, the questionnaire, and the dialing; other systems combine elements from multiple vendors to take advantage of some specializations.

A good CATI system can connect dozens or hundreds of workstations in multiple locations and can offer interviewers and their supervisors the facility to work simultaneously sharing the same system. CATI was first introduced in the early 1970s and is now commonplace in market and social-science research.

A typical system offers customized tools for survey instrument development, call scheduling, display of survey items, recording of survey responses, monitoring and supervision of interviewers' work, keeping a record of calls for each case in the released sample, online data editing and processing, preparation and export of data sets, and other automatic record keeping. A typical system allows one quickly to develop a survey questionnaire in multiple languages, thus enabling the conduct of multilingual interviews. The system can automatically record call outcomes such as no answer, answering machine, disconnect, busy, or fax/modem; it also can dispatch only the connected calls to interviewers or schedule callbacks where required. During the interview itself, the system can automatically execute simple or complex skip patterns without interviewer intervention, and conduct subsampling, if required. It offers the facility to view the call history of a case and to add to it. Integrated CATI systems offer the facility to produce frequency tabulations, survey statistics, response rates, and productivity reports, which are useful for ongoing monitoring of progress and of key indicators of quality. They also admit data exports in a wide range of formats, thus enabling external analysis of the data and reporting of progress.

The two components of a CATI system which are of most relevance to RDD surveys are the sample management component and the dialer technology. Survey costs are largely driven by interviewer time, and in a RDD survey, it is not unusual for many dozens of dials to be required for each completed interview. Dialer technology provides efficiencies by reducing the amount of time the interviewer is involved in each dial, whereas the case management system (see Section 3.3) can provide efficiencies by reducing the number of dials needed to obtain the same result.

Automated dialers can assist the dialing process at both ends of the phone call. For example, at dial initiation, they can automatically deliver and dial the next number when an interviewer comes free, and at call outcome, they can detect outcome tones such as a busy line or a fax machine, and automatically apply the appropriate disposition to the case and file it away. Under a fully manual system, it takes approximately 40 seconds for an interviewer to dial and disposition an engaged or disconnected number; yet an automatic dialer can do this in less than a quarter of this time.

Dialers can be classified into two distinct groups, based on whether they are capable of predictive dialing or not. Predictive dialers are distinguished from nonpredictive (or preview) dialers in that they do not connect cases to an interviewer until after a connection is established. This means they can dial more numbers than there are interviewers available and deliver to the interviewers only those calls which need an interviewer (i.e., those that connect to an answering machine or a live person) while handling in the background a comparable number of dials which do not need interviewer attention at all. In this manner, interviewer time is eliminated from handling unproductive calls, increasing efficiency and lowering cost.

Different predictive dialers use different algorithms, but all rely essentially on predicting two probabilities: the probability of a number being answered and the probability of an interviewer being free to take a call. One can assess the optimal speed of the dialer by using a "supply and demand" analysis, where the demand side is due to the dialer dialing numbers and creating a demand for interviewer labor, whereas the supply side is created by the pool of interviewers who are free to accept a call. When the dialer runs faster than its optimal speed and demand exceeds supply, calls will be abandoned, that is, the dialer will need to hang up on a connection because it cannot find a free interviewer to whom to pass the connection. When the dialer runs slower than the optimal speed and supply exceeds demand, interviewers will sit idle and efficiencies will be lost. The fastest dialing will occur under the following situations:

- Relatively few connections occurring among the dials made immediately preceding the current dial
- Tolerance setting for the risk of abandonment set relatively high
- Many interviewers logged into the system
- Fairly short average connection length and overall relatively little variation in the connection lengths.

However, there are two distinct drawbacks to predictive dialing. One is the risk of abandoning calls, which can be a nuisance for respondents and can ultimately translate into respondent complaints or additional refusals. The other drawback is that since the dialer will not assign a case to an interviewer until after the connection is established, the first opportunity an interviewer will have to review call notes associated with the case is when they hear someone saying "hello." Particularly when trying to convert refusals, it is important that interviewers be informed on the details of previous refusals, but predictive dialers prevent this transfer of information and thus diminish response rates.

Hybrid dialing delivers most of the efficiencies of predictive dialing while eliminating these drawbacks. The dialer starts in predictive mode for virgin cases and continues subsequent callbacks in that mode, but then shifts to preview mode for all callbacks once contact has been established. Given that it is only upon contact that a household is identified and call notes written, the risk of abandoning a call to a known household is eliminated, and the interviewer can prepare properly for the call by reading the call notes left by the interviewer who handled the previous contact.

It is important to note that hybrid dialing mixes the two modes of dialing – predictive and preview – within a single CATI survey in real time. This is quite different from splitting the sample to start it in predictive mode (to clear out disconnects) and then switching the numbers once connected into a separate survey to run in preview mode, or vice versa. Splitting samples in such a manner is inherently less efficient because it means fewer interviewers logged in for the predictive algorithm to maximize its speed and substantial time required for sample management and call history reconstruction.

Hybrid also enhances the interviewer's task, which in turn translates to an improved respondent experience. With a traditional preview dialer, interviewers can spend at least one-third of their time listening to dials being placed or ringing out, which can lead to loss of attention and preparedness for those few crucial seconds when an interviewer first interacts with a potential respondent. The hybrid dialer enables interviewers to spend a much larger proportion of their time actively engaged in respondent interactions.

### *3.2. Sample preparation and release*

Once the sample is selected for an RDD survey, three preparatory steps are applied before releasing the sample to the telephone center for calling. The steps involve (a) subdividing the sample into replicates, (b) identifying the unproductive cases in the sample through prescreening of the telephone numbers for nonworking, nonresidential, and cellular telephone numbers, and (c) mailing advance letters to the households in the sample for which address information is available to increase response rates. These steps provide for good management and control of the sample and efficient dialing.

#### *3.2.1. Forming replicates*

In an RDD survey, the sample is usually not released and loaded into the CATI system as a single monolithic batch. Instead, the specified sample in each stratum is divided into a number of replicates that are formed randomly so that each replicate is a random subsample of the full sample. The replicates are then released to the telephone center as needed to spread the interviews for each stratum evenly across the duration of the interview period. Careful release of replicates allows for both efficient use of the interviewer pool and tight control over the number of completed interviews achieved. Recognizing the uncertainty in the level of survey response that will be achieved in an RDD survey, the statistician will often select a larger sample of telephone numbers than is expected to be consumed in the survey. Release of replicates is terminated when the difference between the (real-time) projected number of completed interviews and the target sample size determined at the planning stage is deemed to be acceptable. Unused replicates are not considered released and only released replicates constitute the actual sample for purposes of weighting and the calculation of response rates.

#### *3.2.2. Prescreening replicates*

Because approximately 75% (This percentage, derived from Table 1, assumes an ample data collection period with the survey protocol configured to achieve a relatively high response rate. Different survey conditions could yield a somewhat larger percentage.) of all selected telephone numbers are nonworking, nonresidential, or unresolved, a large part of the interviewers' efforts may potentially go into simply identifying the status of these numbers. To reduce the size of the task and allow interviewers to focus more fully on household interviews, replicates are prescreened before they are loaded into the CATI system to identify as many unproductive telephone numbers as possible. Telephone numbers can be classified as not eligible for interviewing in three ways.

First, the replicates are prescreened for businesses. Typically, this means the replicates are matched against business directories, and any telephone number that matches a directory listing is classified as a business number and is made ineligible to be called in the survey. If the survey topic might be sensitive to the exclusion of households which share a business number, additional matching of the numbers classified as businesses might be made against residential listings and the ineligible status retained only if no match is found in this additional process.

Second, the remaining numbers are matched against residential directories and the matches are classified as eligible to be called in the survey. All remaining nonmatches (did not match the business directories and the residential directories) are run through a predictive dialer, which automatically detects nonworking numbers by unique signal tones issued by the telephony system, by extended periods of silence, or by continuous noise on the telephone line. The dialer also detects fax and modem numbers. Such numbers are classified as ineligible to be called in the survey.

Third, in the United States, a telephone number originally assigned as a landline number can now be ported to a cell telephone at the request of the subscriber. This means that even if the RDD sample is selected from landline telephone exchanges, it may, by the time of the RDD survey, contain a few cell-telephone numbers. This porting of numbers creates a legal problem that must be solved before dialing can commence in the survey. Except for emergency calls or calls made with the prior express consent of the person being called, the Telephone Consumer Protection Act of 1991 prohibits the use of automatic dialers in calling telephone numbers assigned to a cell telephones. (The act does not bar the manual dialing of cell-telephone numbers.) Because the typical RDD survey uses an automatic dialer, the replicates to be released are matched to a commercial database that contains all ported numbers in the nation. Matched numbers are made ineligible to be called in the RDD survey.

Finally, all telephone numbers within the prescreened replicates that are not designated as ineligible are loaded into the CATI system and are made ready for the launch of interviewing operations. In our recent experience, approximately 56–58% of the selected sample of telephone numbers are loaded to the CATI system and are eventually called. The remaining (42–44% of the numbers) are prescreened as ineligible. Clearly, prescreening saves interviewers a lot of work.

### *3.2.3. Sending the advance letter*

When time and resources permit, an advance letter is mailed to the subscriber(s) of all selected telephone numbers for whom a mailing address is available and not removed in the prescreening process. The addresses are obtained by matching the replicates to be released to commercially available databases that contain telephone numbers and corresponding names and addresses. (These databases are populated with information from credit histories and other sources.) The advance letter explains the purpose of the survey and its importance, and identifies the sponsor and other pertinent facts about the task of completing the survey. The letters are usually mailed a fixed number of days before the expected release of a replicate, timed to arrive two to five days before the first dial is anticipated. The choice of mailing class (e.g., express, first class, or bulk) will depend on budget and available lead time. At the time of writing, an advance letter can be sent to about one-third of the cases in an overall sample in the United States. Approximately 50% of the cases loaded to the CATI system are sent an advance letter.

A well written and presented advance letter increases the rate of cooperation in a RDD survey. De Leeuw et al. (2007) in a meta-analysis found that cooperation was on average 11 percentage points higher among cases that receive an advance letter than among cases that do not receive a letter. The size of increase for any one study will vary depending on a range of factors including topic, letter, and timing, and because there is self-selection involved in which cases present themselves with a complete address for mailing the advance letter, we cannot conclude with certainty that the higher rate is "caused" entirely by the letter.

### *3.3. Case management*

Once all sample preparation steps have been completed, the sample is released to CATI for actual interviewing operations. Usually, the sample is released by replicate or batches of replicates and the release may occur on a daily basis or according to some other schedule. The size of the release of virgin replicates is coordinated with the number of interviewer hours available for interviewing, the distribution of those hours by shift and time zone, and in consideration of the number of pending cases already scheduled for an additional call attempt. The number of virgin cases to be released is determined by taking the difference between the total number of calls that can be handled by the telephone center in a given time period and the number of pending cases already scheduled for a call in that period. If staffing exceeds the amount of work currently in the system, a release of virgin replicates is made. The size of the release may vary by stratum, with relatively more virgin replicates released in any strata that are lagging in achieving their target numbers of completed interviews.

Once replicates are released, the next tasks are to schedule the individual cases to be called, to track and report on their status, to re-schedule them for additional calls, as needed, and generally to bring as many cases to a completed status as possible.

#### *3.3.1. Tracking of case status*

To effectively manage the sample throughout the data collection period, it is essential to be able to track the status of cases in real time, each and every day. Such management requires four distinct types of codes assigned at the case level, for every case in the released sample, and the survey manager must have the capability to tabulate frequencies and cross-tabulations of these codes at any moment. The codes are updated with each call attempt. The four types of codes are the Life Cycle Stage (LCS) code, the call-outcome code, the finalization code, and the disposition code.

The *LCS code*, presented in Table 2, describes the overall status of a case in the CATI system. It describes the key stages that a case goes through as it progresses through the telephone interviewing system. It can be used to determine the scheduling and intensity of future call attempts. Before any live contact is established with a case, there is a good chance that no household exists at the end of that line, so the objective is to be able to clear that telephone number out of the system as fast as possible. Once live contact is established, it is usually possible to classify the case as a WRN or as

Table 2
Life cycle stage codes

| Code | Label |
|---|---|
| 0 | Virgin (fresh) cases |
| 1 | No contact |
| 2 | No live contact but possible household |
| 3 | Live contact and likely household |
| 4 | Screened household |

a nonresidential number. The WRNs will likely be worthy of additional call attempts, and once a household has completed the screening interview and is determined to be eligible, it becomes the subject of still more call attempts, with the option to restrict these cases so that they are handled only by interviewers with advanced refusal conversion skills. The marginal cost of obtaining a completed interview from an already-screened household is generally relatively low. Thus, cases in LCS 4 receive the highest and most focused level of effort.

The LCS code follows a ratchet system that can move forward but not backward. For example, a case that was busy on the last call and on all previous calls will be in LCS 1. However, a case that was busy on the last call but was previously a refusal at the interview level will be in LCS 4, which captures the fact that the case has been identified as an eligible household.

The *call-outcome code* describes the state of a case based on the outcome of the last call attempt. Table 3 gives some examples of possible call-outcome codes. The combination of the LCS and call-outcome codes suggest what rule should be used to determine the schedule for the next call attempt, if any. If the call outcome is a busy signal, it is evident that someone is at home, and it is probably worthwhile to callback within the hour or half-hour. On the other hand, if the call outcome is a refusal, it is probably worthwhile to delay the next callback for several days, assuming the data collection period is long enough, to allow for a cooling-off period.

The *finalization code* is a simple indicator of whether a case has been finalized or not. Completed interviews, ineligible households, and nonworking or nonresidential telephone numbers are considered finalized, meaning no additional callbacks are planned for them. Cases not finalized require additional callbacks. In an RDD survey, the statistician may develop and use alternative rules to designate when a case is finalized. For example, a case may become final after

- a specified total number of call attempts
- a specified number of call attempts within each of several day-parts or shifts (weekday evenings, weekend days, and so forth)
- a specified number of call attempts after reaching LCS 4 status.

The *disposition code* for a case summarizes the current evidence with respect to its resolution as a WRN or not, its screening interview, and (if eligible) its main interview. The disposition code is a synthesis of the current LCS code, the sequence of call-outcome codes for all call attempts to date, and the current finalization code. We defer further discussion of disposition codes until Section 3.4.

### *3.3.2. Shift and resource attributes*

Call scheduling is driven by a number of factors, such as shift and resource attributes, time zone, and current LCS and call-outcome codes. There may be several shift types used in the RDD survey, including weekday days, evenings, and nights, and weekend days, evenings, and nights. A shift attribute is assigned to the next call for each case in the active sample.

There may be several resource types employed in the RDD survey, including regular interviewers, interviewers with refusal conversion skills, special language (e.g., Spanish) interviewers, and special language refusal converters. A resource attribute is assigned to the next call for each case in the active sample.

Table 3
Illustrative call-outcome codes

| Code | Label | Usage Notes |
|---|---|---|
| 01 | Engaged/busy | Non-autodisposition engaged/busy |
| 02 | No reply [no answer] | Non-autodisposition ring no answer/no reply |
| 03 | New phase | Virgin sample |
| 04 | Refusal HUDI | Used when a respondent hangs up during the introduction (HUDI). Respondent does not speak to interviewer at all. |
| 05 | Refusal (gatekeeper) | Used when an adult, noneligible respondent refuses. Only available if specific respondent has already been selected. |
| 06 | Refusal (soft) | Adult refuses to participate in the study after full introduction. |
| 07 | Refusal (hostile) | Used when respondents use profanity towards interviewer in a threatening manner or threatens any legal or governmental action. Reviewed by supervisors before being finalized. |
| 08 | Made soft appointment | Used when a household member gives callback time for a respondent. |
| 09 | Made hard appointment | Used when respondent gives you a specific callback time for them. |
| 10 | Unable to proceed: general call back | Used when no specific appointment information was given but recontact is viable. |
| 11 | Unable to proceed: supervisor review requested | Used when Supervisor review is required. |
| 12 | Privacy manager (known housing unit) | Used if unable to successfully bypass a privacy manager machine - known housing unit. |
| 13 | Privacy manager (unknown if housing unit) | Used if unable to successfully bypass a privacy manager machine - unknown if housing unit. |
| 16 | Answering machine: message left (known housing unit) | Used when a call machine is reached and a message is left: known housing unit. |
| 17 | Answering machine: message left (unknown if housing unit) | Used when a call machine is reached and a message is left - unknown if housing unit. |
| 18 | Business/government | Used when a business or home business line (not used for personal calls) or government office is reached. |
| 19 | Dorm/prison/hostel | Used when a dorm is reached. |
| 20 | Cell phone/mobile/GPS phone | Used when number dialed belongs to a cell phone or other mobile device. |
| 21 | Call forwarding | Used when number dialed is permanently forwarded to a different number. |
| 22 | Fast busy | Used when a fast busy signal is received. |
| 23 | Disconnect/temporarily disconnected | Used when a number is permanently disconnected, temporarily disconnected, or has been changed. |
| 25 | Fax/modem/data line | Used when a fax/modem signal is received. |
| 32 | HH ineligible (no eligible person in HH) | Screener complete, no household members eligible |

The next call for a case in the active sample is scheduled in light of its current LCS and call-outcome codes, the number of times the case was already called in the various shift types, its time zone, and the availability of an appropriate resource in the interviewer pool.

### 3.3.3. *Three phases of the data-collection period*

Interviewing operations proceed through three distinct phases during the data-collection period. During the initial start-up phase, the active sample will be dominated by virgin replicates. At this time, few refusal converters are needed. During the longer middle phase, the active sample is characterized by a mixture of virgin and pending cases. The final phase, also called the close-down period, begins at the point in time when the last replicate is released. Because no virgin replicates are introduced during this period, and as noncontact cases are retired, the active sample will come to be dominated more and more by difficult and eligible cases. At this time, a larger number of specialized interviewers are needed to deal with partially completed interviews, refusals and hidden refusals, appointments, and cases that require a specialized language.

### 3.3.4. *Staffing and staff scheduling*

Once the middle phase of data collection is reached, analysis of the most productive times to call and the dynamics of the call scheduling rules (particularly if shift types are used) will produce a distinct pattern of dials that must be performed at different times of the day and week. For example, many companies concentrate RDD dialing into weekday evenings, when households are more likely to be at home, leaving the day only lightly staffed.

This pattern is further complicated by the number of time zones that a study is spanning from a single site. For example, if two-thirds of the sample is in the Eastern time zone and the remainder is in the Central time zone, and if the sample size requires 10 booths to be active at 9:00 am Eastern time, then by 10:00 am Eastern time another 5 will need to be added for a total of 15. If a sharp distinction is drawn between midweek evening and midweek days with volumes increasing four-fold, the booth requirements at 5:00 pm will jump to 45 once the Eastern evening dialing starts, and up to 60 at 6:00 pm when the Central evening dialing starts, dropping to only 20 at 9:00 pm Eastern time when the Eastern dialing closes for the day. If only five interviewers were dialing at 9:00 am, a proportion of the Eastern sample would not be dialed in that first hour and would back up for later dialing at a less optimal time of day or even to the next day. Conversely, if 20 interviewers started dialing at 9:00 am, then the day work would be exhausted by mid-afternoon, and either the interviewers would sit idle, or cases requiring a night time call would be dialed far too early with a much lower chance of success. While the patterns of calling by time zone cited here are merely illustrative, call center managers need to understand their own survey-specific sample sizes and interviewer capacity to schedule and manage the staff.

## 3.4. *Case disposition*

For an RDD survey, just as for most surveys, the statistician must provide an explicit definition of what constitutes a completed interview. The definition may hinge on such factors as whether key items were completed, whether a majority of items were completed, or whether a certain range of sections of the questionnaire were completed before break-off, if any.

Given this definition, telephone numbers in the released sample should move through the CATI system until they finalize as resolved-nonhousehold, ineligible-household, or eligible-household-completed-interview. The cases that are not finalized directly are eventually finalized as unresolved, unscreened-households, or eligible-household-incomplete-interview.

After going through the prescreening, calling, and interviewing process, all cases in an RDD survey must end up receiving final disposition codes based on their prescreening statuses, life cycle, and call-outcome codes. The final disposition codes must be defined at a sufficient level of detail so that the cases can be treated appropriately in estimation, analysis, and reporting of response rates.

The American Association for Public Opinion Research has developed standard definitions for disposition categories for RDD surveys (AAPOR, 2006). At the highest level, the AAPOR codes are:

(1) Interview includes fully complete interviews or partial interviews with all necessary questions answered.

(2) Eligible, noninterview includes cases that are resolved as WRNs and are screened as eligible for the survey but have not completed the interview.

(3) Unknown eligibility, noninterview includes cases that are unresolved for WRN status or are resolved as WRN but the screener interview to determine eligibility was incomplete.

(4) Not eligible includes cases that are not WRNs and cases that are WRNs but are screened as ineligible.

For a listing of the detailed AAPOR codes, see Table 1 in the document www.aapor.org/uploads/standarddefs_4.pdf.

While the detailed categories of AAPOR can be very helpful to data-collection managers in diagnosing problems and opportunities and to survey statisticians in planning sample sizes and data-collection operations for future surveys, they provide more detail than is necessary to conduct estimation and analysis for the current survey. For the latter purposes, we find that a simpler set of final disposition categories is sufficient and fully acceptable. We present our set of final disposition categories in Table 4 along with a cross-walk that demonstrates the link between the AAPOR detailed categories and our final disposition categories. The next subsection shows how these final categories support the calculation of response rates for an RDD survey, and Section 4 demonstrates how the final categories support the survey estimation procedure.

To give an illustration of the distribution of an RDD sample by final disposition categories, we have analyzed recent RDD surveys conducted at NORC. Figure 1 presents the recent trend in case dispositions found in our surveys. WRNs comprise categories U1, J, ER, and C. Categories V and UN are unresolved, whereas all other categories are resolved. Notice the downward trend in the proportion of the sample classified as resolved WRNs. This trend is due to the evolving telephony infrastructure in the United States.

Table 4
Cross-walk between final disposition categories and AAPOR categories

| Final Categories | Meaning | AAPOR Categories |
|---|---|---|
| V | Virgin, cases released to telephone center but never dialed | 3.11 |
| UN | Unresolved telephone number | 3.10 (except 3.11) |
| D | Nonworking, out-of-scope | 4.20 + 4.30 + 4.40 |
| NR | Nonresidential, out-of-scope | 4.50 |
| U1 | Known household, screening incomplete | 3.20 |
| J | Screened household, not eligible | 4.70 |
| ER | Eligible respondent, incomplete interview, or refusal | 2.0 |
| C | Completed interviews | 1.0 |

*Note*: AAPOR categories 3.90, 4.10, and 4.80 are not used. 3.90 is ambiguous and should be classified to either UN or U1, as the case may be; 4.10 should not exist in a properly managed survey; 4.80 does not apply in the case of strict probability sampling discussed here.

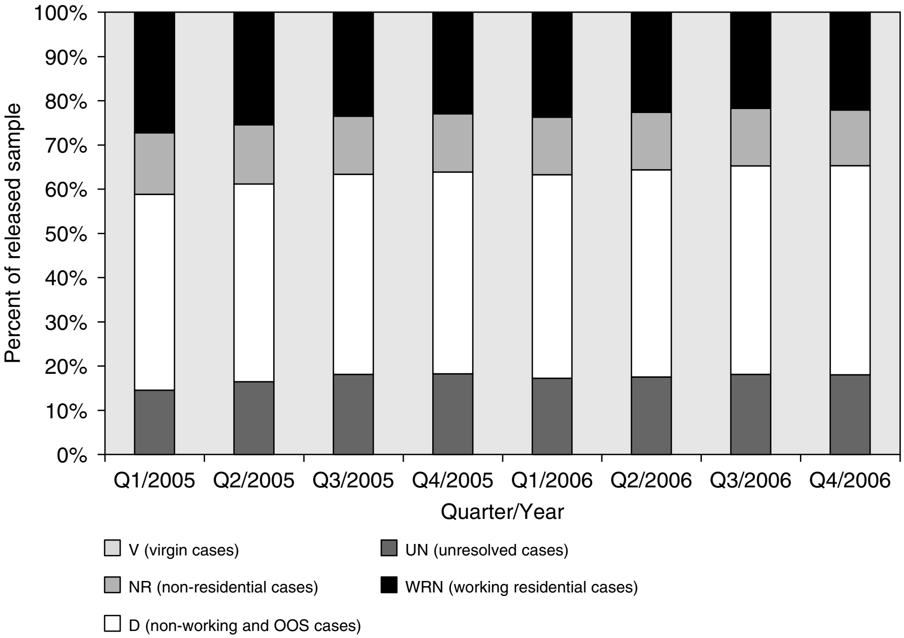


Fig. 1. Recent trends in the classification of cases in RDD surveys

### 3.5. *Measures of response rates*

It is important to report the response rate for an RDD survey, just as it is for any other survey. This rate stands as an indicator of the risk of nonresponse bias that may be present in survey estimators. Features of RDD surveys make the calculation of the response rate less than completely straightforward. To illustrate the principles, we take the situation in which a brief screening interview is administered to households to ascertain eligibility for the survey, followed by a main interview administered only to eligible households. Given this situation, what is the appropriate response rate?

The *response rate* is a summary measure used to designate the ratio of the number of completed interviews to the number of eligible units in the sample. It is a measure of the result of all efforts, properly carried out, to execute the RDD survey. Since this definition is supported by the Council of American Survey Research Organizations (CASRO), we often call it the CASRO response rate. Assumptions or estimation is usually required to work out the denominator of the response rate.

In general, the response rate is defined by

$$r = \frac{C}{C + ER + \lambda_1 U1 + \lambda_2 \phi UN} = \frac{\text{completed interviews}}{\text{eligible cases in the released sample}}, \tag{6}$$

where $\lambda_1$ is the unknown proportion of unscreened households that are in fact eligible for the survey, $\phi$ is the unknown proportion of unresolved telephone numbers that are in fact WRNs, and $\lambda_2$ is the unknown proportion of unresolved WRNs that are in fact eligible for the survey. Equation (6) also agrees with response rate 3 defined in AAPOR (2006). Because the rates in the denominator of the response rate are unknown, we estimate them from the sample itself by

$$\hat{\lambda}_1 = \frac{C + ER}{C + ER + J} = \frac{\text{eligible cases}}{\text{eligible and ineligible cases}}, \tag{7}$$

$$\hat{\phi} = \frac{C + ER + J + U1}{C + ER + J + U1 + D + NR} = \frac{\text{WRNs}}{\text{resolved cases}}, \tag{8}$$

and

$$\hat{\lambda}_2 = \hat{\lambda}_1. \tag{9}$$

For simplicity, we are using a system of notation wherein the same symbol is used to represent both the final disposition category, the set of survey cases classified in the category, and the cardinality of the set. For example, C designates both the category of completed interviews, the set of cases with a completed interview, and the number of cases that achieved a completed interview.

Given assumptions (6)–(8), the estimated response rate becomes

$$\hat{r} = \frac{C}{C + ER + \hat{\lambda}_1 U1 + \hat{\lambda}_2 \hat{\phi} UN} = c_R c_S c_I, \tag{10}$$

where

$$\begin{aligned} c_R &= \frac{C + ER + J + U1 + D + NR}{C + ER + J + U1 + D + NR + UN} \\ &= \frac{\text{resolved telephone numbers}}{\text{total telephone numbers in released sample}} \end{aligned} \tag{11}$$

is the *resolution completion rate*;

$$c_S = \frac{C + ER + J}{C + ER + J + U1} = \frac{\text{screening interviews}}{\text{WRNs}} \tag{12}$$

is the *screener completion rate*; and

$$c_I = \frac{C}{C + ER} = \frac{\text{completed interviews}}{\text{eligible households}} \tag{13}$$

is the *interview completion rate*. The completion rates are useful not only for computing the response rate but also for planning, monitoring, and managing the RDD survey.

Massey (1995) and Ezzati-Rice et al. (2000) propose an alternative to assumptions (6)–(8) that takes account of both nonresponse and undercoverage, if any, in the number of identified eligible households in the survey. Some authors prefer to quote weighted response rates. To implement this idea, one could let $C$, ER, and so on be the sum of the base weights of the members of the corresponding set, instead of simply the cardinality of the set. The use of the weights could be important for an RDD survey that employs differential sampling rates from stratum to stratum.

## 4. Analysis of RDD surveys

In this section, we discuss the procedures used to develop *weights* for the survey respondents. Properly calculated weights are needed to provide essentially unbiased estimators of population parameters. Given a large-scale RDD survey, an estimator of the population total, $Y$, is of the general form

$$\hat{Y} = \sum_{i \in C} W_i Y_i, \tag{14}$$

where $C$ is the set of completed interviews, $Y_i$ is the characteristic of interest for the $i$th completed interview, and $W_i$ is the survey weight for the $i$th completed interview. Estimators of proportions, ratios, regression coefficients, and the like are calculated as functions of estimated totals.

Weighting begins with the construction of the probabilities of selection for the telephone numbers in the released sample. The *base weights*, or the reciprocals of the probabilities of selection, implement the Horvitz and Thompson (1952) estimator of the population total. Several adjustments to the weights are required, each being designed to compensate for missing data at the several steps in the survey response process. The steps usually include nonresolution of telephone numbers as to their residential or nonresidential status; nonresponse of households to the screening interview, if any, designed to determine the persons eligible for the survey; and nonresponse of eligible respondents to the main interview. In addition, weights should be adjusted to correct for any multiple probabilities of selection and to compensate for any households missing from the sampling frame because they do not have landline telephones. Finally, weights are usually benchmarked to external population control totals. In what follows, we define all these steps in weighting. Some steps may not apply to all RDD surveys.

### 4.1. Base weights

The list-assisted RDD survey design is essentially a simple random sample without replacement of telephone numbers within stratum, with independent sampling from one stratum to the next. Let $N_h$ be the size of the population of all landline telephone numbers in stratum $h$ and let $n_h$ be the size of the sample selected and released. Then, the base weight for the $k$th telephone number in the set of released telephone numbers, $A$, is defined by

$$W_{1k} = 1/\pi_{1k} = N_h/n_h, \quad \text{if } k \in A, \tag{15}$$

where $\pi_{1k}$ is the probability of selection given the sampling design. The base weight is a constant for all released telephone numbers in a stratum.

## *4.2. Adjustment for nonresolution of telephone numbers*

As we have seen, the RDD frame is highly inefficient with more than 70% of the telephone numbers being out-of-scope, either nonworking or nonresidential. The first step in the survey response process is to identify the WRN status of all released telephone numbers. Even after repeated call attempts, some telephone numbers will inevitably remain unresolved, meaning that there was not enough evidence collected to classify them as residential, nonresidential, or nonworking. To compensate for the missing classifications, a nonresponse adjustment is conducted within cells. The adjustment is applied by forming adjustment cells within each stratum and assuming that the WRN rate within a cell is the same for both the resolved and unresolved cases. The weights of the unresolved telephone numbers are distributed to the weights of the resolved cases in the same adjustment cell. For the $k$th resolved telephone number within the $\ell$th adjustment cell, the nonresolution adjusted weight is defined by

$$W_{2k} = W_{1k}/\pi_{2\ell}, \quad \text{if } k \in B \cap \ell, \tag{16}$$

where $B$ is the subset in $A$ of resolved telephone numbers, and

$$\pi_{2\ell} = \sum_{k \in B \cap \ell} W_{1k} \Bigg/ \sum_{k \in A \cap \ell} W_{1k} \tag{17}$$

is the weighted resolution completion rate for the $\ell$th adjustment cell.

## *4.3. Adjustment for nonresponse of households to the screening interview*

Once a WRN is identified, it is screened for eligibility. The target population of many surveys does not include all persons within the household. A given RDD survey may only target persons of a specific age, sex, education level, or health condition. The brief screening interview, conducted prior to the main interview, is intended to identify the members of the target population living in households. Despite repeated callbacks, the screener is sometimes left missing due to refusal or noncontact. To account for such nonresponse, the weights of the WRNs with completed screening interviews are adjusted to account for the nonresponding WRNs within cells. The screener nonresponse adjusted weight of the $k$th screener complete in the $m$th nonresponse adjustment cell is defined by

$$W_{3k} = W_{2k}/\pi_{3m}, \quad \text{if } k \in S \cap m, \tag{18}$$

where $S$ is the subset in $B_1$ consisting of screener completes, $B_1$ is the subset in $B$ consisting of the resolved WRNs, and

$$\pi_{3m} = \sum_{k \in S \cap m} W_{2k} \Bigg/ \sum_{k \in B_1 \cap m} W_{2k} \tag{19}$$

is the weighted screener completion rate for the $m$th adjustment cell.

### 4.4. Adjustment for selection of eligible respondents

In some surveys, once the eligible persons are identified within the household, one or more of them are selected for participation in the survey. Let $E$ be the subset in $S$ of households with eligible persons, $P$ be the set of eligible persons identified in completed screening interviews, and $P_1$ be the subset in $P$ of persons selected for the survey interview. Then, an adjustment for subsampling is defined by

$$W_{4i} = W_{3k}/\pi_{4i}, \quad \text{if } i \in k \text{ and } i \in P_1, \tag{20}$$

where $i$ designates the selected, eligible person in household $k$, $\pi_{4i}$ is the conditional probability of selecting the $i$th eligible person given that their household screening interview was completed, and $W_{3k}$ is the screener-nonresponse adjusted weight from the previous weighting step. For example, if one eligible person was selected at random from a household containing three eligible persons, then $\pi_{4i} = 1/3$. If the survey calls for interviewing all eligible persons within the household, then $\pi_{4i} = 1$ and $W_{4i} \equiv W_{3k}, i \in k$.

### 4.5. Adjustment for nonresponse of eligible persons to the main interview

Following screening and subsampling, the survey's main interview is administered to the selected eligible persons. At this juncture, further nonresponse is likely due to refusals and not-at-homes. An interview nonresponse adjustment is applied within cells to account for the eligible cases that fail to provide a completed interview. The interview nonresponse-adjusted weight for the $i$th eligible person with a completed interview in the $q$th adjustment cell is defined by

$$W_{5i} = W_{4i}/\pi_{5q}, \quad \text{if } i \in C \cap q, \tag{21}$$

where $C$ is the subset in $P_1$ of eligible persons who completed the main interview, and

$$\pi_{5q} = \sum_{i \in C \cap q} W_{4i} \Bigg/ \sum_{i \in P_1 \cap q} W_{4i} \tag{22}$$

is the interview completion rate within the $q$th cell.

Table 5 presents a summary of a typical response pattern in an RDD survey, showing how the units with various disposition codes are treated in different weighting steps.

Table 5
Summary of hierarchical response pattern and mix of units for a typical RDD survey

| Disposition Codes | Released Sample of Telephone Numbers | Resolution Status (Telephone Numbers) | WRNs (Telephone Numbers) | Screening Status (Telephone Numbers) | Eligibility Status (Households) | Subsampling Status (Eligible Persons) | Interview Status (Eligible Persons) |
|---|---|---|---|---|---|---|---|
| C: complete interview | $A$ = set of all telephone numbers in the released sample | $B$ = set of all resolved telephone numbers | $B_1$ = set of all resolved WRNs | $S$ = set of WRNs that responded to the screening interview | $E$ = set of households with one or more eligible persons | $P_1$ = set of selected, eligible persons | $C$ = set of respondents to the main interview |
| ER: incomplete interview | | | | | | | $P_1 \cap C^c$ = selected, eligible persons that did not complete the main interview |
| Eligible persons subsampled out | | | | | | $P \cap P_1^c$ = set of nonselected eligible persons | |
| J: screened household, no eligible person | | | | | $S \cap E^c$ = set of ineligible households | | |
| U1: known household, incomplete screening interview | | | | $B_1 \cap S^c$ = set of WRNs that did not complete the screening interview | | | |
| NR: nonresidential, out-of-scope | | | $B \cap B_1^c$ = set of resolved non-working or non-residential numbers | | | | |
| D: nonworking, out-of-scope | | | | | | | |
| I: answering machine (status unresolved) | | $B^c$ = set of all unresolved telephone numbers | | | | | |
| NC: noncontact (status unresolved) | | | | | | | |
| U2: possible household (status unresolved) | | | | | | | |

### 4.6. Forming adjustment cells

A different set of nonresponse adjustment cells may be formed and used at the different steps in the RDD response process. As would be typical of any sample survey, cells in an RDD survey are formed to achieve two ends: (i) units within a cell should be alike with respect to the survey characteristics of interest, and (ii) the survey response rates should vary from cell to cell. Cells are usually, though not necessarily always, nested within the sampling strata. The directory-listed status is often found to be a significant correlate of the nonresponse mechanism at each of the resolution, screening, and interviewing steps and is almost always included in the cell structure within stratum.

At the resolution step, only sampling frame variables are available for cell formation, including such variables as directory-listed status; broad geographic location including census region and state; and environmental variables obtained at the census tract level corresponding to the approximate mapping between the telephone exchange and the tract. The latter include variables related to the distribution of the population with respect to age, sex, race/ethnicity, housing tenure, and income.

At the screening step, both frame variables and variables collected at the resolution step are available for cell formation, and at the interviewing step, frame variables, resolution variables, and variables collected at the screening step are available for cell formation. Usually, the resolution step adds little information, and the cell structure for the screener nonresponse adjustment is forced to rely mainly on frame variables. Depending on the RDD survey, some screening interviews collect substantial new information and some do not. Sometimes the screening interview is minimized to include only essential variables so as not to risk additional nonresponse at this step. The cell structure for the interview nonresponse adjustment may be relatively more or less articulated depending on how many additional useful variables are obtained in the screening interview.

### 4.7. Adjustment for multiple probabilities of selection

Many households have two or more landlines that may be used for voice communications (excluding the lines used only for fax or computer communications), and the interview respondents in these households have multiple chances of selection into the survey. The increased probabilities of selection were not known at the time the base weights were formed and, therefore, an adjustment for these probabilities is now in order. Without an adjustment, the extant weights would provide an upward biased estimator of the population total. Although we have not emphasized this point until now, at this step in the weighting process, we have an essentially unbiased estimator of the total population of eligible person/WRN pairs. An eligible person with two landlines appears twice in this total. The real parameter of interest, however, is the total (and functions thereof) of the population of eligible persons. A weight adjustment is needed to convert the weighted estimator to be an estimator of the real parameter of interest.

The number of voice landlines is collected during the interview, and the appropriate adjusted weight is given by

$$W_{6i} = W_{5i}/\min(t_i, t_o), \tag{23}$$

where $t_i$ is the number of telephone lines in the household of the $i$th eligible respondent. The value $t_o$ is used to cap the number of telephone lines used in the weight adjustment, both to control variability and to guard against reporting bias. For example, some surveys take $t_o = 3$.

It is possible that a household with multiple landlines is selected more than once in the sample. With time and expense, one can undertake efforts to identify the lines. To identify the landlines in the sample that are linked to the same household, all landlines of the responding household can be collected and the sample file of responding landlines can be checked to establish the link. However, to avoid the response burden and the costs of collecting and processing the information, and considering the negligible chance of selecting the household more than once in the sample, it is usually assumed in the weighting process that only one landline has been selected per household.

### *4.8. Adjustment for noncoverage of nonlandline households*

Since the frame of a RDD survey only includes landline telephone numbers, the eligible persons living in households without landline service (including both nontelephone households and households with only cell-telephone service) are not covered by traditional RDD surveys. To compensate for this type of undercoverage, a certain poststratification adjustment can be used (Keeter, 1995). The basic idea is to use the households with an interruption in landline telephone service of one week or more during the past year to represent not only themselves but also the households without a landline telephone. Keeter (1995), Brick et al. (1996), and Srinath et al. (2002) showed that the socioeconomic characteristics of persons who live in households with interruptions of one week or more in landline telephone service within the past 12 months are similar to those who live in nontelephone households. This finding resonates with common sense because if a survey is conducted at a point in time when the household service is interrupted, then the household is necessarily considered as a part of the population of nontelephone households. Therefore, interviewed persons living in households with a recent interruption in landline telephone service can be used to represent persons living in nontelephone households.

Given this method, two cells are formed within a stratum by the telephone interruption status of the household as follows: 1) interruption of more than one week during the past 12 months; and 2) no interruption of more than one week during the past 12 months. Let $T_{h1}$ denote the total number of eligible persons in the population either without a landline telephone or with a landline telephone with an interruption in service, and let $T_{h2}$ denote the total number of eligible persons in the population with a landline telephone without an interruption in service, all in stratum $h$. Then, the adjusted weight for the $i$th person in $C$ is defined by

$$W_{7i} = \left( T_{hg} \Bigg/ \sum_{i \in C \cap U_{hg}} W_{6i} \right) W_{6i}, \quad \text{if } i \in C \cap U_{hg}, \tag{24}$$

where $g$ indexes the cell within stratum $h$, and $U_{hg}$ denotes the population of eligible persons in households in cell $(h, g)$. The population totals $T_{h1}$ and $T_{h2}$ are typically unknown and must be estimated from a census or a reliable reference survey.

### 4.9. Calibration to population control totals

The aim of the previous step is to reduce bias in the survey estimator due to the noncoverage of the population without landline telephone service. The RDD surveys may also be subject to differential coverage of the population by race/ethnicity and other factors. Like almost any census or survey, some categories of persons are underreported at a higher rate than others. The penultimate step in RDD surveys is to correct for the differential undercoverage by calibrating the weights to independent population control totals. A poststratification (Holt and Smith, 1979) or raking-ratio type calibration (Brackstone and Rao, 1979) is commonly applied.

The requisite population control totals are obtained from a census or a reference survey and are typically available for socioeconomic characteristics such as age, sex, race/ethnicity, income or poverty status, education attainment, and telephone interruption status. The population control totals may be obtained for the population as a whole or within each stratum.

To illustrate, we give the adjustment corresponding to a raking-ratio calibration. Introduce the following additional notation:

$a$ = iteration within the raking procedure
$b$ = dimension (or classification variable) of the raking structure within the iteration
$c$ = category (or value or the classification variable) within the dimension within the iteration
$L^b$ = number of categories within the $b$th dimension
$T_c^b$ = population control total for the $c$th category within the $b$th dimension of the raking structure
$\varphi_{ci}^b$ = 1, if the $i$th eligible person is in the $c$th category of the $b$th dimension of the raking structure
= 0, otherwise

Then, the adjusted weights at the $b$th dimension of the $a$th iteration are given by

$$W_{8i}^{a,b} = \left( \sum_{c=1}^{L^b} \varphi_{ci}^b \frac{T_c^b}{\sum\limits_{i \in C} \varphi_{ci}^b W_{8i}^{a,b-1}} \right) W_{8i}^{a,b-1}, \tag{25}$$

for eligible persons with a completed interview $i \in C$. The raking procedure cycles through each of the dimensions of one iteration before moving forward to the next iteration. The entry weights for the first dimension of an iteration are the exit weights following the last dimension of the previous iteration. The entire process opens with the weights from the prior step in weighting, $\{W_{7i}\}$ for $i \in C$. The process iterates until a user-specified convergence criteria is achieved.

For example, if population control totals are available for five age groups, two sex groups and four race/thnicity groups, then we have the raking structure

| Dimension ($b$) | Classification Variable | Number of Categories ($L^b$) |
|---|---|---|
| 1 | Age | 5 |
| 2 | Sex | 2 |
| 3 | Race/ethnicity | 4 |

### *4.10. Trimming of extreme weights*

Since the weights in an RDD survey are derived through the foregoing series of adjustments, a few weights in some strata may end up being very large in comparison to other weights in the stratum. To avoid any undue influence of large weights on survey estimates and to control the sampling variance of the estimator, the extreme weights may be trimmed by using a suitable truncation or Winsorization procedure (see Fuller, 1991; Kish, 1992; Potter, 1990). After trimming, the calibration adjustment is reapplied to ensure the consistency of the resulting weights with the population control totals. In some cases, the process may have to be iterated a few times to achieve final weights that are at once void of extreme values and also consistent with the population control totals. Let $\{W_{9i}\}$ or simply $\{W_i\}$ denote the resulting, final weights.

### *4.11. Estimation and variance estimation*

For a RDD survey, the survey weights are used to produce survey statistics and variance estimates in the usual way. To see this, expand the notation to let $\mathbf{Y_{hij}} = (Y_{1hij}, \ldots Y_{phij})'$ be a $p$-variate characteristic of interest reported by the $j$th eligible person selected in the $i$th household in the $h$th stratum, and $W_{hij}$ be the corresponding final weight. Suppose the goal is to estimate a parameter $\theta = g(\mathbf{Y})$ of the eligible population, where $\mathbf{Y}$ is the vector of population totals and $g(\cdot)$ is a well-behaved differentiable function. Then, the estimator of the parameter of interest is $\hat{\theta} = g(\hat{\mathbf{Y}})$, where

$$\hat{Y}_r = \sum_{(h,i,j)\in C} W_{hij} Y_{rhij} \tag{26}$$

is a typical element of $\hat{\mathbf{Y}}$, $r = 1, \ldots, p$, and $C$ is the set of eligible respondents. The usual Taylor series estimator of the variance is

$$v(\hat{\theta}) = \sum_h \frac{n_h}{n_h - 1} \sum_i \left( \sum_j W_{hij} \hat{V}_{hij} - \frac{1}{n_h} \sum_{i'} \sum_{j'} W_{hi'j'} \hat{V}_{hi'j'} \right)^2 \tag{27}$$

$$\hat{V}_{hij} = \sum_{r=1}^{p} \frac{\partial g(\hat{\mathbf{Y}})}{\partial y_r} Y_{rhij}, \tag{28}$$

where $\sum_h$ denotes a sum over strata, $\sum_i$ is a sum over the respondents' households within stratum, $n_h$ is the number of such households, and $\sum_j$ is a sum over respondents within household. For a superior estimator of variance that takes into account the calibration done in the 9th step in weighting, see Sections 6.12 and 6.13 of Wolter (2007). Notice that the households act as the primary sampling units in an RDD survey.

Alternatively, a replication-type estimator of variance may be used, such as the jackknife estimator

$$v_J(\hat{\theta}) = \sum_h \frac{n_h - 1}{n_h} \sum_i \left( \hat{\theta}_{(hi)} - \hat{\theta}_{(h+)} \right)^2, \tag{29}$$

where

$$\hat{\theta}_{(hi)} = g(\hat{\mathbf{Y}}_{\mathbf{(hi)}}) \tag{30}$$

is the estimator derived from the sample after omitting the $(h, i)$th household,

$$\hat{\theta}_{(h+)} = \frac{1}{n_h} \sum_i \hat{\theta}_{(hi)}, \tag{31}$$

$$\hat{Y}_{r(hi)} = \sum_{h'} \sum_{i'} \sum_{j'} W_{(hi)h'i'j'} Y_{rh'i'j'} \tag{32}$$

is a typical element of $\hat{\mathbf{Y}}_{(\mathbf{hi})}$ for $r = 1, \ldots, p$, and the replicate weights are defined by

$$\begin{aligned} W_{(hi)h'i'j'} &= W_{h'i'j'}, && \text{if } h' \neq h \\ &= W_{h'i'j'} \frac{n_h}{n_h - 1}, && \text{if } h' = h \text{ and } i' \neq i \\ &= 0, && \text{if } h' = h \text{ and } i' = i. \end{aligned} \tag{33}$$

For large RDD surveys, it will be prohibitive to compute replicate weights for the drop-out-one-completed-interview version of the jackknife. To better manage survey costs, one may use the survey replicates defined in Section 3 and compute the drop-one-replicate (or random group) version of the jackknife (see Wolter, 2007, Chapter 4).

One may use the estimated variances and covariances together with ordinary normal theory to make inferences regarding relationships and parameters of interest.

### *4.12. Difference in frame and actual location*

In some cases, a telephone number is selected from one stratum, but the interview reveals that the respondent (or the respondent's housing unit) is actually in a different stratum. Actual location may be collected in the interview and used to define the estimation domains for analysis. (Actual location could also be used earlier in the implementation of the last several weight adjustments.)

Since the sample selection probability may vary substantially from one stratum to another, the movement of one or more respondents to a new estimation domain may add variation in the sampling weights within the domain. To protect against any such extra variation in weights, the weight of the reclassified respondent may be truncated if it is very large compared with all other weights in the domain. A multiple of the average weight in the domain is sometimes used as a cap for the weights.

Let $\delta_i^d$ be an indicator variable for the $d$th estimation domain based on actual location collected in the interview. Then the estimator of the domain total is defined by

$$\hat{Y}^d = \sum_{i \in C} W_i \delta_i^d Y_i. \tag{34}$$

Variance estimation occurs as defined in Section 4.10, with the variable $Y_i^d = \delta_i^d Y_i$ replacing the original variable$Y_i$.

# Introduction to Part 2

Paul Biemer

After the data are collected, a number of processing steps must be performed to convert the survey data from their raw, unedited state to a verified, corrected state ready for analysis, and/or for dissemination to the users. If the data are collected by paper-and-pencil interviewing (PAPI) methods, they must be converted into a computer-readable form. Data collected by computer-aided interviewing (CAI) do not require this step, but may require additional cleaning steps to remove data remnants left after an erroneous branch. Responses to open-ended questions may need to be classified into categories using a coding scheme so that these responses can be tabulated. Additional operations may be performed on the data to reduce survey error and missing data.

For example, the data may be "cleaned" by eliminating inconsistencies and addressing unlikely or unusual responses (e.g., outliers). Survey weights may be computed to account for unequal selection probabilities. These weights may be further refined by a series of postsurvey adjustments that are intended to reduce coverage error bias, nonresponse bias, and sampling variance. Some survey variables (for example, household income) may have numerous missing values, and plausible values may be imputed for them. After these steps are completed, the data contents file should be well-documented. Data masking and de-identification techniques may also be conducted on the file to protect the confidentiality of the respondents. The next section provides a brief overview of these data processing activities.

## 1. Overview of data processing steps

The data processing steps vary depending on the mode of the data collection for the survey and the technology available to assist in the data processing. The steps involved for processing PAPI questionnaires, shown in Fig. 1, are discussed initially. The steps for CAI are essentially the same except for the data entry step.

Prior to keying, paper questionnaires must undergo a *scan editing* process that involves several steps. First, as the survey organization receives the questionnaires, their identification numbers are entered into the receipt control system and the questionnaires are inspected for obvious problems, such as blank pages or missing data for key items that must be completed for questionnaires to be usable. Questionnaires determined to be incomplete may be sent back to the field for completion. In mail surveys, incomplete questionnaires might be routed to a telephone follow-up process for completion. In cases where there is no follow-up of nonresponse, the questionnaires may be passed on

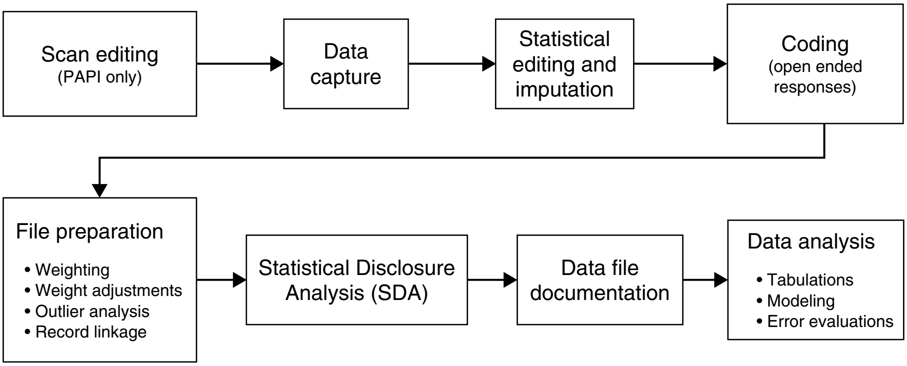


Fig. 1. Processing steps for a typical survey.

to the next data processing step (i.e., data capture or keying). Ultimately, questionnaires that are not minimally complete are coded as nonresponding units. As part of the scan editing process, questionnaires may be grouped into small batches called work units to facilitate the subsequent processing steps.

For the data capture step, paper questionnaires are *digitized* (i.e., converted into a computer-readable form). Data can be entered manually using keying equipment or automatically using scanning or optical character recognition devices. For the latter, messy questionnaires may have to be copied onto new, clean forms so that the scanner can read them properly. Keying usually involves some form of quality control verification. For example, each questionnaire may be keyed independently by two different keyers. Any discrepancies between the first and second keyed entries are then rectified by the second keyer. Alternatively, acceptance sampling methods (typically, a single sampling plan) may be applied to each work unit. Here, only a sample of questionnaires within each work unit is rekeyed. If the number of discrepancies between the two keyings exceeds some threshold value, the entire work unit is rekeyed. As a result of these verification methods, the error rate for keying is usually quite low for closed-ended responses: less than 0.5%. However, for verbal responses such as names and addresses, the error rates are substantially higher: 5% or more (Biemer and Lyberg, 2003).

Note that for CAI questionnaires, interviewers or respondents perform this data capture step as they enter their data directly into the computer. Typically there is no quality control operation to identify the keying errors during this step. However, some evidence suggests (see, for example, Dielman and Couper, 1995; Lepkowski et al., 1998) that keying errors for CAI are quite small and inconsequential.

### *1.1. Editing and imputation*

Editing is a process for verifying that the digitized responses are plausible and, if not, modifying them appropriately. Editing rules can be developed for a single variable or for several variables in combination. The editing rules may specify acceptable values for a variable (e.g., an acceptable range of values) or acceptable relationships between two or more variables (e.g., an acceptable range for the ratio of two variables). Typically, editing identifies entries that are definitely in error (called *critical edits*) or are highly likely to be in error (called *query edits*). All critical edits must be corrected, while various rules may be applied to determine which query edits to address to reduce the cost of the editing process. This approach is sometimes referred to as selective editing (Granquist and Kovar, 1997).

Some surveys specify that respondents should be recontacted if the number of edit failures is large or if key survey items are flagged as erroneous or questionable. Thus, missing, inconsistent, and questionable data can be eliminated by the respondent's input; however, this is not always done either to save costs or because of the impracticality of respondent recontacts. In that case, values may inserted or changed by means of deducing the correct value based on other information on the questionnaire or from what is known about the sample unit from prior surveys, a process called *imputation*. Consistency checks, selective editing, deductive editing, and other editing functions can be performed automatically by specially designed computer software (discussed in Chapter 13).

In Chapter 9, methods for editing are examined under the rubric of statistical data editing (SDE). SDE involves two steps: *error localization* (identifies errant or missing data entries) and imputation (supplies a value for the errant or missing data item). The latter topic is discussed in much greater detail in Chapter 10. This chapter provides a comprehensive discussion of imputation methods, focusing primarily on methods for imputing a single value (as opposed to methods for multiple imputations). In addition, issues of inference in the presence of imputed values are explored.

### *1.2. Coding*

Coding is a procedure for classifying open-ended responses into predefined categories that are identified by numeric or alphanumeric code numbers. For example, the open-ended question "What is your occupation?" may have thousands of different responses. To be able to use this information in subsequent analysis, each response is assigned one of a much smaller number (say 300–400) of code numbers that identify the specific occupation category for the response. So that occupation categories are consistent across different surveys and different organizations, a standard occupation classification (SOC) system is used. A typical SOC code book may contain several hundred occupation titles and/or descriptions with a three-digit code number corresponding to each. In most classification standards, the first digit represents a broad or main category, and the second and third digits represent increasingly detailed categories. Thus, for the response "barber," a coder consults the SOC code book and looks up the code number for "barber." Suppose the code number is 411. Then the "4" might correspond to the main category "personal appearance workers," 41 might correspond to "barbers and cosmetologists," and 712 to "barber." In automated coding, a computer program assigns these code numbers to the majority of the cases while the cases that are too difficult to be accurately coded by computer are coded manually. A discussion of methods for coding open-ended responses can be found in Biemer and Lyberg (2003).

### *1.3. File preparation*

The file preparation step results in a file that is ready for data analysis. This step consists of a number of activities including weighting, weight adjustment, outlier analysis, and record linkage. For sampling with unequal probabilities, base (or selection) weights must be computed for the sample units. Weight adjustments can also be applied to compensate for unit nonresponse and frame coverage errors. Often the weights are developed in three steps. First, the base weight is computed for each unit as the inverse of the probability of selection of the unit. Next, the base weight is adjusted to compensate for unit nonresponse by a response propensity adjustment. This adjustment factor is usually computed as the inverse of the estimated probability of responding to the survey. Finally, additional weight adjustments might be performed to adjust for frame coverage error depending on availability of external information. These so-called calibration adjustments are intended to achieve additional improvements in the accuracy of the estimate. Chapter 8 provides a general introduction to survey unit nonresponse and the need for weighting survey data. It reviews various methods for computing response rates, examines response rate trends, and considers the relationship between response rates and nonresponse bias. Methods for weighting, especially to reduce nonresponse bias, are covered in some detail. Also covered are methods for variance estimation and confidence interval estimation in the presence of nonresponse.

Related to the topics of both weighting and imputation is outlier analysis, the subject of Chapter 11. In some surveys, a few units can account for up to 10% of an estimate of the population total. The situation is even worse if some of these extreme units are combined with large survey weights. Like data editing, the goal of outlier analysis is to identify these extreme values and confirm or correct them. Extremely large or small values of a survey variable that cannot be confirmed with the respondent may be set to missing and imputed (see Chapter 10).

Beaumont and Rivest (Chapter 11) distinguish between two types of outliers: those due to reporting errors (referred to as nonrepresentative) and those that are correct values but represent an extremely small part of the population (referred to as representative). Nonrepresentative outliers can be handled at the data collection and/or editing stages of a survey process using outlier detection techniques. Representative outliers offer the greater challenge to statisticians because whether these outliers are included or excluded in the calculations of the sample means or totals can dramatically impact the magnitude of these statistics. Although including representative outliers is statistically correct and produces design unbiased estimators of totals, they can noticeably increase the standard errors of the estimates. On the other hand, estimators that limit the influence of large values produce more stable estimates, but are biased. As Beaumont and Rivest show, the art of outlier treatment in survey sampling lies in the management of this bias-variance trade-off. Chapter 11 examines these issues in detail and discusses the major methods for dealing with representative outliers.

Another potential step in preparing files for data analysis is appending administrative or possibly census block or tract-level data to the survey data records. This might be done prior to the weighting step to provide additional auxiliary data for the weight adjustments. In addition, such data supplements can provide contextual variables to enrich data analysis. Linking survey records to external, auxiliary records requires the techniques of record linkage, which are discussed in Chapter 14. In countries using population registries, such linkages across data systems are facilitated by the existence of a unique identifier for each population member. However, in many other applications, such linkages must rely on a few fields such as first name, last name, and date of birth. In these situations, special techniques have been developed to achieve high levels of accuracy with a known level uncertainty. Other applications of record linkage methodology include the construction of multiple frames to avoid duplication of frame units, or the evaluation of survey or census coverage error.

### *1.4. Statistical disclosure analysis*

Chapter 15 takes up the important topic of preserving the confidentiality and privacy of survey respondents in the public release of survey data files. Virtually all national statistical institutes (NSIs) and many other survey organizations have policies regarding the release of macrodata and microdata to external users. *Macrodata* refer to files containing tabulations, counts, and frequencies. *Microdata* refer to files containing records that provide data about individual persons, households, establishments, or other units. *Disclosure protection* refers to efforts made by a survey organization or data supplier to reduce the risk that a specific unit in the population is identified as a unit in the data file, when such a disclosure could reveal information about the unit that is generally unknown. Thus, for any proposed release of tabulations or microdata to the public, the acceptability of the level of risk of disclosure must be evaluated. Statistical disclosure control (SDC) is a set of statistical techniques that help to evaluate the risk of reidentification and, if the risk is deemed too high, to reduce the risk by altering the data.

Chapter 15 begins with the basic concepts, goals, and essential approaches of statistical disclosure analysis. It casts SDC as an optimization problem that trades the risks of disclosure against the utility of the data to analysts. As an example, stripping the microdata records of all geographic identifiers (including primary sampling unit indicators) is often necessary to reduce disclosure risks. However, such identifiers are needed to appropriately estimate the design variances of the estimators. The chapter discusses these issues for both microdata and macrodata releases.

### *1.5. Data documentation and analysis*

The final processing step is data documentation in which a type of data file users' manual is created. This document describes the methods used to collect and process the data and provides detailed information on the variables on the file. For example, each variable on the data file might be linked to one or more questions on the questionnaire. If variables were combined, recoded, or derived, the steps involved in creating these variables are described. The documentation might also include information regarding response rates for the survey, item nonresponse rates, reliability estimates, or other information on the total survey error of key variables.

As noted in Chapter 13, new technologies have opened up many possibilities to integrate these data processing steps. Therefore, for some surveys, the sequence of steps might be very different from those described earlier. For example, it is possible to integrate data capture and coding into one step; likewise, data capture and editing can be integrated with coding. It is also possible to integrate editing and coding with data collection through the use of CAI technology. The advantage of integration is that inconsistencies in the data or insufficient information for coding can immediately be resolved with the respondent, which reduces follow-up costs and also may result in better information from the respondent. Many other possibilities for combining the various data processing steps may be feasible. The goal of integration is to increase the efficiency of the operations while improving data quality.

Chapter 13 reviews the major software packages available for all the processing steps listed in Fig. 1, including data collection (via CAI). This chapter addresses one of the most frequent errors made by data analysts who are not familiar with survey data, that is, incorrect use of weights. Such errors range from ignoring the weights completely to regarding them incorrectly as frequency weights in standard statistical packages such as SAS and SPSS. The chapter also considers a number of widely available survey analysis software packages such as SUDAAN, STATA, WesVar, and the special survey analysis modules of SAS and SPSS.

## 2. Data quality and data processing

As is clear from the previous discussion, the data can be modified extensively during data processing. Hence, data processing has the potential to improve data quality for some variables while increasing the error for others. Unfortunately, knowledge about the errors introduced in data processing is very limited in survey organizations and, consequently, such errors tend to be neglected. Operations are sometimes run without any particular quality control efforts, and the effects of errors on the overall accuracy as measured by the mean squared error (MSE) are often unknown, except perhaps for national data series of great importance.

As an example, although editing is intended to improve data quality, it misses many errors and can even introduce new ones. Automation can reduce some errors made by manual processing, but might introduce new errors. For instance, in optical recognition data capture operations, the recognition errors are not uniformly distributed across digits and other characters, which can introduce systematic errors (i.e., biases). For these reasons, quality control and quality assurance measures should be a standard part of all data processing operations.

The evaluation of measurement errors in surveys, including data processing errors, is the topic of Chapter 12. As this chapter explains, knowledge of the magnitudes of measurement bias and variance can serve multiple purposes.

- Information on the errors related to alternative data collection methods can be used to improve data collection methodology for future surveys.
- Estimates of the reliability and validity of survey questions can lead to improved questionnaire design.
- Information on the measurement error properties of survey variables used in the data analysis is important for data users and analysts who need to understand the limitations of the data to account for them in a proper way.

Chapter 12 presents five modeling approaches that are appropriate for the study of measurement error, three of which focus primarily on classification errors and two on the error in continuous data. The chapter begins with the model first espoused by Hansen et al. (1964), which can be applied to any type of variable. Much of the chapter is spent on examining the essential concepts and methods underlying latent class analysis (LCA) of measurement error, including Markov latent class analysis (MLCA) for panel data. The chapter closes with a discussion of some common approaches for the assessment of measurement error in continuous data, using structural equation modeling techniques.

*Sample Surveys: Design, Methods and Applications, Vol. 29A*
ISSN: 0169-7161

DOI: 10.1016/S0169-7161(08)00008-4

8

# Nonresponse and Weighting

*J. Michael Brick and Jill M. Montaquila*

Broadly defined, *nonresponse* is the failure to obtain a valid response from a sampled unit. It is of concern to survey methodologists and practitioners since complete response is assumed by the randomization or design-based theory that allows for inference from a sample to the target population. Nonresponse has the potential to introduce bias into survey estimates and reduce the precision of survey estimates. As a result, survey practitioners make efforts to minimize nonresponse and its effects on inferences from sample surveys. However, even with the best of efforts there will be nonresponse, so it is essential to understand its potential effects and methods that can be used to limit these effects.

We begin by discussing nonresponse in surveys, the reasons for nonresponse, and methods used to increase response rates in surveys. In Section 2, we define response rates, review methods of computing response rates, and then examine the trends in response rates over time. Section 3 examines the relationship between response rates and nonresponse bias, and methods for modeling response propensities and estimating bounds on nonresponse bias. Section 4 reviews weighting in surveys. It begins with a brief review of the reasons for weighting in surveys and the steps in weighting, before concentrating on weighting methods to reduce nonresponse bias including standard weighting class adjustment methods and calibration weighting. Section 5 examines variance estimation and confidence interval estimation in the presence of nonresponse, including a discussion of some of the software available for doing this. We conclude with a discussion that includes some areas that need further research.

## 1. Nonresponse in surveys

While all types of nonresponse result in missing data, it is useful to classify nonresponse by the pattern of missingness. When a sampled unit fails to respond at all to the data collection efforts, this type of missing data is called unit nonresponse (the failure of a sampled unit to respond to the survey). Item nonresponse is another form of missing data that occurs when a unit responds to some of the data items in the survey but fails to answer one or more items. A third type of nonresponse is partial or wave nonresponse. Partial nonresponse occurs when only a portion of the survey is completed. For example, a survey may involve sampling households and screening for persons eligible for the main or extended interview. If the screening interview is completed and an eligible person is sampled but does not respond, this might be considered partial nonresponse. Partial nonresponse also occurs when a respondent to an initial wave of a longitudinal survey fails to respond to a subsequent wave. In longitudinal or panel surveys, wave nonresponse is often an important component of nonresponse.

These classes or types of nonresponse are related to the main methods used to deal with survey nonresponse in data files. Typically, some form of weighting adjustment is used to compensate for unit nonresponse. Imputation, on the other hand, is usually the method chosen to deal with item nonresponse. While this is common practice, weighting adjustment can be used for item nonresponse and imputation for unit nonresponse. Partial nonresponse is a hybrid in the sense that it may be treated as either unit nonresponse or item nonresponse. Other factors, such as processing costs and accuracy of the compensation method, play an important role in determining whether the partial nonresponse is handled by weighting or imputation. For example, the Survey of Income and Program Participation, a longitudinal survey conducted by the U.S. Census Bureau, imputes for wave nonresponse if the preceding and subsequent waves are obtained, but weights for all other forms of wave nonresponse.

Another type of missing data occurs when eligible units in the target population are not included in the sampling frame; this is undercoverage or noncoverage rather than nonresponse. Noncoverage is not studied in this chapter.

The remainder of this chapter focuses on unit nonresponse and methods for adjusting the unit nonresponse. Wave nonresponse is discussed in Chapters 5 and 34, and item nonresponse in Chapter 10.

### *1.1. Reasons for nonresponse*

Unit nonresponse may occur for various reasons, but most nonresponse may be classified into two broad categories: accessibility issues and amenability issues. *Accessibility* refers to the ability to make contact with the sampled unit. For example, accessibility issues may be the inability to find anyone at home for an in-person or telephone survey, or the inability to trace a unit successfully from a list sample. *Amenability* refers to the unit's willingness to cooperate with the survey request after contact has been made. A third, and generally less significant, cause of unit nonresponse is loss due to administrative issues, such as mail questionnaires that are received too late or are lost in processing. Stoop (2005) discussed various theories of survey participation and provided an extensive literature review to address those theories.

Much of the nonresponse literature examines the effect of the characteristics of the respondents and interviewers on amenability. Goyder (1987), Groves and Couper (1998), and Stoop (2005) reviewed this extensive literature and examined the differences in response levels by factors, such as the sex, age, and geography of the respondent. The differences by demographic characteristics are relatively consistent across surveys. The characteristics of interviewers, on the other hand, rarely predict those who are likely to obtain higher response rates. Groves and Couper (1998) reviewed this literature as well as the interaction between respondents and interviewers. Below, some of the other factors that affect either amenability or accessibility are briefly mentioned.

Many studies have found that mode of data collection is an important factor in determining the level of nonresponse, especially in household surveys. The highest response rates are often attained by face-to-face surveys. Telephone survey response rates are usually lower than in-person response rates. Mail surveys generally have the lowest response rates. Over two decades ago, Goyder (1985) noted, "Many texts caution against expecting mailed questionnaire response from a general population to exceed about 30 percent" and cited variations among authors in the range of "typical" response rates. Krosnick (1999) referred to mail surveys as "undesirable" due to low rates of response. Tourangeau et al. (2000) discussed some of the psychological reasons that make respondents more amenable to interviewer-administered surveys compared with self-administered modes. Accessibility also varies substantially by mode. For example, households without telephones are not accessible in telephone surveys.

The type of unit sampled also affects response rates. Organizational or establishment surveys generally have lower response rates than surveys of individuals (Tomaskovic-Devey et al., 1994). Among a number of factors studied, Tomaskovic-Devey et al. (1994) found that motive to respond had the most significant effect on response rates in organizational surveys, although authority to respond and capacity to respond were also associated with the quality of survey responses. Lynn and Sala (2004) concluded that a flexible approach to interviewing in establishment surveys, including offering multiple modes for response, can boost response rates considerably. In reporting on U.S. military personnel surveys, Newell et al. (2004) stated that "the primary reasons why Navy surveys are not being returned is a belief that they have no impact, general apathy over the survey process, and survey length."

The introduction of new technologies may also affect the level of nonresponse. The potential of technology to affect response rates is most prominent in telephone surveys, where answering machines, caller ID devices, and privacy managers are technological barriers to obtaining high response rates. Despite the popularity of these devices, their effects are still largely unknown. Studies have found that households with answering machines or caller ID devices were harder to reach, but once contacted, they were as likely to respond as those without answering machines (Piazza, 1993; Link and Oldendick, 1999). However, Smith (1995) noted that "refusals and the related problem of call screening and answering machines were considered the most important problems over the next 10 years." Tuckel and O'Neill (2002) reported that although refusal rates appeared to have leveled off, noncontact rates continued to increase, causing nonresponse to increase.

Other societal changes also influence the response levels. In the United States, a National Do Not Call Registry (often referred to as the "Do-Not-Call list") was implemented in 2003 to prohibit telephone solicitations for households that have registered their telephone numbers. Although survey research firms are exempted from the Do-Not-Call restrictions, some have hypothesized that telephone survey response rates suffer because respondents are unaware of this exemption and view the survey request as an intrusion. However, others hypothesized that because of the reduction in unsolicited calls due to the Do-Not-Call list, respondents should be more willing to comply with the survey request. Link et al. (2006) provided some empirical evidence suggesting the Do-Not-Call list has had little or no effect on state-level response rate.

The importance of the topic of the survey to the sampled person has long been identified as an important determinant of nonresponse. Groves et al. (2004) used this idea to propose a leverage-salience theory of response. They pointed out that items directly related to the survey topic are most susceptible to nonresponse bias, and that people who are interested in or have characteristics related to the survey topic may be more likely to cooperate. To the extent that this is true, comparisons of easier and harder to interview respondents might provide confirmation of the theory. Some studies have attempted to examine this theory empirically. For example, Voigt et al. (2003) found "intermediate and late responders" to be slightly younger, more likely to be non-White, and less educated than "early responders." We discuss these models in more detail in a subsequent section.

### *1.2. Methods to increase response rates*

As discussed above, response rates differ across survey modes. However, response rates may also vary widely across surveys even when the surveys are conducted using the same mode. One explanation for these differences in response rates is that different levels of effort are expended in the survey to obtain the responses. For example, Edwards et al. (2006) noted a wide variation in effort among in-person surveys. Such differences in levels of effort also exist in mail and telephone surveys.

Because of the potential for differences among initial cooperators, refusers, and reluctant (or "late" or "resistant") respondents, a number of data collection and estimation strategies have been used to attain high response rates, with the ultimate goal of minimizing the effects of nonresponse bias for the estimates (e.g., Holbrook et al., 2008). Some standard strategies have been used for decades, while others have been introduced more recently to make use of emerging technologies and to combat declines in response rates. Some examples of such strategies are given below.

For in-person and telephone surveys, efficient contact or calling strategies that result in multiple contact attempts, on varying days of the week at various times of the day, have been an important part of the effort to obtain high response rates for many years. The field period must be long enough to allow sufficient numbers of contact attempts (e.g., Groves and Couper, 1998; Weeks, 1988). Following good questionnaire design principles (e.g., well-written introductions and clear, well-executed skip patterns), setting limits on the length of the interview, and using well-trained interviewers contribute substantially to effective data collection efforts.

Translation of questionnaires into multiple languages (or the use of interpreters) and using multiple modes of administration are relatively newer strategies. Tactics such as refusal conversion, multiple mailings (including, in the case of telephone or in-person surveys, both advance letters and mailings to follow up with nonrespondents), and incentives (Cantor et al., 2008) are becoming more widely used in attempts to elicit cooperation. For telephone surveys, answering machine messages are used to introduce the study and attempt to mitigate any effects of call screening. Heberlein and Baumgartner (1978) found that for mail surveys, the number of follow-ups and topic salience were the two most significant predictors of response.

Nonresponse follow-up has been found to be an important tool in understanding and assessing the nonresponse bias. Teitler et al. (2003) implemented a nonresponse follow-up study to examine how characteristics and costs per case vary by mode. They used proxy data available from an early stage of interviewing to examine nonresponse and found large differences between respondents and nonrespondents in education and race.

They used this information to inform their nonresponse weighting strategy, by using education and race as auxiliary variables in a weighting adjustment for nonresponse. The goal was to reduce the bias associated with the differences between respondents and nonrespondents. A more complete discussion of the use of weighting adjustments to reduce nonresponse bias is presented later in this chapter.

## 2. Response rates

For the past several decades, the response rate has been one of the most important and widely used indicators of survey quality. Two reasons why response rates became so important are that response rates are measurable, and the common assumption that high nonresponse was more likely to be associated with high levels of nonresponse bias. In the next section, we discuss the evidence showing the relationship between response rates and nonresponse bias is not as clear as had been assumed. First, we define unit response rates and discuss trends in response rates in surveys over time.

### *2.1. Computation of response rates*

The *unit response rate* (often simply referred to as the *response rate*) is the ratio of the number of respondents to the number of eligible units. In some cases, this rate is easily defined and computed, whereas in other cases the denominator must be estimated. Lohr (1999) pointed out that sometimes organizations use nonstandard definitions of response rates to show higher "response rates" than would be achieved by standard methods. The American Association of Public Opinion Research (AAPOR) published a booklet specifically developed to help standardize methods of computing response rates. The computations presented here are those described in AAPOR (2006).

Response rates may be weighted or unweighted. The unweighted rate, computed using the raw sample counts, provides a useful description of the success of the operational aspects of the survey. The weighted rate is computed by summing the weights (usually the reciprocals of the probabilities of selection) for both the numerator and the denominator. Since the weights allow for inference of the sample data (including response status) to the population level, the weighted rate gives a better description of the success of the survey with respect to the population sampled. In establishment surveys, weighting the response rates by the size of the establishment (e.g., number of employees, number of students, dollar volume) reflects the importance of the institutions to the population total (otherwise the response of an institution with one employee would account for the same as an institution with 10,000 employees).

Let $s$, $r$, and $nr$ denote the set of sampled units, the set of respondents in the sample, and the set of nonrespondents in the sample, respectively. Further, let $e$, $ie$, and $ue$ denote the sets of units in the sample that are known to be eligible, known to be ineligible, and have unknown eligibility, respectively, so that $e = r \cup nr$ and $s = e \cup ie \cup ue$. The unweighted unit response rate, which we denote $\hat{\phi}^{(u)}$, is

$$\hat{\phi}^{(u)} = \frac{\sum\limits_{r} 1}{\sum\limits_{e} 1 + \alpha \sum\limits_{ue} 1}, \tag{1}$$

where $\alpha$ is an estimate of the proportion of cases with unknown eligibility that are eligible. The weighted unit response rate, denoted $\hat{\phi}^{(w)}$, is computed as follows:

$$\hat{\phi}^{(w)} = \frac{\sum_{r} w_i}{\sum_{e} w_i + \alpha \sum_{ue} w_i}, \tag{2}$$

where $w_i = \pi_i^{-1}$ is the base weight or reciprocal of the selection probability of unit $i$. If the response rate for an establishment survey is weighted by the size of an institution, then $w_i$ in (2) is replaced by $w_i s_i$, where $s_i$ is the size of institution $i$.

Thus, for the computation of response rates and for adjusting of nonresponse, it is necessary to classify each unit in the sample as a respondent, a nonrespondent, an ineligible, or a case with unknown (undetermined) eligibility. Simply put, respondents are those eligible units who complete the survey. However, this class is complicated by *partial completes*, units who respond to part of the survey request but not the entire request. (These may include, e.g., units who break-off during a telephone interview, those who neglect to complete one page of a multipage mail questionnaire, or those who respond to certain items but fail to respond to other items). Survey practitioners establish rules for handling partial completes that address which items or sets of items must have been completed, or what sections of the interview had to have been finished, to classify a partial complete as a respondent.

*Nonrespondents* are eligible sample units for which a response was not obtained. These may include units who could not be located; those who were located but could not be contacted; those who could not complete the survey due to reasons, such as a language barrier, illness, or extended periods away from home; and those who refuse to participate.

*Ineligible units* are sampled units that are not part of the target population. In an establishment survey, these include units that have gone out of business. In a household survey, units that are vacant during the data collection period would be ineligible. In a list sample, the people on the list may include those who are eligible and those who are not.

In multistage samples, ineligible units may be nested within eligible units. For example, in a school-based survey of students enrolled in fourth grade, ineligible students (students enrolled in grades other than fourth grade) are nested within eligible schools (schools containing students enrolled in fourth grade). Another example is the case of a screening survey, in which a household screener is administered to identify members of a rare population. In such cases, if the larger unit (in these examples, the school or the household) does not respond, it is important to recognize that only a proportion of units within that larger unit are eligible, or that only a proportion of larger units may contain eligible units, whichever the case may be.

Units with unknown eligibility (or *undetermined units*) are typically units whose eligibility could not be determined due to failure to make contact with the unit. An example of unknown eligibility is a "no answer" telephone number in a random digit dial (RDD) household telephone survey—a case for which every call attempt resulted in a ring, but no answer. In general, it cannot be determined if a "no answer" telephone number is assigned to a household and thus eligible for the survey.

An important issue in the computation of response rates is the estimation of $\alpha$, the proportion of undetermined units assumed to be eligible. (See AAPOR, 2006). Bounds on the response rate may be obtained by considering the two extremes, $\alpha = 0$ and $\alpha = 1$. The Council of American Survey Research Organizations (CASRO) method response rate is computed by allocating the units with unknown eligibility in the same proportion observed among cases with known eligibility status, that is, $\alpha = \frac{\sum_e w_i}{\sum_e w_i + \sum_{ie} w_i}$. Various other approaches have been used, including the use of survival analysis methods with call record data to estimate $\alpha$ and using auxiliary data to classify units. See Smith (2002) for a detailed discussion of methods of estimating $\alpha$, including the CASRO method, the survival method, and other methods based on references to auxiliary data.

## *2.2. Trends in response rates*

Achieving high response rates in sample surveys is a longstanding concern that has become increasingly difficult in recent years. As early as 1946, Hansen and Hurwitz expressed concerns with low mail survey response rates and proposed a mixed-mode approach of following up a subsample of mail survey nonrespondents with in-person attempts. More recently, Atrostic et al. (2001) and Curtin et al. (2005) showed declining response rates in several household surveys conducted in the United States. As noted earlier, this trend toward lower response rates is happening despite the introduction of additional procedures aimed at increasing response.

Although much of the literature on trends in response rates focuses on trends in U.S. surveys, the results are similar internationally. Smith (1995), Groves and Couper (1998), and Holbrook et al. (2003) have cited numerous studies demonstrating declines in response rates in both U.S. surveys and those in other countries. De Leeuw and de Heer (2002) found declines in surveys in developed countries. Synodinos and Yamada (2000) reported overall response rate declines of about 10% over the last quarter of the 20th century in Japanese surveys. In their study, Synodinos and Yamada noted increasing difficulty in contacting respondents and hypothesized this may be due to longer work hours and longer commutes. They also observed an increase in refusal rates over that same period.

De Heer (1999) noted differences in response rates and trends among labor force surveys and expenditure surveys in different countries, but attributed these differences mainly to differences in survey practices among various survey organizations. Stoop (2005) also noted considerable differences among European countries in terms of both response rates and reasons for nonresponse for the European Social Survey (see Stoop, 2005, Chapter 10).

Steeh et al. (2001) argued that a decline in response rates observed over the 1960s and 1970s leveled off during the 1980s and 1990s, even though the composition of nonrespondents changed over that period; refusal rates declined while noncontact rates increased. In a study of mail survey nonresponse, Brennan and Hoek (1992) concluded that people tended to be consistent in whether or not they responded to mail survey requests, and that refusers differed from nonreturners in their likelihood to respond to subsequent mail survey requests.

Earlier we noted that societal changes may affect response rates. These changes can have both long-term and short-term effects on response rates. For example, Goyder (1985) noted the long-term effect of the increases in literacy rates following World War II on mail questionnaire response rates. Harris-Kojetin and Tucker (1999) found evidence of short-term relationships between refusal rates in the U.S. Current Population Survey and political and economic conditions.

## 3. The relationship between response rates and nonresponse bias

Thus far, we have discussed unit nonresponse and response rates. Of course, nonresponse is primarily of interest in surveys because missing data introduce the potential for bias. Two models, a deterministic and stochastic model, have been developed to relate nonresponse rates to nonresponse bias. Lessler and Kalsbeek (1992) and Särndal and Lundström (2005) reviewed these models.

### *3.1. Two models of response*

The first approach, the deterministic model, treats response as a fixed outcome, so that the population can be partitioned into respondent and nonrespondent strata (Cochran, 1977). Nonresponse bias of an estimated mean under this model is

$$\text{bias}(\overline{y}_{\text{NHT}}) = (1 - R)(\overline{Y}_r - \overline{Y}_m), \tag{3}$$

where $R$ is the (nonstochastic) proportion of units in the respondent stratum (the expected value of the weighted response rate), $\overline{Y}_r$ is the mean in the stratum of respondents, and $\overline{Y}_m$ is the mean in the stratum of nonrespondents. The bias depends on the relative sizes of the strata and the differences in the characteristic of the respondents and nonrespondents.

The second approach is based on a response propensity model that assumes response is stochastic, similar to a second phase of sampling (Särndal and Swensson, 1987). An important difference from two-phase sampling is that in the stochastic model of response, the response probabilities or propensities ($\phi_i$) are unknown. The response propensity model requires that $\phi_i > 0$ for all $i$ to allow unbiased estimation. Nonresponse cannot be accommodated entirely within the randomization framework because the theory requires that all the probabilities of selection must be known. Consequently, models must be used to address nonresponse bias.

Platek et al. (1978), Kalton and Maligalig (1991), Bethelehem (1988), and others have examined the relationship between response rates and nonresponse bias using response propensity models. As an example, consider the respondent ratio mean (the estimator of the total based on respondents divided by the sum of the weights of respondents). The nonresponse bias for this estimator is

$$\text{bias}(\overline{y}^{*}_{\text{NHT}}) \approx \overline{\phi}^{-1} \sigma_\phi \sigma_y \rho_{\phi,y}, \tag{4}$$

where $\overline{\phi}$ is the mean of the response propensities, $\sigma_\phi$ is the standard deviation of $\phi$, $\sigma_y$ is the standard deviation of the $y$, and $\rho_{\phi,y}$ is the correlation between $\phi$ and $y$. The estimated respondent mean is unbiased if $\phi$ and $y$ are uncorrelated. This expression clearly demonstrates that nonresponse bias is defined for a specific statistic, even though the unit response rate is the same for all statistics from the survey.

In fact, the relationship between the response propensities (*response rates* are observable outcomes while *response propensities* are unobservable) and the nonresponse bias depends on both the type of statistic and the estimator. For example, Groves et al. (2004) gave an expression for the nonresponse bias of the difference between two means. Bethlehem (1988) and Kalton and Maligalig (1991) examined the bias of the poststratified estimator and raking estimator, respectively. Brick and Jones (2008) extended these results to other types of statistics. The expressions for the bias are similar to (4); for both estimators, the bias involves covariances between the response propensities and characteristics within partitions defined by the auxiliary data used in the estimators.

### *3.2. Response propensity modeling*

Response propensity models are a valuable way to think about nonresponse bias and to motivate weighting methods to reduce bias. As the stochastic view of nonresponse has become more accepted, many researchers seek to develop models to estimate response propensities. If the posited response propensity model holds and the propensities can be accurately estimated, then the estimated propensities can be used to adjust the weights and create estimates with lower nonresponse biases. A variety of nonresponse weighting schemes that are either based on response propensity models or are consistent with some response propensity model are discussed below.

Despite the promise of response propensity models, the ability to eliminate nonresponse bias in survey estimates is limited. Response propensities are unknown, and the only observables are whether or not the sampled unit responds. Accurately estimating probabilities based on zero-one observations is often difficult and unreliable. This is because response propensity models are based upon largely speculative theories of individual or institutional behavior regarding such actions as responding to a survey.

Särndal and Lundström (2005) pointed out that nonresponse bias can be reduced if powerful auxiliary variables are available. A problem, especially in household surveys, is that variables that are highly predictive of response propensities are rarely available. Eliminating nonresponse bias for all statistics produced from a survey is a difficult if not impossible task, especially when surveys generate estimates that were not even contemplated when the survey was designed or when the survey weights were prepared.

Since perfect estimation of response propensities is infeasible in practice, other ways of taking advantage of the relationship between response propensities and the statistics being estimated may be considered. One approach is to establish approximate bounds on the potential bias (Montaquila et al., 2008). For example, assume the statistic is a proportion ($P$) and we wish to evaluate the potential for nonresponse bias. To bound the bias, suppose the response propensities for all the units with the characteristic are $\phi_1$, and are $\phi_2$ for all those without the characteristic. In this case, (4) simplifies to $\text{bias}(p^*_{\text{NHT}}) \approx P(1-P)(\phi_1-\phi_2)\{P\phi_1+(1-P)\phi_2\}^{-1}$. Since the bias depends only on $\lambda = \phi_2\phi_1^{-1}$, it can be further simplified to $\text{bias}(p^*_{\text{NHT}}) \approx P(1-P)(1-\lambda)\{P+(1-P)\lambda\}^{-1}$. The bias is negative when $\lambda > 1$ and positive when $\lambda < 1$. Since the bias is a monotonic decreasing function of $\lambda$ (for $0 < p < 1$), a bounded estimate of the bias can be computed by choosing a bounding value for $\lambda$. For example, if $\phi_2$ is very likely to be no more than 1.5 times $\phi_1$ (e.g., response rates of 30% and 45% for those with and without the characteristic), then the bound on the bias in the respondent proportion for a 50% characteristic is $-10\%$ and the relative bias is $-20\%$. The bounds computed using this approach may be large, especially for values of $P$ less than 50%. These bounds may still be useful, especially when compared with bounds computed using the deterministic approach to nonresponse. Upper bounds are also obtained by specifying $\lambda$.

Specifying $\lambda$ requires speculating on the causes of nonresponse. As discussed earlier, the two major causes of nonresponse in surveys are related to *accessibility* (the ability to contact the sampled unit so that they can be surveyed) and *amenability* (the likelihood of gaining cooperation from the sampled units that are contacted). In estimating a proportion, a large value of $\lambda$ is likely only if the causes of nonresponse are related to the presence of the characteristic. Groves (2006) proposed a "common cause" model for nonresponse, in which the relationship is part of a causal sequence. In this model, the probability of responding to the survey and the distribution of the characteristic differ because the same factor(s) affects both. It might be possible to infer relationships of this nature between response propensities and characteristics from empirical studies. It could also be part of the reason that these studies tend to find weak relationships between response rates and nonresponse bias (e.g., Groves, 2006).

Causal relationships could be related to either accessibility or amenability. For example, consider people who travel a great deal of time; these persons might be expected to have low response propensities due to the inability to contact them, and they might have characteristics that are different from other persons for topics such as time use or travel. Another example is a list survey where the contact information for sampled units comes from different sources. If persons on the list with poor contact information are those who have infrequent dealing with medical providers, then estimates of health may be biased due to this common relationship. Another example is bias in estimates of economic status when persons who do not speak the native language or persons who are illiterate cannot be surveyed and are likely to have lower than average economic status.

With respect to amenability, the topic and sponsorship of the survey are aspects that have been shown to cause or at least be highly correlated with differential response propensities for those with and without a characteristic. Groves et al. (2000) incorporated this hypothesis into their leverage-saliency theory. For example, sample persons who are involved in a socially undesirable activity, such as using illegal drugs, may be more likely to refuse a survey that is sponsored by a drug enforcement agency; sampled businesses who do not provide a service to their employees, such as insurance, may be less likely to cooperate in a survey if the introductory letter indicates that the survey will be about insurance benefits. Groves et al. (2004) and Groves et al. (2006) conducted experiments to study this theory and found at least some support for the theory. However, even when they attempted to produce nonresponse bias intentionally, the observed differences in response propensities were minor.

Although one might be tempted to surmise from these empirical investigations that large values of $\lambda$ are unlikely to result from differences in response propensities associated with amenability, there are many exceptions with very large values of $\lambda$. For example, Abraham et al. (2006) found that respondents in a labor force survey who reported being a "volunteer" were much more likely to respond to a follow-up time use survey than those who had not volunteered. Differences in amenability could bias estimates of time spent in volunteering. Notice that even in the aforementioned case, nonresponse bias may be large only for a few statistics related directly to amenability, and for many other estimates of time use, the bias may be small or negligible.

Nonresponse biases due to inaccessibility are often easier to recognize. The survey literature has many studies showing factors related to accessibility. For example, households with many members are easier to contact than those with only one or two members; households with children are easier to contact, young adults are harder to contact, and persons who are socially isolated are even harder to contact. If the factors associated with accessibility are related to a statistic being estimated, then nonresponse bias resulting from a substantial value for $\lambda$ is likely. In a telephone survey of cell or mobile phone numbers in the United States, Brick et al. (2006) found substantial nonresponse bias in estimating the percentage of households with a cell phone but no landline (cell-only households). The bias in this statistic was attributed to the difference in accessibility of the cell-only households and the numbers linked to persons with both types of telephone service. They conjectured that persons with cell-only service rely on their cell phone for much of their phone conversations and are likely to answer their cell phone calls, whereas in some households with both types of service the cell phone is often used only for emergencies so accepting incoming calls on their cell phone is rare.

When deciding on the ratio of the response propensities for bounding the bias, the effects of both accessibility and amenability must be considered simultaneously. Sometimes the nonresponse bias due to one of the effects is partially offset by the nonresponse bias due to the other. An example is given by Lin and Schaeffer (1995) where nonresponse bias due to refusals was in the opposite direction of that due to inability to contact the sampled unit; Montaquila et al. (2008) provided another example.

If the direction of the bias due to amenability and accessibility is consistent, then large nonresponse biases can occur, and this is the situation that is of most concern in practice. For this reason, understanding the sources of nonresponse bias and the likely direction of the biases is essential; it enables researchers to predict which estimates are likely to have large nonresponse biases and those likely to have small nonresponse biases. Knowledge of the sources of nonresponse can also help in the construction of weighting schemes that have the potential to reduce the biases of the estimates. For example, Kennickell and McManus (1993) reported results from a sample of wealthy individuals, where the sampling frame contained information about the income and wealth of the individuals. The goal of the survey was to estimate statistics related to overall financial assets. Kennickell and McManus observed that response rates differed depending upon the known values of wealth of the individuals. The wealthiest individuals were both harder to contact and less likely to respond when contacted. The results of this combination were very large nonresponse biases in estimates of assets and financial characteristics. The biases could not be reduced or eliminated without having the data from the sampling frame.

We conclude this discussion of the relationships between response propensity and nonresponse bias by discussing a couple of the assumptions of response propensity models. One of the assumptions is that $\phi_i > 0$ for all $i$. If some units will not respond to a particular survey, then a better model would be a hybrid of the deterministic (for those with $\phi_i = 0$) and stochastic models. Deming (1953) explicitly considered units with zero response propensities. He called those who never participate as "permanent refusers" while others have used the term "hard-core nonrespondents." We prefer "persistent nonrespondents" because the units may be either those that cannot be contacted or those that refuse under all circumstances pertinent to the survey. The existence of persistent nonrespondents essentially adds a noncoverage component to the nonresponse bias, and thus distorts the relationship between bias and response propensity discussed earlier. Any nonresponse bias analysis based only on the response probability model assumptions does not account for the possibility of the existence of persistent nonrespondents.

A second assumption that is at least implied in the model is that response propensities are specific to both the units sampled and the survey conditions. For example, the same units may have different response propensities depending on key survey conditions, such as the mode, the content of the survey, the length of the survey, the survey sponsor, and the use of incentives. As noted earlier, practitioners manipulate these factors to increase response rates. For example, in a nonresponse follow-up, the mode may be changed, say, from mail to face-to-face to increase the response rate. One way of thinking about this is that the mode switch essentially changes the response propensities of the respondents by modifying the survey conditions.

## 4. Weighting for nonresponse

As mentioned in the previous section, weighting adjustments are often used to reduce nonresponse bias in the estimates from sample surveys. To provide some context for the discussion of nonresponse weighting adjustments, we first briefly describe weighting in surveys generally. We then focus most of the discussion on nonresponse weighting adjustments.

### *4.1. Reasons for weighting in surveys*

Weighting can be viewed as a means of adjusting for missing data in a broad sense, with unit nonresponse being just one form of missing data that use weighting adjustments. For example, weighting compensates for missing data associated with units in the population that are not sampled (*sampling weights*) and for missing data due to units that are not in the sampling frame (*noncoverage weight adjustments*).

Assigning a weight to a record in the data file is essentially a computational tool for implementing an estimation scheme to deal with missing data. The sampling or *base weight* (the inverse of the probability of selection) may be attached to each sampled unit. Assuming no other missing data, the product of the weight and the $y$-characteristic summed across all sampled units yields the familiar Horvitz–Thompson (Narain, 1951; Horvitz and Thompson, 1952) estimate of a total. Many estimators used in practice are implemented by devising a scheme for weights specific to an estimator and attaching a weight to each record in the data file. Since the weights implement an estimation scheme, the same procedures can be used for many statistics. The benefits of being able to implement complex estimation schemes using the same computational procedure with weights are quite significant.

The rationale for using weights presented here is based on Kish (1992). The first reason for weighting is to account for data that are missing due to sampling. Varying probabilities of selection are frequently used in sample surveys. For example, optimum allocation or the desire to produce precise estimates for domains or subgroups may result in differential selection probabilities (i.e., different units having different probabilities of selection). To produce consistent and approximately unbiased estimates in the design-based paradigm, weights are attached to the records to account for these differential probabilities of selection.

While the first rationale for weighting deals with biases in the estimates due to sampling, the second is to reduce the variance of the estimates due to sampling. For example, suppose some auxiliary variables are known for all sampled units and for the population in aggregate. Estimation schemes such as calibration may reduce the variance of the estimates by using the auxiliary variables. These schemes are usually implemented by adjusting the base weights to create a new set of weights.

A third reason for weighting is to reduce the bias of the estimates from data missing due to nonresponse and noncoverage. Nonresponse and noncoverage are qualitatively different from data missing due to sampling because design-based theory relies solely on selection probabilities for making inferences. Nonresponse and noncoverage are outside the design-based framework and can only be handled by making model assumptions. Nonresponse is sometimes considered another stage of sampling; a stage in which the probabilities of nonresponding are unknown. As a result, weight adjustments for nonresponse often mimic methods used to adjust for differential sampling probabilities.

Noncoverage refers to data missing because the units were excluded from the sampling frame and thus had no chance of being selected. Models for noncoverage weighting adjustments are not as well developed, but one approach assumes the units omitted from the sampling frame are similar to units that are covered and uses a pseudorandomization model to justify adjusting the weights. An important difference is that nonresponse weighting adjustments can be based on auxiliary data known from the sampling frame, while noncoverage adjustments rely on auxiliary variables known for the total population.

A fourth reason for using weights is to force the estimates from the survey to be consistent with those known from another reliable source. Poststratification or calibration weighting adjustments are used to accomplish this goal, but here the primary goal is to make the estimates consistent with known totals from external sources (sometimes referred to as *control totals*), rather than to reduce variance or biases due to nonresponse or noncoverage. Matching controls provides face validity for the survey when the control totals are well known and widely accepted as being accurate. The commonly used epidemiological method called *standardization* is used to force the survey estimates to match specified control totals and improve the comparability of estimates across time and domains.

A final reason for weighting is to combine samples and produce estimates that are more accurate than those available from any of the samples individually. For example, multiple frame surveys (Hartley, 1974, Chapter 4) may have overlapping samples for some segments of the population. By combining the overlapping samples with weights, more precise estimates of the overlap and the total population may be produced. *Composite weights* are often used in this setting. Another objective may be to combine samples over time to produce reliable estimates for a domain. The American Community Survey is designed to cumulate samples over three- and five-year periods to produce estimates of sufficient precision for small geographic areas in the United States, and weights are used to achieve this result.

In any one survey, weighting adjustments may be used for one or more of the reasons mentioned above. Naturally, some trade-offs are made in deciding which of the adjustments should or should not be used. The traditional criterion is to minimize the mean square error of the estimates, but mean square error is difficult to minimize because the magnitude of the bias of the estimate is often unknown. Practitioners must decide whether adjustments that are likely to increase the variance of the estimates are likely to reduce the potential bias sufficiently to warrant their use.

Until recently, major statistical analysis software packages made little or no attempt to deal with sample surveys and assumed simple random sampling without weighting. Because weighting has become such a common practice in sample surveys, most standard statistical packages now handle weights at least for producing estimates. As a result, most statistical packages can use weights to produce design-unbiased first-order estimates. Of course, the weights alone are not sufficient to produce design-unbiased second-order estimates; for that, other information is required on the sampling scheme and estimation method. Consequently, specialized survey software is still needed to compute design-consistent standard errors in most cases. Some general statistical software packages, such as Stata®, have been making great strides in developing software that incorporates both weights and the sample design and estimation scheme appropriately, but there is more work to be done.

## *4.2. Steps in weighting*

A typical sequence of weighting steps is presented by Brick and Kalton (1996). The weighting begins with the creation of base weights and these weights are then adjusted sequentially for the reasons listed in the previous section. The base or sampling weight compensates for nonsampled units and is used to produce the Horvitz–Thompson estimator of the population total. Heuristically, the base weight can be conceived of as creating a data set in which the sampled unit represents itself and $(\pi_i^{-1} - 1)$ other units in the population. This approach is consistent with randomization theory since the selection probabilities are known for all sampled units and no model assumptions are required.

After creating the base weights, the weights may be adjusted for nonresponse and then some sort of calibration adjustment may be applied to the nonresponse-adjusted weights. As noted above, the calibration adjustments may be used to accomplish several goals simultaneously—reducing sampling error, reducing residual nonresponse bias and noncoverage bias, and matching known control totals.

While these weighting steps may be prototypical, the specific sequence of weight adjustments varies depending on the type of survey and the auxiliary data available for adjustment. For example, the data available for weighting in household surveys (Chapter 16) are quite different from the data available in business surveys (Chapter 17). One dimension that often affects the weighting is the type of sample, and weighting approaches used in different sampling schemes are described elsewhere (e.g., Chapters 3, 4, 5, 6, 7, 31, and 34). Similarly, samples drawn to examine specific topic areas are frequently characterized by types of auxiliary data that can be used in weighting. Sample designs for several topic areas are covered elsewhere (e.g., Chapters 16, 17, 18, 19, 20, 22, and 38).

Calibration methods are used in nonresponse weighting, as well as for other purposes as mentioned above. We discuss calibration uses for nonresponse weighting in this chapter, but general issues related to calibration weighting methods are discussed in Chapter 25. Other weighting methods that are not specifically for nonresponse are not discussed in this chapter. See Chapter 26 on estimating functions and Chapter 30 on empirical likelihood methods, for example. Another relevant weighting method is the trimming of weights to deal with outliers when the outliers are due to a few large weights. Chapter 11 covers this issue in some detail.

### *4.3. Nonresponse weighting adjustments*

Expressions like (4) show that to reduce bias, both response propensities and the characteristics being estimated must be taken into consideration in designing weighting adjustments. Researchers have long recognized this and have suggested modeling key statistics being produced from a survey as an approach to nonresponse weighting (e.g., Kalton, 1983). Little (1986) labeled weighting methods based on modeling the distribution of survey characteristics as predicted mean stratification. He contrasted this approach with response propensity stratification, which he defined as using the modeled response propensities to create categories or classes that are used in the nonresponse adjustment. Response propensity stratification is thus a special case of using the modeled response propensities in adjusting for nonresponse. Little (1986) pointed out that predicted mean stratification may reduce both bias and variance while response propensity stratification may only reduce bias. In practice, both the response propensity and survey characteristics are modeled in adjusting for nonresponse.

The key to modeling both response propensities and survey characteristics is the information available for modeling. The two types of information are variables known for all sampled units but not for the entire population (*sample-based*) and variables or functions of variables known for the entire population (*population-based*). Brick and Kalton (1996) used the sample and population terminology while Lundström and Särndal (1999) referred to these as *InfoS* and *InfoU* auxiliary data. Although the data requirements differ, there is virtually no difference in terms of nonresponse bias reduction for population and sample-based estimators. Population-based estimators do have advantages in terms of variance reduction and accounting for noncoverage bias.

One method of nonresponse weighting adjustment is to directly model the response propensities for the sampled units using the inverse of the estimated propensities as the weighting adjustment. The estimator is

$$\hat{y}_1 = \sum_r w_i \hat{\phi}_i^{-1} y_i, \tag{5}$$

where $\hat{\phi}_i$ might be estimated by logistic or probit regression modeling. Other approaches have also been used. For example, Da Silva and Opsomer (2004, 2006) used a nonparametric method.

An early method of directly estimating response propensities was suggested by Politz and Simmons (1949). In their method, the interviewer asked how often the respondent would be at home on different days to estimate the propensity of a respondent to be contacted. The inverse of this estimate is the weighting adjustment factor. Notice that this adjustment only accounts for noncontact. Another example of this type of nonresponse adjustment weight was explored by Bartholomew (1961).

The Dunkelburg and Day (1973) approach, in which the weights of the difficult-to-complete cases are increased to account for the nonrespondents, is closely related to adjusting the weights as if the additional level of effort is equivalent to a two-phase weighting adjustment. For example, Waksberg et al. (1993) subsampled nonrespondents and then adjusted the weights of those completed in the subsample to account for all nonrespondents eligible for subsampling.

Little (1986) noted that direct estimates of response probabilities, such as those derived from logistic regression models, may lead to unstable estimates if some of the estimated probabilities are close to zero. This may also apply to other methods of computing direct estimates of the response propensities. For example, in the Politz and Simmons (1949) model, the adjustments vary by a factor of six. Little (1986) suggested sorting the sample by estimated response propensities, forming five categories based on the quintiles of the response propensity distribution, and assigning the same weighting adjustment to all sampled units within a category.

### 4.4. *Nonresponse calibration weighting*

Lundström and Särndal (1999) proposed using calibration estimators as a unifying approach to both sample-based and population-based nonresponse weighting adjustment. Deville and Särndal (1992) introduced calibration estimation in the full response case and showed that many standard estimators, such as the *generalized regression estimator* (GREG) and poststratified estimator, are calibration estimators. (See also Bethlehem, 2002, Chapter 18.) Lundström and Särndal (1999) extended calibration estimators to encompass estimators that adjust for unit nonresponse. They define a calibration estimator as

$$\hat{y}_{\text{cal}} = \sum_r w_i^* y_i, \tag{6}$$

subject to the calibration equation $\sum_r w_i^* \mathbf{x_i} = \mathbf{X}$, where the sum is over the respondents, $w_i^*$ is the adjusted weight, $\mathbf{x_i}$ is a vector of auxiliary variables, and $\mathbf{X}$ is a vector of totals of those auxiliary variables. Since the weights are not uniquely defined by these conditions, further assume $w_i^* = w_i v_i$, where $w_i$ is the base weight. While nonlinear relationships between the weighting adjustment and the vector of auxiliary variables can be considered, they focus on the linear relationship, where the squared difference between the original weights and the calibrated weights is minimized. With full response, this linear calibration estimator is the GREG estimator. We take advantage of the ability to write many nonresponse weighting estimators as calibration estimators below.

The weighting class adjustment mentioned above is a calibration estimator and it can be either sample-based or population-based. The estimator is

$$\hat{y}_2 = \sum_r w_i v_{2i} y_i, \tag{7}$$

where $v_{2i} = x_c / \sum_r w_i \delta_i(c)$, and $\delta_i(c) = 1$ if $i$ is in cell $c$ and $= 0$ otherwise. For the sample-based version, $x_c = \sum_s w_i \delta_i(c)$, while for the population-based version $x_c = N_c$, where $N_c$ is the number of units in the population in cell $c$. Little and Vartivarian

(2003) argue that including the survey weights in the estimation of the propensities is either incorrect or unnecessary. For example, if weighting classes are used, they prefer $\hat{\phi}_c = n_c/r_c$, the unweighted ratio of the number sampled to the number of respondents in cell $c$ to $v_{2c} = \hat{\phi}_c = \sum_s w_i\delta_i(c)/\sum_r w_i\delta_i(c)$, the calibrated and weighted version. The weighted version clearly has some desirable properties, especially for estimating totals rather than means or proportions.

While the development of the weighting class adjustment was motivated here by response propensity modeling, the cells should be formed considering variables that are predictive of response and are correlated to the key statistics being produced, including domains. The bias of an estimated mean will be reduced if (1) the response propensities of the units within the cells are approximately equal ($\phi_i \approx \phi_c$ for $i \in c$), (2) the value of $y$ is approximately constant within cells ($y_i \approx \overline{y}_c$ for $i \in c$), or (3) the response propensities and $y$'s are approximately uncorrelated within cells ($\rho_{y_i,\phi_i} \approx 0$ for $i \in c$). The weighting class adjustment approach can be applied when important variables are quantitative by categorizing these variables. Forming three to five categories from a quantitative variable typically extracts most of the information from the variable.

Kalton (1983) and Särndal and Lundström (2005) discussed the methods for choosing variables that satisfy these conditions. When there are many possible auxiliary variables, some of the methods for choosing those for nonresponse weighting include the use of substantive experts who define important variables, logistic and other forms of regression modeling, and categorical search algorithms. Rizzo et al. (1996) examined many of these methods in some detail for nonresponse adjustment of weights in a later wave of a household panel survey. These methods of choosing variables are not specific to the weighting class adjustment approach and can be applied to other nonresponse weighting adjustments.

A limitation of the weighting class approach is that some variables may not be able to be fully utilized in creating the cells. This is especially problematic when many auxiliary variables are available, such as in surveys with rich sampling frames and in longitudinal surveys. Practitioners often require that cells have a minimum number of respondents and avoid large weighting adjustments that could increase the variances of the estimates. In this situation, the weighting class approach inhibits using all the auxiliary data effectively.

Raking and the two-way classification method are alternative calibration methods that support including as many auxiliary variables as needed by controlling only to marginal totals. Raking to marginal totals traces back to Deming and Stephan (1940) while the two-way classification is the linear estimator suggested by Särndal and Lundström (2005). Särndal and Lundström indicated that there is little difference in the weighting adjustment regardless of whether raking or the two-way classification is used. Brick and Jones (2008) provided further evidence of the similarity of these methods.

An advantage of the raking approach is that the adjustment factors are constrained so that no weights can be negative, although they can be less than unity. The adjustment factors based on the linear estimator may be negative, and this may be a more serious problem for users. It is possible to constrain the variation in the adjustment factors to avoid negative weights, but this adds some complexity to the computations. Kalton and Flores-Cervantes (2003) showed the effect of constraining the adjustment factors on the weights. Other practical issues that may be considered are the iterative computation of the raking estimator and that convergence in high-dimensional raking problems may be slow. The two-way classification requires inverting a matrix, and there may be difficulties with the inversion in high-dimensional cases.

Neither the raking nor the two-classification estimator admits a simple expression for the weights, but both can be written as

$$\hat{y}_3 = \sum_r w_i v_{3i} y_i, \tag{8}$$

where $v_{3i}$ is computed so that it satisfies the calibration equations defined by the marginal constraints.

As an example of raking, consider the case of two marginal constraints with population-based margins. The adjustment factor for all respondents in level $h$ of the first (row) variable and level $k$ of the second (column) variable can be written as $v_{3hk} \approx \hat{\alpha}_h \hat{\beta}_k$, where $\hat{\alpha}_h$ is the adjustment for level $h$ of the row variable and $\hat{\beta}_k$ is the adjustment for level $k$ of the column variable. The factors are the product of all the adjustments that are made to the specified row and column over the iterations, until the process converges. With more than two raking dimensions, the adjustment factor for a cell defined as the intersection of all the dimensions is still the product of the dimension adjustments. The only difference between the sample-based and the population-based versions of the raking process is the nature of the marginal totals, analogous to the difference in the weighting class adjustment. In the population-based version, the marginal totals are known universe counts (e.g., $N_{h+}$ and $N_{+k}$) while in the sample-based version, the marginal totals are computed for the full sample (e.g., $\hat{N}_{h+} = \sum_k \sum_{i \in r} w_{hki}$ and $\hat{N}_{+k} = \sum_h \sum_{i \in r} w_{hki}$).

The nonresponse bias of an estimated mean that has weights adjusted by either the raking or two-way classification method will be reduced under conditions similar to those of the weighting class approach. To better appreciate the similarity, we rewrite the conditions in terms of a main effects models (if $\phi_i \approx \phi_c$, then $\phi_i \approx \alpha_c$, for $i \in c$; if $y_i \approx \overline{y}_c$, then $y_i \approx \alpha'_c$, for $i \in c$, and if $\rho_{y_i,\phi_i} \approx 0$, then $\rho_{y_i-\alpha'_c,\phi_i-\alpha_c}=0$, for $i \in c$). Raking and two-way or multiway classification methods correspond to main effects models of higher dimensions. Thus, the conditions for approximately unbiased estimates, using only two auxiliary variables to simplify the presentation, are as follows: (1) the response propensities of the units can be expressed as $\phi_{hki} \approx \alpha_h \beta_k$ for $i \in hk$; (2) the value of $y$ is approximately given by $y_{hki} \approx \alpha'_h + \beta'_k$, for $i \in hk$; or (3) the response propensities and $y$'s are approximately uncorrelated after accounting for the main effects ($\rho_{y_i-\alpha'_h-\beta'_k,\phi_i-\alpha_h\beta_k} \approx 0$ for $i \in hk$). See Kalton and Maligalig (1991), Oh and Scheuren (1983), Holt and Elliot (1991), Särndal and Lundström (2005), and Brick and Jones (2008) for discussions related to these conditions.

Adding other main effects through use of the raking (or the multiway classification approach) permits more information than might be possible with the weighting class estimator. As the conditions above clearly show, nonresponse bias due to interactions is not accounted for in the weighting adjustments. With many auxiliary variables available, it is possible to create margins that are very extensive to reduce the potential for missing important interactions. For example, suppose weighting class adjustments by age, by sex, and by region are created. This classification could be used as one marginal constraint, and other auxiliaries like education by income added as another margin. The goal is to include important interactions between the auxiliaries, such as education and income, in the weighting while avoiding nonresponse biases that might occur if they are omitted.

There are potential dangers in making the marginal constraints too extensive and too numerous. The standard advice is to avoid controlling the weights in this process by creating too many levels and variables. Some of this advice is based on empirical experience where problems have been encountered. For example, Brick et al. (2003) described issues that arose in surveys when high-dimensional raking was attempted and the cross classification of the dimensions resulted in many cells with no observations. The applications raised questions about both the convergence of the weighting procedure and the properties of the adjustment factors in these settings. Current practice varies, but many practitioners impose larger minimum cell size requirements on the levels of the margins than for cells with the weighting class approach. More research into these issues is needed to guide practice.

## 5. Variance and confidence interval estimation

Using weights adjusted for nonresponse has implications for both point and interval estimation. The previous section focused on nonresponse weighting adjustment to reduce the bias of the estimates.

If the weight adjustment is successful in reducing the nonresponse bias, then the confidence interval coverage rates may be substantially improved. Cochran (1977) discussed the deleterious effect of bias in confidence interval coverage rates and showed that the relative bias (ratio of the bias of the estimate to its standard error) should be less than about 10% to give confidence intervals with coverage rates close to the nominal levels. Bias causes the intervals to be off-center, lowering the overall coverage rate and making the coverage very asymmetric. For example, if the bias in a mean is half the standard error of the mean, then instead of the nominal 95% coverage rate, the actual rate is about 92%, and almost all of the noncoverage is at the lower end of the interval (if the bias is positive). The substantial effect of bias on the confidence interval coverage rate (and the analogous effects on error rates in tests of hypotheses) is the reason why so much emphasis is placed on methods to reduce bias, even when those methods may increase the standard errors of the estimates.

Of course, unnecessary increases in the variability of the weights, where the variability in weights does not result in reductions in bias, should also be avoided. If the estimates of variance account for the nonresponse adjustment, then this type of variation in weights will cause confidence intervals to be wider than necessary and could lead to conservative intervals (covering at greater than the nominal level). On the other hand, if the variance estimation technique does not account for the nonresponse adjustment, then the variance estimates may be biased downward causing the coverage intervals to be too short and anticonservative. Many researchers have warned about the effect of weight adjustments on the variance of the estimates, but there is relatively little in the literature about the actual computation of variances with nonresponse-adjusted weights.

The two main methods of computing variances for complex sample surveys are replication methods and Taylor series linearization. These methods are also used when the weights have been adjusted for nonresponse. Below, we review these methods and the ability to account for the nonresponse adjustments with the methods.

The most frequently used replication methods are the jackknife, the balanced repeated replication, and the bootstrap. All of these methods involve the same general steps: (1) replicate samples are selected from the full sample (using different replicate sampling techniques for the different replication methods), (2) replicate estimates are computed from each replicate sample using the same procedures used in the full sample to compute the estimate of interest, (3) the sum of the squared deviations between the replicate estimates and the full sample estimate is computed, and (4) an estimate of the variance of the full sample estimate is computed by scaling the sum of the squares by the constants appropriate for the replication method. There are many potential variations in the specifics of these steps that produce variance estimates with slightly different asymptotic properties. For example, in step 3, the deviation can be from the average of the replicate estimates rather than the full sample estimate. See Wolter (2007) and Shao and Tu (1995) for details on replication for the full response case. Fuller (1998) gave a detailed discussion of replication for two-phase samples, where the second phase may be the uncontrolled unit nonresponse.

Rust and Rao (1996) described how replicate weights can be used to implement replication methods efficiently. Since the replicate samples are selected only once for any particular survey, replicate weights are attached to each respondent record and stored on the data file. These replicate weights can then be used to produce the replicate estimates for nearly all, or certainly the vast majority of, estimates of interest using the same estimation scheme used to create the full sample estimate, thus greatly simplifying the computational tasks associated with replication.

Creating replicate weights also helps facilitate the ability to include the effect of weight adjustments in the variance estimation process. For example, Yung and Rao (2000) showed that if poststratification weighting adjustments are performed separately and independently for each replicate, then the variance estimator will be consistent. Valliant (1993) obtained similar results. Although the theory for other weighting adjustments has not been presented formally, the inclusion of all or almost all steps of weighting adjustments in the calculation of replicate weights is now a commonly accepted practice. When the replicate weights are adjusted in the same manner as the full sample weights, no other information is needed for producing variance estimates for any statistic that can be written as a function of the sum of weighted estimates. This feature has enabled replication advocates to claim that this procedure “accounts” for all the weighting steps, including nonresponse adjustment. However, there are issues that must be handled appropriately to deal with nonresponse. Mantel et al. (2000) noted that when using the bootstrap to replicate nonresponse adjustments special precautions for some bootstrap replicates must be considered to avoid computations with no respondents in the denominator of the adjustments.

The Taylor series linearization approach for estimating variances of complex sample survey estimates is relatively simple. In most surveys, even designs with multiple stages of sampling and unequal probabilities of selection, an expression for the variance of a linear statistic is usually straightforward for with-replacement sampling methods. For nonlinear statistics, the same variance estimation procedures are used, but a linear substitute based on the first-order Taylor series approximation is used instead of the nonlinear statistic. In large multistage samples, the variances are approximated simply if the first-stage sampling units are selected with-replacement or if the proportion of units sampled is very small. The same conditions are required for replication methods to give consistent variance estimates. See Wolter (2007) and Binder (1983) for descriptions of the linearization method.

The poststratified estimator is essentially a ratio or regression estimate, so the linearization approach to deal with this type of weighting adjustment is to compute the variance for a linearized ratio or regression statistic. Any nonresponse weighting adjustment that can be written as a calibration estimator with the linear distance function can be approximated using this type of approach as discussed by Lundström and Särndal (1999). Raking adjustments are more complicated and difficult to handle with linearization methods. Linearization methods using residuals can be used, but there are operational issues. The main difficulty is that all the marginal control totals need to be included along with the sample data for variance estimation. One option that is used in practice approximates the variance of the raked estimate by treating one of the marginal controls as defining poststrata while ignoring the other marginal controls. Implementation still requires including the "poststratification" control totals for variance estimation.

Two commercially available software programs that compute linearized variance estimates from complex samples and support weight adjustments in some way are SUDAAN® and Stata®. Both of these packages require the control totals to be input to compute poststratified estimates. Other packages with the same types of requirements include Bascula (Nieuwenbroek and Boonstra, 2002), POULPE (Caron, 1998), and CLAN97 (Andersson and Nordberg, 1998). CALMAR2 (Sautory, 2003) supports controlling for several dimensions, but for a number of reasons including concerns about confidentiality and disclosure, the full set of calibration totals are rarely included with the survey data when linearization is used.

The difficulties associated with reflecting the full set of weight adjustments in the linearization method raises the important question of whether it matters on a practical level; there has been very little research on this topic. Valliant (2004) conducted a simulation study of the effect of weighting adjustments on variance estimates; he compared variance estimates and confidence intervals computed with both replication and linearization methods. He concluded that replication methods do account for the effects of weight adjustments on the variances of the estimates to a greater extent than linearization methods, but in a limited sense. In his study, replication methods with fully replicated weight adjustments tended to overestimate variances but the estimated confidence intervals had coverage rates close to the nominal level. The linearization estimators he reviewed did not fully account for the adjustments, and these gave underestimates of variances and confidence intervals that covered at lower than the nominal level. Valliant's results are consistent with expectation, but other simulation and empirical research would be beneficial.

## 6. Discussion

Nonresponse is a major concern in making inferences from sample surveys. Almost all surveys are subject to unit nonresponse; the trend toward lower response rates is being consistently observed across countries, modes of data collection, content areas, and sponsorships. The lower response rates are the result of decreases in both accessibility and amenability, but the specific reasons seem to differ greatly by survey. Efforts to increase response rates by using tactics such as more call attempts or incentives, have limited effectiveness. Surveys often find that even with these additional efforts, response rates are just declining more slowly than in comparable surveys that do not use these tactics.

The decline in response rates has heightened concern about the potential for nonresponse bias. A stochastic response model is often used to bridge the gap in the randomization model that exists due to the failure to observe responses from all sampled units. The stochastic model postulates positive response propensities for all sampled units, but these propensities must be estimated, resulting in theoretical and practical difficulties. The theoretical problem is that randomization requires that all probabilities are known. The practical problem is that estimating response propensities based on observing responses and nonresponses in one trial is a process that may be subject to substantial errors.

Nonresponse bias is the most serious effect of attaining response rates less than 100%. The precision of the estimates is also affected by nonresponse, but this loss in accuracy is easy to handle. In the past, response rates have been used to provide some sort of metric to guide users on the magnitude of nonresponse bias. More recently, it has been shown that response rates may not be very good measures of nonresponse bias. Within a survey, the nonresponse bias for different statistics may vary greatly, even though there is only one unit response rate for the survey. This has encouraged researchers to examine the relationships between the survey characteristics being estimated and the reasons for nonresponse. One component of this examination is modeling response propensities using data available from the sampling frame and from other sources. Another is modeling the statistic itself in terms of these same types of auxiliary variables. These models may be useful in predicting when nonresponse bias is likely to be significant and when it is likely to be negligible, an important consideration for both users and producers of sample surveys. They also provide information that can be used in forming weighting adjustments.

Many nonresponse weight adjustment procedures are calibration estimators. The totals used to calibrate the weights may be either sample-based (using auxiliary variables available for the sampled units) or population-based (using auxiliary variables from external sources). Standard estimators defined under full response, such as poststratification, raking, and GREGs, can be modified and used to deal with unit nonresponse. Both sample-based and population-based estimators may be effective in reducing nonresponse bias, but those based on population totals have added advantages in that they may reduce other sources of error, such as noncoverage bias and variance.

If reasonable models of response propensities and the key survey statistics can be formed from available auxiliary variables, then weighting adjustments derived from these models have the potential to reduce nonresponse bias. The key to bias reduction lies with the auxiliary data and the way the auxiliary data are used in the adjustments. Either implicitly or explicitly, calibration adjustments specify a model of the relationships between the response propensities and the auxiliary data, and the characteristics themselves and the auxiliary data. If the auxiliary data are effective in accounting for strong relationships, then the residuals between the response propensities and the estimates after fitting the models should have low correlations. If these correlations are low, then nonresponse bias in the estimates should be minimal. Unfortunately, this modeling is often difficult in practice. In some cases, the auxiliary data are limited or are not predictive of response propensities or the key survey statistics. However, an increasing awareness of the importance of collecting *paradata* (survey process data) may result in increased availability of auxiliary data that are predictive of response propensities.

One area of active research is focused on improving the modeling that underpins nonresponse weighting adjustments. The adoption of the stochastic model of nonresponse took many years to gain general acceptance in the survey community. As the model has become more accepted, new questions about its application in the survey setting are being posed. Some basic issues that still need to be addressed involve the nature of response propensities and whether the assumptions applied in observational studies are appropriate in the survey sampling context. Developments in these areas may result in changes in the ways weight adjustments are performed.

More research is needed on the implications of weight adjustments on the estimated variance and confidence intervals for survey statistics. Valliant (2004) provided an important first step in this area, but many practical and theoretical issues remain to be studied for both replication and linearized variance estimation techniques.

*Sample Surveys: Design, Methods and Applications, Vol. 29A*
ISSN: 0169-7161

DOI: 10.1016/S0169-7161(08)00009-6

# Statistical Data Editing

*Ton De Waal*

## 1. Introduction

Users of statistical information are nowadays demanding high-quality data on social, demographic, industrial, economic, financial, political, and cultural aspects of society with a great level of detail and produced within a short span of time. National statistical institutes (NSIs) fulfill a central role in providing such high-quality statistical information. Most NSIs face this challenge while their financial budgets are constantly diminishing.

A major complicating factor is that collected data generally contain errors. The data collection stage in particular is a potential source of errors. For instance, a respondent may give a wrong answer (intentionally or not), a respondent may not give an answer (either because he does not know the answer or because he does not want to answer this question), and errors can be introduced at the NSI when the data are transferred from the questionnaire to the computer system, etc. The occurrence of errors in the observed data makes it necessary to carry out an extensive process of checking the collected data, and, when necessary, correcting them. This checking and correction process is referred to as statistical data editing (SDE).

To check and correct data, two steps have to be carried out. First, the erroneous records and the erroneous fields in these records have to be localized. This is called the error localization step. Second, the localized erroneous fields and the missing fields have to be imputed, that is, the values of the erroneous fields have to be replaced by better, preferably the correct, values and the values of the missing fields have to be estimated. This is called the imputation step. The error localization step only determines which fields are considered erroneous; the imputation step determines values for these fields as well as for the missing ones.

Although the error localization step and the imputation step are closely related in theory, in practice, they are usually treated as two separate steps in the statistical process. In this chapter, we will also treat them as two distinct steps. We will use the phrase SDE in the sense of localizing errors, unless stated otherwise. Imputation of missing data is discussed in Chapter 10.

Traditionally, statistical agencies have always put much effort and resources into SDE, as they considered it a prerequisite for publishing accurate statistics. In traditional survey processing, SDE was mainly an interactive activity where all individual records were checked with the aim to correct all data in every detail. It has long been recognized, however, that it is not necessary to correct all data in every detail. Several studies (see, e.g., Granquist, 1984, 1997; Granquist and Kovar, 1997; Pannekoek and De Waal, 2005) and many years of practical experience at several NSIs have shown that in general it is not necessary to remove all errors from a data set to obtain reliable publication figures. The main products of statistical offices are tables containing aggregate data, which are often based on samples of the population. This implies that small errors in individual records are acceptable. First, because small random errors in individual records generally tend to cancel out, that is, their sum generally tends to be negligible in comparison to the corresponding publication figure. Second, because if the data are obtained from a sample of the population, there will always be a sampling error in the published figures, even when all collected data are completely correct. In this case, an error in the results caused by incorrect data is acceptable as long as it is small in comparison to the sampling error. To obtain data of sufficiently high-quality, it is usually sufficient to remove only the most influential errors.

In the past, and often even in the present, too much effort was spent on correcting errors that did not have a noticeable impact on the ultimately published figures. This has been referred to as "overediting." Overediting not only costs budget but also a considerable amount of time, making the period between data collection and publication unnecessarily long. Sometimes overediting even becomes "creative editing": the editing process is then pursued to such an extent that unlikely, but correct, data are " corrected," or discarded and replaced. Such unjustified alterations can be detrimental for data quality.

To improve the efficiency of the editing process, modern techniques such as selective editing, automatic editing and macroediting can be applied instead of the traditional microediting approach, where all records are extensively edited manually. In this chapter, we discuss these editing techniques. A crucial role in several (versions) of these techniques is often played by so-called edit rules. We describe the use of these edit rules in Section 2. We then continue our discussion with interactive editing in Section 3. We examine the possibility of editing during the data collection phase in Section 4. In the next three sections, we examine modern editing techniques: selective editing in Section 5, automatic editing in Section 6, and macroediting in Section 7. In Section 8, we discuss a strategy for SDE based on combining different editing techniques. Section 9 ends the chapter with a brief discussion of possible future developments with respect to SDE.

For more information on SDE in general, we refer to Ferguson (1994), and for SDE for business surveys to EDIMBUS (2007). An important international project on SDE and imputation was the EUREDIT project. The EUREDIT project aimed at improving the efficiency and the quality of automatic methods for SDE and imputation at NSIs. For the main findings of this project, we refer to EUREDIT Project (2004a,b). Several software packages for SDE have been developed. We refer to Chapter 13 for a discussion of these software packages.

## 2. The use of edit rules

At NSIs, edit rules, or edits for short, are often used to determine whether a record is consistent or not. An example of an edit is

$$T = P + C, \tag{1}$$

where $T$ is the turnover of an enterprise, $P$ its profit, and $C$ its costs. Edit (1) expresses that the profit and the costs of an enterprise should sum up to its turnover. Such an edit is referred to as a balance edit. Another example is $T \geq 0$, expressing that the turnover of an enterprise should be non-negative. Edits like these are referred to as non-negativity edits. A third example is $P/T \leq 0.5$. Such an edit expressing that the ratio of two variables should be less (or greater) than a certain threshold is referred to as a ratio edit. Examples of edits for categorical (discrete) data are that children of the head of household cannot be married to each other and that a person can have only one (biological) mother.

To construct a set of edits, one usually starts with the "hard" (or logical) edits, which hold true for all correctly observed records. Balance edits are usually hard edits. After the hard edits have been specified, one generally uses subject-matter knowledge and statistical analyses to add a number of "soft" edits, which hold true for a high fraction of correctly observed records but not necessarily for all of them. In many cases, ratio edits are soft edits. The thresholds of soft ratio edits have to be carefully determined, so correct records do not, or only very rarely, violate these edits, while the edits are powerful enough to pinpoint erroneous records. Another example of a soft edit is that a mother must be at least 15 years older than any of her children. In most cases where this edit is violated, the record under consideration is indeed incorrect. Only in extremely rare cases, this edit is violated by a correct record.

Records that are inconsistent with respect to the edits, that is fail one or more edits, are considered to contain errors if hard edits are violated and are considered to be suspicious if only soft edits are violated. The values in an erroneous record have to be modified in such a way that the resulting record is a better approximation of the true data of the corresponding respondent. Suspicious records are either examined further or are treated as erroneous records.

A consistent record, that is a record that satisfies all edits, is not necessarily (considered to be) error-free. For instance, (some of) the values in a consistent record may be outliers with respect to the bulk of the data. Such outlying values are often considered suspicious and are, hence, checked in the editing process, even if all edits are satisfied.

To avoid overediting, one should in particular be careful not to specify too many soft edits. In general, users tend to apply more soft edits than necessary to the data (see Di Zio et al., 2005a).

## 3. Interactive editing

The use of computers in the editing process started many years ago. In the early years, their role was restricted to checking which edits were violated. For each record, all violated edits were listed. Subject-matter specialists then used these lists to correct the records. That is, they retrieved all paper questionnaires that did not pass the edits and corrected these questionnaires, for instance by recontacting the respondent or by comparing the respondent's data to data from similar respondents. After they had corrected the data, these data were again entered into the computer, and the computer again checked whether the data satisfied all edits. This iterative process continued until (nearly) all records passed the edits.

A major problem with respect to this approach was that during the manual correction process, the records were not checked for consistency. As a result, a record that was "corrected" could still fail one or more specified edits. Such a record hence required more correction. It was not exceptional that some records had to be corrected several times. It is therefore not surprising that editing in this way was very costly, both in terms of budget as well as in terms of time (see, e.g., Federal Committee on Statistical Methodology, 1990; Granquist and Kovar, 1997).

Subject-matter specialists have extensive knowledge with respect to their area of expertise. This knowledge should be used as well as possible. This aim can be achieved by providing subject-matter specialists with efficient and effective data editing tools. Survey-processing systems such as Blaise and CSPro (see Chapter 13 for a discussion of such software systems) are often used to edit data at NSIs. When such systems are used, the specified edits can be checked during or after data entry, and, if necessary, the data may immediately be corrected. This is referred to as interactive or computer-assisted editing. The introduction of systems such as Blaise and CSPro led to a substantial efficiency improvement of the editing process. In this section and in the next, we use Blaise as an example of a survey-processing system due to the wide use of the system.

When Blaise is used to edit the data, it is no longer necessary to edit the data in several iterations, each consisting of a checking phase and a correction phase. When data are corrected, new error signals due to failed edits, if any, are immediately shown on the computer screen. The error signals, in combination with the data themselves, direct the subject-matter specialist to potential errors in the data. For instance, Blaise can calculate the number of times each field is involved in a failed edit. Fields that are most often involved in failed edits are usually the most likely ones to be in error. To correct data, the subject-matter specialists often check the paper questionnaires or scanned versions thereof as this can help them to identify errors in the data.

Data from paper questionnaires can be entered either by fast data entry personnel or by subject-matter specialists. In the former case, the data are keyed in without attempting to edit them at this stage. Later, subject-matter specialists edit the keyed raw data. Alternatively, data can directly be entered by subject-matter specialists. This costs more time than letting the data be keyed in by data entry personnel. However, subject-matter specialists can enter and correct data at the same time. The extra time required to enter the data is often (more than) compensated for by the fact that each record is treated, that is entered or edited, only once. A practical drawback of keying in data and editing them at the same time is that the raw, unedited, data are not available for later analyses, for instance, analyses with respect to the efficiency and effectiveness of the editing process itself.

An alternative to keying in data is scanning the paper questionnaires in combination with optical character recognition. For paper questionnaires for which the answers mainly consist of numerical data, this often leads to data of similar quality as keyed-in data. Paper questionnaires for which optical character recognition does not give good results are often scanned anyway to help the subject-matters specialists during the interactive editing process.

Interactive editing can be used to edit both categorical and numerical data, and it is nowadays a standard way to edit data. The number of variables, edits, and records may, in principle, be high. Survey managers generally consider data edited in an interactive manner to be of high statistical quality. For more on interactive editing by means of systems like Blaise, we refer to Pierzchala (1990).

The fundamental problem of interactive editing is that, even though each record has to be edited only once, still all records have to be edited. We have already mentioned that this can, and often does, lead to overediting. Instead of editing all records, one could consider editing only the ones with influential errors. This is referred to as selective editing and is discussed in Section 5. In Section 4, we first discuss the most efficient editing technique of all: no editing at all, but instead ensuring that correct data is obtained during the data collection phase.

## 4. Editing during the data collection phase

Blaise not only applies edits but also so-called routing rules. Frequently, different questions are posed to different kinds of respondents. For instance, it is not useful to ask a male respondent whether he has ever been pregnant as the answer to this question would not provide any additional information. Blaise ensures that each respondent is asked the questions that are applicable to this kind of respondent. Owing to this functionality, Blaise is an excellent system for CAPI (computer-assisted personal interviewing), CATI (computer-assisted telephone interviewing), CASI (computer-assisted self interviewing), and CAWI (computer-assisted web interviewing).

When CAPI is used to collect the data, an interviewer visits the respondent and enters the answers directly into a laptop. When CATI is used to collect the data, the interview is carried out during a telephone call. When CASI or CAWI is used to collect the data, the respondent fills in an electronic questionnaire himself. The difference between these two modes is that for CAWI, an electronic questionnaire on the internet has to be filled in, whereas for CASI, an off-line electronic questionnaire has to be filled in. When an invalid value for a question is given or an inconsistency between the answers of two or more questions is noted during any of these data collection modes, this is immediately reported by Blaise. The error can then be resolved by asking the respondent these questions again. For CASI and CAWI, generally not all edits that could be specified are actually specified since the respondent might get annoyed and may refuse to complete the questionnaire when the edits keep on reporting that the answers are inconsistent.

In many cases, data collected by means of CAPI, CATI, CASI, or CAWI contain fewer errors than data collected by means of paper questionnaires as random errors that affect paper questionnaires can be detected and avoided at collection. For face-to-face interviewing, CAPI has in fact become the standard. CAPI, CATI, CASI, and CAWI may hence seem to be ideal ways to collect data, but, unfortunately, they too have their disadvantages.

A first disadvantage of CATI and CAPI is that CATI and, especially, CAPI are very expensive. A second disadvantage of CATI and CAPI is that a prerequisite for these two data collection modes is that the respondent is able to answer the questions during the interview. For a survey on persons and households, this is often the case. The respondent often knows (good proxies of) the answers to the questions or is able to retrieve the answers quickly. For a survey on enterprises, the situation is quite different. Often, it is impossible to retrieve the correct answers quickly, and often the answers are not even known by one person or one department of an enterprise. Finally, even in the exceptional case that one person knew all answers to the questions, the NSI would generally not know the identity of this person. For the above-mentioned reasons, many NSIs frequently use

CAPI and CATI to collect data on persons and households but only rarely for data on enterprises.

Pilot studies and actual applications have revealed that CASI and CAWI are indeed viable data collection modes, but also that several problems arise when these modes are used. Besides IT problems, such as that the software, and the internet, should be fast and reliable and the security of the transmitted data should be guaranteed, there are many practical and statistical problems. We have already mentioned the practical problem that if the edits keep on reporting that the answers are inconsistent, the respondent may get annoyed and may refuse to fill in the rest of the questionnaire. An example of a statistical problem is that the group of people responding to a web survey may be selective (see, e.g., Bethlehem, 2007). Another important problem for CAWI and CASI is that data collected by either of these data collection modes may appear to be of higher statistical quality than data collected by means of paper questionnaires, but in fact are not. When data are collected by means of CASI and CAWI, one can enforce that the respondents supply data that satisfy build-in edits or one can avoid balance edits by automatically calculating the totals from their components. As less edits are failed by the collected data, the collected data may appear to be of higher statistical quality. This may not be the case, however, as respondents can be less accurate when filling in the entries in an electronic questionnaire, especially if totals are computed automatically (see Børke, 2008; Hoogland and Smit, 2008).

NSIs seem to be moving toward the use of mixed-mode data collection, where data are collected by a mix of several data collection modes. This obviously has consequences for SDE. Some of the potential consequences have been examined by Børke (2008), Hoogland and Smit (2008), and Van der Loo (2008). For more information on computer-assisted data collection in general, we refer to Couper et al. (1998).

## 5. Selective editing

### *5.1. Introduction to selective editing*

Selective (or significance) editing (see, e.g., Farwell and Raine, 2000; Hedlin, 2003; Hidiroglou and Berthelot, 1986; Hoogland, 2002; Latouche and Berthelot, 1992; Lawrence and McDavitt, 1994; Lawrence and McKenzie, 2000) is an umbrella term for several methods for identifying the influential errors in a data set, that is, the errors that have a substantial impact on the publication figures. The aim of selective editing is to split the data into two streams: a critical and a noncritical stream. The critical stream consists of records that are the most likely ones to contain influential errors; the noncritical stream consists of records that are unlikely to contain influential errors. Only the records in the critical stream are edited interactively. The records in the noncritical stream are either not edited or edited automatically (see Section 6).

The scope of most techniques for selective editing is limited to (numerical) business data. In these data, some respondents can be more important than other respondents, simply because the magnitude of their contributions is higher. Social data are usually count data where respondents contribute more or less the same, namely their raising weight, to estimated population totals. In social data, it is therefore difficult to differentiate between respondents. Selective editing has gradually become a popular method for editing business data and increasingly more NSIs use selective editing techniques.

Many selective editing methods are relatively simple ad hoc methods based on common sense, although also complicated outlier detection techniques have been used in the context of selective editing (see Di Zio et al., 2008). The most often applied basic idea is to use a score function (see, e.g., Hidiroglou and Berthelot, 1986; Van de Pol and Molenaar, 1995). We distinguish two important components to construct a score function: the influence component and the risk component. The influence component measures the relative influence of a record on a publication figure. The risk component usually measures the deviation of the observed values from "anticipated" values. How to define suitable anticipated values depends on the specific data set. For a cross-sectional survey, one could, for instance, use means or medians in certain groups of records. For longitudinal surveys, one could, for instance, use values from a previous period, possibly multiplied by an estimated trend. For some variables, anticipated values may be obtained from available register data.

A score function for an entire record is referred to as a global score function. Such a global score function is often based on local score functions. A local score function is a score function for a single variable within a record. It is usually defined as a distance between observed and anticipated values of a variable $y$ in the record under consideration, taking the influence of this record into account.

An example of a local score function is

$$w_i|y_i - \hat{y}_i|, \tag{2}$$

where $y_i$ denotes the observed value of variable $y$ in record $i$, $\hat{y}_i$ the corresponding anticipated value, and $w_i$ the raising weight of record $i$. This local score function can be considered as the product of a risk component, $|y_i - \hat{y}_i|/\hat{y}_i$, which measures the relative deviation of the observed value to the anticipated value, and an influence component, $w_i\hat{y}_i$, which measures the anticipated impact on the publication figure.

A global score function combines the local scores to a measure on the record level, so one can decide whether to edit the record in an interactive manner or not. Local scores can be combined into a global score by, for instance, taking a (weighted) sum of the local scores or by taking the maximum of the local scores (see Subsection 5.2). A record is considered suspicious if the value of the global score function exceeds a certain cutoff value (see Subsection 5.3).

The local score function (2) is suited for simple estimators for population totals. In principle, one can also develop local score functions for more complex estimators than simple estimators for totals. This can be done by linearization of the estimators, that is, by taking the first-order Taylor series. For more details, we refer to Lawrence and McKenzie (2000).

### 5.2. *Combining local scores into a global score*

When combining several local scores into a global score, one, first of all, needs to take into account that different variables may have a different order of magnitude or may be measured in different units. This problem can be overcome by scaling the variables. There are several options to scale variables, such as dividing the observed value by the mean value, by the standard error, or by the mean squared error of the variable under consideration (see Lawrence and McKenzie, 2000). From now on, whenever we refer to a local score, we will in fact mean the scaled local score.

The currently most general approach to combine local scores into a global score seems to be the use of the so-called Minkowski metric (see Hedlin, 2008). In our case, the Minkowski metric is given by

$$\mathrm{GS}_r(\mathbf{LS_r}, \alpha) = \left(\sum_{i=1}^{n} \mathrm{LS}_{r,i}^{\alpha}\right)^{1/\alpha}, \tag{3}$$

where $\mathrm{GS}_r$ denotes the global score for a record $r$, $\mathrm{LS}_{r,i} \geq 0$ the local score of the $i$th variable, $n$ the total number of variables, $\alpha > 0$ a parameter, and $\mathbf{LS_r} = \left(LS_{r,1}, \ldots, LS_{r,n}\right)$. The choice of $\alpha$ in (3) determines how the local scores are actually combined into a global score. The influence of large local scores on the global score increases with $\alpha$. For $\alpha = 1$, the global score is simply given by the sum of the local scores, and for $\alpha = 2$, the global score is the well-known Euclidean metric. For the limit where $\alpha$ goes to infinity, the global score $\mathrm{GS}_r(\mathbf{LS_r}, \infty)$ becomes $\max_i \mathrm{LS}_{r,i}$, that is, the maximum of the $n$ local scores.

The cutoff threshold value above which a record is considered to need interactive editing depends on the value of the $\alpha$ parameter and on the data set to be edited. Setting the cutoff threshold value is examined in Subsection 5.3.

The advantage of taking the maximum value of the local scores as the global score is that one is ensured that no influential error on any of the involved variables will slip through. Attempting to avoid the occurrence of influential errors in any of the involved variables may have the drawback that one has to edit many records interactively. Hedlin (2003), however, argues by using a model for the occurrence of influential errors in the data that this may not be the case. It depends on the data set to be edited how valid the assumptions underlying this model are.

The Minkowski metric is a flexible function that encompasses many well-known metrics used for selective editing purposes. By choosing the $\alpha$ parameter, one can basically select a metric varying from taking the sum of all local scores to taking the maximum of the local scores. However, more complex metrics cannot be selected. Such a more complex metric may be deemed necessary if there are many variables with many complex interactions between them (see also Subsection 5.5). Presently, no good generally applicable technique for combining local scores into a global score using such complex metrics seems available.

### *5.3. Determining cutoff thresholds*

After a method to determine the global score has been decided upon, a cutoff threshold should, in principle, be fixed. All records with a global score above the cutoff threshold are selected for interactive editing, whereas the records with a global score below the cutoff threshold are not edited interactively.

The most common approach to set the cutoff threshold is to perform a simulation study, using a raw (unedited) data set and the corresponding clean (edited) version. These data sets are generally from a previous period. The simulation study consists of calculating the global scores of the records in the raw data set and prioritizing the records in order of these scores. For several percentages $p$, one then simulates that the first $p\%$ of these records are edited and the remaining records are not. This is done by replacing the first $p\%$ of the records in the prioritized raw data set with the corresponding records in the clean version of the data.

A natural criterion to determine the quality of a selective editing procedure in the simulation study approach is the absolute pseudobias (see Latouche and Berthelot, 1992), which measures the absolute deviation between the raw value and the clean value. The difference between the raw value and the clean value is called the pseudobias rather than the bias as one cannot be sure that the clean value is indeed the correct value. For the records selected for interactive editing, the corresponding absolute pseudobias is zero. Based on the simulation study, the cutoff threshold is selected so that the sum of the absolute pseudobias is acceptably low compared with other errors in the data, such as the sampling error and the coverage error.

As Lawrence and McKenzie (2000) argue, the simulation study approach is also a way to check the effectiveness of the edits and the editing process itself. The simulation study allows one, for instance, to check if the records with high global scores indeed contain influential errors.

In some cases, the simulation study approach may not be applicable, for instance, because data from a previous period are not available. Lawrence and McKenzie (2000) suggest using a model for the editing process to determine the cutoff threshold in such cases. Given that model, one can then estimate the bias due to not editing some records as a function of the cutoff threshold. By specifying a maximum for the estimated bias, the corresponding cutoff threshold can be determined.

Lawrence and McKenzie (2000) use a relatively simple model for the editing process, but note that it can be extended to more complicated cases. An obvious drawback of the model-based approach is that is dependent on the model assumptions. Lawrence and McKenzie (2000) therefore propose to use the model-based approach only at the beginning of a new editing process to find a first cutoff threshold and later use the results of the current editing process to improve this threshold by means of a simulation study.

In practice, one sometimes does not fix a cutoff threshold before selective editing but instead only uses the global scores to prioritize the records. One then edits the records in order of priority until budget or time constraints tell one to stop.

### *5.4. The edit-related approach*

Hedlin (2003) proposes an edit-related approach to selective editing rather than the above sketched approach, which he refers to as estimate-related. The underlying idea of the edit-related approach is that influential errors will lead to violated edits. In the edit-related approach, one measures how many edits are failed by a record and by how much they fail. For each edit, the amount of failure is measured in some way. For a balance edit, one can, for instance, measure the amount of failure as the absolute difference between the observed total and the sum of its observed components. The amount of failure may be of very different orders for different (types of) edits. Hedlin (2003) therefore proposes using the Mahalanobis distance to combine amounts of failure into a global score per record.

The edit-related approach has the advantage that it does not focus on a single target variable. It also has the advantage that, unlike the estimate-related approach, it can be applied to categorical data. The edit-related approach has the drawback that it is dependent on the specified edits. In a study, Hedlin (2003) found that the estimate-related approach performed better than the edit-related approach.

Hybrid approaches where an estimate-related approach is combined with an edit-related approach are also possible. For instance, Hoogland (2002) discusses such a hybrid approach that uses anticipated values to estimate the risk of a record and at the same time takes the violations of the edits as well as the number of missing values of that record into account.

### *5.5. Experimental approaches*

At Statistics Netherlands, some more advanced, experimental, approaches have been examined that to some extent try to capture complex interactions between different variables. One such approach is based on logistic regression. In this logistic regression approach, one tries to estimate either the probability that a record contains influential errors or the probability that a specific variable in a record contains an influential error. In both cases, records or variables that are likely to contain influential errors need interactive editing.

We describe the case where we aim to estimate the probability $\pi$ that a specific variable contains an influential error. For this, we need a training data set consisting of both the raw (i.e., unedited) data and the clean (i.e., edited) data from a previous period. To each record in the unedited data set, we assign a probability $\pi$ that this record contains an influential error. The assigned probability is high for records for which the edited version differs much from the raw version, and is low for records for which the edited version is close to the raw version. Based on the training data, we then fit a logistic model defined by

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p, \tag{4}$$

where the $x_1$ to $x_p$ are predictor variables and $\beta_0$ to $\beta_p$ the model parameters.

For a new data set that needs to be edited, one then uses the model given by (4) with model parameters estimated using the training data. For each record in the data set to be edited, we hereby obtain an estimate $\hat{\pi}$ that the variable under consideration contains an influential error. If, for a certain record, $w\hat{\pi}$ is above a certain threshold value, where $w$ is the raising weight of this record, the variable under consideration is considered to require to interactive editing for this record.

For the record-level case, one can construct a model similar to (4), with the main difference that $\pi$ denotes the probability that a record with attributes $x_1$ to $x_p$ contains an influential error, and hence requires interactive editing.

Another approach that has been studied at Statistics Netherlands is the use of classification and regression trees (see Breiman et al., 1984) for selective editing. The idea of this approach is to grow a classification or regression tree that predicts the occurrence of an influential error in a record or a specific variable.

In general, a tree-based model classifies the data in terms of the values of a set of categorical predictor variables. A tree-based model is a binary tree that is generated by successively splitting a training data set into smaller subsets. These subsets are increasingly more homogeneous with respect to a selected response variable. This response variable may be either categorical or numerical. The process of recursively splitting the data set into two subsets continues until a stopping criterion is met. The terminal nodes in this tree form homogeneous clusters.

Homogeneity of a cluster can be measured in many ways. If the response variable is categorical, homogeneity may, for instance, be measured by the so-called Gini index (see Breiman et al., 1984). If the response variable is numerical, homogeneity may, for instance, be measured by ordinary least squares.

In the context of selective editing, there are several options for selecting the response variable. First, one has to choose whether one wants to generate a tree, either a classification tree or a regression tree, for a single variable or an entire record. Second, one has to choose between generating a classification tree or a regression tree. If one constructs a categorical response variable that expresses whether one considers a record or variable to need interactive editing or not, a classification tree has to be generated. If one constructs a numerical response variable that expresses the magnitude and impact on the publication figures of the errors in a record or variable, a regression tree has to be generated. By combining the possibilities, one obtains four different options, namely a classification tree for a single variable, respectively for an entire record, and a regression tree for a single variable, respectively for an entire record.

To generate a classification or regression tree, one again needs a training data set. In our case, the training data set consists of a raw (unedited) data set from a previous period, together with information from the editing process applied to this data set. In the cases where the aim is to generate a classification tree, we use the information whether a record is considered to require interactive editing, based on the changes that were made to this record during the editing process, as our response variable. In the cases where the aim is to generate a regression tree, we use the magnitude and impact on the publication figures of the changes that were made to this record during the editing process as our response variable.

After the generation of a tree, and hence generation of classification rules for constructing homogeneous clusters of records, a data set to be edited is supplied to the tree. The tree is then used to decide whether a variable or record needs to be edited interactively (in the case of a classification tree) or to estimate the magnitude and impact of the error in a single variable or an entire record on the publication figures (in the case of a regression tree).

At Statistics Netherlands, Van Langen (2002) and Sanders (2002) have carried out limited evaluation studies for the logistic regression approach and the tree-based approach, respectively. They both used a single data set of the Dutch Structural Business Statistics on the Construction Industry of which four versions were available: raw and clean versions for use as training data and other raw and clean versions for use as evaluation data. The simulation studies showed that in most cases, the logistic regression approach and the tree-based approach performed worse than a traditional approach based on an estimate-related global score. An exception was the approach based on a regression tree for an entire record. This approach turned out to be slightly more powerful than a traditional approach based on an estimate-related global score. However, given the complexity of the method and the low transparency of the decision rules generated, the approach based on generating a regression tree for an entire record has thus far not been implemented in editing processes at Statistics Netherlands.

## 6. Automatic editing

### *6.1. Introduction to automatic editing*

When automatic editing is applied, records are edited by computer without human intervention. In that sense, automatic editing is the opposite of the traditional approach to the editing problem, where each record is edited manually. Automatic editing can be applied to both categorical and numerical data. To automate the SDE process both the error localization step and the imputation step have to be automated. In this section, we focus on discussing the former step.

We can distinguish two kinds of errors: systematic error and random error. A systematic error is an error reported consistently by (some of) the respondents. It can be caused by the consistent misunderstanding of a question by (some of) the respondents. Examples are when gross values are reported instead of net values and particularly when values are reported in units instead of, for instance, the requested thousands of units (so-called "thousand-errors"). Random errors are not caused by a systematic deficiency but by accident. An example is an observed value, where a respondent by mistake typed in a digit too many.

Systematic errors, such as thousand-errors, can often be detected by comparing a respondent's present values with those from previous years, by comparing the responses to questionnaire variables with values of register variables, or by using subject-matter knowledge. Other systematic errors, such as transpositions of returns and costs and redundant minus signs, can be detected and corrected by systematically exploring all possible transpositions and inclusions/omissions of minus signs. Rounding errors—a class of systematic errors where balance edits are violated because the values of the involved variables have been rounded—can be detected by testing whether failed balance edits can be satisfied by slightly changing the values of the involved variables. Once detected, a systematic error is often simple to correct. We treat systematic errors in more detail in Subsection 6.2.

Generally speaking, we can subdivide the methods for automatic error localization of random errors into methods based on statistical models, methods based on deterministic checking rules, and methods based on solving a mathematical optimization problem. Methods based on statistical models, such as outlier detection techniques (see Chapter 11 of this book for more on outlier detection and treatment) and neural networks (see Nordbotten, 1995, for one of the first attempts to apply neural networks in the context of SDE), are extensively discussed in the literature. We therefore do not discuss these techniques in this chapter.

Deterministic checking rules state which variables are considered erroneous when the edits in a certain record are violated. An example of such a rule is if component variables do not sum up to the corresponding total variable, the total variable is considered to be erroneous. Advantages of this approach are its transparency and its simplicity. A drawback of this approach is that many detailed checking rules have to be specified, which can be time and resources consuming to do. Another drawback is that maintaining and checking the validity of a high number of detailed checking rules can be complex. Moreover, in some cases, it may be impossible to develop deterministic checking rules that are powerful enough to identify errors in a reliable manner. A final disadvantage is the fact that bias may be introduced as one aims to correct random errors in a systematic manner.

The automatic error localization problem for random errors can be formulated as a mathematical optimization problem in several ways. Freund and Hartley (1967) were among the first to propose such a formulation. It is based on minimizing the sum of the distance between the observed data and the "corrected" data and a measure for the violation of the edits. After "correction," some of the edits may still be failed. A second formulation, based on minimizing a quadratic function measuring the distance between the observed data and the "corrected" data subject to the constraint that the "corrected" data satisfy all edits, has later been proposed by Casado Valero et al. (1996).

A third approach for the automatic error localization problem for random errors is based on first imputing missing values and potentially erroneous values for an inconsistent record by means of hot-deck donor imputation (see Chapter 10), using a number of donor records. Subsequently, an imputed record that satisfies all edits and that is "closest" to the original record according to some distance function is selected. The values in the original record that differ from the corresponding values in the selected imputed record are considered to be erroneous. This paradigm forms the basis of nearest neighbor imputation methodology (NIM; see Bankier et al., 2000). Since the hot-deck donor imputation approach underlying NIM is much more suited for social data than for economic data, NIM has thus far mainly been used for demographic data. In some cases, NIM has been used in combination with methodology based on the Fellegi–Holt paradigm, which is discussed below (see Manzari, 2004).

The most often used approach for the automatic error localization problem for random errors is based on the paradigm of Fellegi and Holt (see Fellegi and Holt, 1976). This paradigm is, in fact, only one of three principles for automatic edit and imputation proposed by Fellegi and Holt in 1976. These three principles are as follows:

(1) the data in each record should be made to satisfy all edits by changing the fewest possible items of data (fields);
(2) as far as possible, the frequency structure of the data file should be maintained;
(3) imputation rules should be derived from the corresponding edit rules without explicit specification.

In the context of error localization, the first one of these principles is referred to as the "Fellegi–Holt paradigm." With regards to error localization, it is the most important principle of the three. The other two principles relate to imputation after errors have been localized. In their second principle, which was originally formulated in the context of categorical data only, Fellegi and Holt basically note that imputation should result in the preservation of the distribution of the true data, and in their third principle, that error localization and imputation should be applied in combination and not as completely separate processes.

In due course, the Fellegi–Holt paradigm has been generalized to: the data of a record should be made to satisfy all edits by changing the values of the variables with the smallest possible sum of reliability weights. That is, for each record $(x_1, \ldots, x_n)$, we wish to ensure the existence of, and often want to find, a synthetic record $\left(\hat{x}_1, \ldots, \hat{x}_n\right)$ such that $\left(\hat{x}_1, \ldots, \hat{x}_n\right)$ satisfies all edits, and

$$\sum_{k=1}^{n} w_k \delta\left(x_k, \hat{x}_k\right) \tag{5}$$

is minimized, where $\delta\left(x_k, \hat{x}_k\right)$ equals 1 if $x_k$ is missing or differs from $\hat{x}_k$ and 0 otherwise, and $w_k \geq 0$ is the so-called reliability weight of the $k$th variable ($k = 1, \ldots, n$). A reliability weight of a variable expresses how reliable one considers the values of this variable to be. A high reliability weight corresponds to a variable of which the values are considered trustworthy, and a low reliability weight corresponds to a variable of which the values are considered not so trustworthy. The (generalized) Fellegi–Holt paradigm can be applied to numerical data as well as to categorical data.

A variable $k$ ($k = 1, \ldots, n$), for which $x_k$ is missing or differs from $\hat{x}_k$, is considered to be erroneous. Such a variable has to be imputed later using a suitable imputation method (see Chapter 10). The existence of a synthetic record satisfying all edits ensures that the variables considered erroneous can indeed be imputed consistently, that is, such that all edits can be satisfied. The synthetic record is generally not used as the "corrected" record.

For business surveys, an overview of algorithms for solving the error localization problem based on the Fellegi–Holt paradigm has been given in De Waal and Coutinho (2005). Algorithms for categorical data have been proposed by Fellegi and Holt (1976), Garfinkel et al. (1986), Winkler (1998), Bruni et al. (2001), and Bruni and Sassano (2001). The algorithms described in the first two articles have been examined in detail by Boskovitz (2008). Algorithms for solving the error localization problem in a mix of categorical and continuous data have been proposed by Sande (1978), Schaffer (1987), De Waal (2003a,b, 2005), and De Waal and Quere (2003). This latter algorithm is sketched and illustrated in Subsections 6.3 and 6.4.

Assuming that only few errors are made, the Fellegi–Holt paradigm obviously is a sensible one. Provided that the set of edits used is sufficiently powerful, application of this paradigm generally results in data of higher statistical quality, especially when used in combination with other editing techniques. This is confirmed by various evaluation studies such as Hoogland and Van der Pijll (2003).

A drawback of using the Fellegi–Holt paradigm is that the class of errors that can safely be treated is limited to random errors. A second drawback is that the class of edits that can be handled is restricted to "hard" edits. "Soft" edits cannot be handled as such, and, if specified, are treated as hard edits. Especially, in the case of automatic editing, one should be careful not to specify too many soft edits to avoid overediting (see Di Zio et al., 2005a).

### *6.2. Approaches to automatic editing of systematic errors*

As already mentioned, a well-known class of systematic errors consists of so-called thousand-errors. These are cases where a respondent replied in units rather than in the requested thousands of units. The usual way to detect such errors is by considering "anticipated" values, which could, for instance, be values of the same variable from a previous period or values available from a register. One then calculates the ratio of the observed value to the anticipated one. If this ratio is higher than a certain threshold value, say 300, it is assumed that the observed value is 1000 times too large. The observed value is then corrected by dividing it by 1000. A minor practical problem occurs when the anticipated value equals zero. Usually, this problem can easily be solved in practice.

Al-Hamad et al. (2008) note more important problems with this standard procedure. The main problem they note is that the anticipated value itself has to be of sufficiently high-quality. If this value is incorrect, the procedure may not detect a thousand-error in the observed value if it is present.

They propose an alternative procedure, which simply consists of comparing the number of digits of the observed value to the anticipated value. In this way, thousand-errors (and larger errors) may be identified as those records for which the difference between the number of digits of the observed and anticipated value is 3 or more. In a study, they found that this alternative rule slightly outperformed the standard rule based on taking ratios.

A more complex approach for detecting and correcting thousand-errors, or more generally unity measure errors, that is, any error due to the erroneous choice by some respondents of the unity measure in reporting the amount of a certain variable, has been proposed by Di Zio et al. (2005b). That approach uses model-based cluster analysis to pinpoint various kinds of unity measure errors. The model applied consists of a finite mixture of multivariate normal distributions.

A second kind of systematic error that can relatively easily be corrected occurs when a respondent adds a minus sign to a value that is subtracted. The questionnaire of the Dutch Structural Business Survey (SBS) contains a number of combinations of items where costs have to be subtracted from returns to obtain a balance. If a respondent adds a minus sign to the reported costs, the value becomes wrongfully negative after data processing. Such an error where a respondent by mistake adds or deletes a minus sign is called a sign error. An obvious way to correct a sign error is by taking the absolute value of the reported value.

The situation becomes more complicated when a respondent may also have interchanged returns and costs on the questionnaire. Scholtus (2008a,b) examines this situation. Part of the Dutch SBS is the so-called results block. In this block of related questions, a respondent has to fill in a number of balance amounts. We denote the balance variables by $x_0, x_1, \ldots, x_{n-1}$. The so-called pretax result is denoted by $x_n$ and equals the sum of $x_0$ to $x_{n-1}$, that is,

$$x_0 + x_1 + \cdots + x_{n-1} = x_n. \tag{6}$$

Some of these balance variables are equal to the difference between a returns variable and a costs variable. That is,

$$x_{k,r} - x_{k,c} = x_k, \tag{7}$$

where $x_{k,r}$ denotes the returns variable and $x_{k,c}$ the corresponding costs variable of the $k$th balance restriction.

We give a simple example of sign errors and interchanged returns and costs. To this end, we consider a record with the following values: $x_{0,r} = 3,250$, $x_{0,c} = 3,550$, $x_0 = 300$, $x_{1,r} = 110$, $x_{1,c} = 10$, $x_1 = 100$, $x_{2,r} = 50$, $x_{2,c} = 90$, $x_2 = 40$, $x_{3,r} = 30$, $x_{3,c} = 10$, $x_3 = 20$, and $x_4 = -140$. This record has to satisfy (6) with $n = 4$ and (7) for $k = 0, 1, 2, 3$. The record can be made to satisfy all edits by changing the value of $x_0$ from 300 to $-300$ and interchanging the values of $x_{2,r}$ and $x_{2,c}$. These are likely to be the correct values as this is the only way to make the record satisfy all edits by means of such simple and natural modifications.

Assuming that if an inconsistent record can be made to satisfy all balance edits (6) and (7) by adding/deleting minus signs and interchanging returns and costs, this is indeed the way the record should be corrected; Scholtus (2008a,b) provides a formulation for correcting a record as a binary linear programming problem. Well-known operations research techniques can be applied to find a solution to this problem.

Assuming that the variables $x_{0,r}$ and $x_{0,c}$, which incidentally are the so-called operating returns and operating costs, respectively, in the case of the Dutch SBS, are not interchanged, Scholtus (2008b) proves that if a solution is found, it is the unique solution under some mild additional conditions.

Balance edits are often violated by the smallest possible difference. That is, the absolute difference between the total and the sum of its components is equal to 1 or 2. Such inconsistencies are often caused by rounding. An example is when the terms of the balance edit $x_1 + x_2 = x_3$ with $x_1 = 2.7$, $x_2 = 7.6$, and $x_3 = 10.3$ are rounded to integers. If conventional rounding is used, $x_1$ is rounded to 3, $x_2$ to 8, and $x_3$ to 10, and the balance edit becomes violated.

From a purely statistical point of view, rounding errors are rather unimportant as by their nature they have virtually no influence on publication figures. Rounding errors may be important, however, when we look at them from the point of view of the SDE *process*. Some statistical offices apply automatic editing procedures for random errors, such as automatic editing procedures based on the Fellegi–Holt paradigm. Such automatic editing procedures are computationally very demanding. The complexity of the automatic error localization problem increases rapidly as the number of violated edit rules becomes larger, irrespective of the magnitude of these violations. A record containing many rounding errors may hence be too complicated to solve for an automatic editing procedure for random errors, even if the number of random errors is actually low. From the point of view of the SDE process, it may therefore be advantageous to resolve rounding errors at the beginning of the editing process.

Scholtus (2008a,b) describes a heuristic method for resolving rounding errors. The method does not lead to solutions that are "optimal" according to some criterion, such as that the number of changed variables or the total change in value is minimized. Instead the method just leads to a good solution. Given that the statistical impact of resolving rounding errors is small, a time-consuming and complex algorithm aimed at optimizing some target function is not necessary anyway. The heuristic method is referred to as the "scapegoat algorithm", because for each record assumed to contain rounding errors, a number of variables, the "scapegoats," are selected beforehand and the rounding errors are resolved by changing only the values of the selected variables. Under certain very mild conditions, the algorithm guarantees that exactly one choice of values exists for the selected variables such that the balance edits become satisfied. Different variables are selected for each record to minimize the effect of the adaptations on published aggregates.

In general, the obtained solution might contain fractional values, whereas most business survey variables are restricted to be integer-valued. If this is the case, a controlled rounding algorithm could be applied to the values to obtain an integer-valued solution (see, e.g., Salazar-González et al., 2004). Under certain additional mild conditions, which appear to be satisfied by most data sets arising in practice, the problem of fractional values does not occur, however. For details, we refer to Scholtus (2008a,b).

Rounding errors often occur in combination with other “obvious” systematic errors. For instance, a sign error might be obscured by the presence of a rounding error. Scholtus (2008a,b) provides a single mathematical model for detecting sign errors and rounding errors simultaneously.

### 6.3. *Example of a Fellegi–Holt-based algorithm*

In this subsection, we sketch an algorithm based on the Fellegi–Holt paradigm to illustrate how such algorithms work. We will first describe the algorithm for numerical data and later describe how the algorithm can be adapted to categorical data. In Subsection 6.4, the algorithm is illustrated by means of an example.

The basic idea of the algorithm we describe in this section is that for each record, a binary tree is constructed. In our case, we use a binary tree to split up the process of searching for solutions to the error localization problem. We need some terminology with respect to binary trees before we can explain our algorithm. Following Cormen et al. (1990), we recursively define a binary tree as a structure on a finite set of nodes that either contains no nodes or comprises three disjoint sets of nodes: a root node, a left (binary) subtree, and a right (binary) subtree. If the left subtree is nonempty, its root node is called the left child node of the root node of the entire tree, which is then called the parent node of the left child node. Similarly, if the right subtree is nonempty, its root node is called the right child node of the root node of the entire tree, which is then called the parent node of the right child node. All nodes except the root node in a binary tree have exactly one parent node. Each node in a binary tree can have at most two (nonempty) child nodes. A node in a binary tree that has only empty subtrees as its child nodes is called a terminal node or also a leaf. A nonleaf node is called an internal node. In each internal node of the binary tree generated by our algorithm, a variable is selected that has not yet been selected in any predecessor node. If all variables have already been selected in a predecessor node, we have reached a terminal node of the tree.

We first assume that no values are missing. After the selection of a variable, two branches, that is, subtrees, are constructed; in one branch, we assume that the observed value of the selected variable is correct, and in the other branch, we assume that the observed value is incorrect. By constructing a binary tree, we can, in principle, examine all possible error patterns and search for the best solution to the error localization problem.

In the branch in which we assume that the observed value is correct, the variable is fixed to its original value in the set of edits. In the branch in which we assume that the observed value is incorrect, the selected variable is eliminated from the set of edits. A variable that has either been fixed or eliminated is said to have been treated (for the corresponding branch of the tree). To each node in the tree, we have an associated set of edits for the variables that have not yet been treated in that node. The set of edits corresponding to the root node of our tree is the original set of edits.

Eliminating a variable is nontrivial, as removing a variable from a set of edits may imply additional edits for the remaining variables. To illustrate why edits may need to be generated, we give a very simple example. Suppose we have three variables $x_1$, $x_2$, and $x_3$, and two edits $x_1 \leq x_2$ and $x_2 \leq x_3$. If we want to eliminate variable $x_2$ from these edits, we cannot simply delete this variable and the two edits but have to generate the new edit $x_1 \leq x_3$ implied by the two old ones for else we could have that $x_1 > x_3$ and the original set of edits cannot be satisfied.

To ensure that the original set of edits can be satisfied, Fourier–Motzkin elimination is used. For inequalities, Fourier–Motzkin elimination basically consists of using the variable to be eliminated to combine these inequalities pairwise (if possible), as we did in the above example. If the variable to be eliminated is involved in a balance edit, we use this equation to express this variable in terms of the other variables and then use this expression to eliminate the variable from the other edits.

In each branch of the tree, the set of current edits is updated. Updating the set of current edits is the most important aspect of the algorithm. How the set of edits has to be updated depends on whether the selected variable is fixed or eliminated. Fixing a variable to its original value is done by substituting this value in all current edits, failing as well as nonfailing. Conditional on fixing the selected variable to its original value, the new set of current edits is a set of implied edits for the remaining variables in the tree. That is, conditional on the fact that the selected variable has been fixed to its original value, the remaining variables have to satisfy the new set of edits. As a result of fixing the selected variable to its original value, some edits may become tautologies, that is, may become satisfied by definition. An example of a tautology is "$1 \geq 0$." Such a tautology may, for instance, arise if a variable $x$ has to satisfy the edit $x \geq 0$, the original value of $x$ equals 1, and $x$ is fixed to its original value. These tautologies may be discarded from the new set of edits. Conversely, some edits may become self-contradicting relations. An example of a self-contradicting relation is "$0 \geq 1$." If self-contradicting relations are generated, this particular branch of the binary tree cannot result in a solution to the error localization problem. Eliminating a variable by means of Fourier–Motzkin elimination amounts to generating a set of implied edits that do not involve this variable. This set of implied edits has to be satisfied by the remaining variables. In the generation process, we need to consider all edits, both the failing edits as well as the nonfailing edits, in the set of current edits pairwise. The generated set of implied edits plus the edits not involving the eliminated variable become the set of edits corresponding to the new node of the tree.

If values are missing in the original record, the corresponding variables only have to be eliminated from the set of edits (and not fixed).

After all variables have been treated, we are left with a set of relations involving no unknowns. If and only if this set of relations contains no self-contradicting relations, the variables that have been eliminated to reach the corresponding terminal node of the tree can be imputed consistently such that all original edits can be satisfied (cf. Theorems 1 and 2 in De Waal and Quere, 2003). The set of relations involving no unknowns may be the empty set, in which case it obviously does not contain any self-contradicting relations. In the algorithm, we check for each terminal node of the tree whether the variables that have been eliminated to reach this node can be imputed consistently. Of all the sets of variables that can be imputed consistently, we select the ones with the lowest sum of reliability weights. In this way, we find all optimal solutions to the error localization problem (cf. Theorem 3 in De Waal and Quere, 2003).

For categorical data, the algorithm is essentially the same as the above-described algorithm. The only difference between the algorithms for the two data types is the way in which variables are eliminated and hence the way in which implied edits are generated. As we have mentioned above, when a variable in numerical data is eliminated, we pairwise apply Fourier–Motzkin elimination. For the case of categorical data, when a variable is to be eliminated, we apply the method originally proposed by Fellegi and Holt (1976) to generate implied edits, using the variable to be eliminated as the so-called generating field.

We again denote the number of variables by $n$. Furthermore, we denote the domain, that is, the set of all allowed values of a variable $i$, by $D_i$. In the case of categorical data, an edit $j$ is usually written in so-called *normal form*, that is, as a collection of sets $F_i^j$ $(i = 1, 2, \ldots, n)$:

$$(F_1^j, F_2^j, \ldots, F_n^j), \tag{8}$$

meaning that if for a record with values $(v_1, v_2, \ldots, v_n)$ we have $v_i \in F_i^j$ for all $i = 1, 2, \ldots, n$, then the record fails edit $j$, otherwise the record satisfies edit $j$. For instance, suppose we have three variables: *Marital status*, *Age*, and *Relation to head of household*. The possible values of *Marital status* are Married, Unmarried, Divorced, and Widowed, of *Age* are "less than 16 years" and "16 years or older," and of *Relation to head of household* are Spouse, Child, and Other. The edit that someone who is less than 16 years cannot be married can be written in normal form as

$$(\{\text{Married}\}, \{\text{less than 16 years}\}, \{\text{Spouse, Child, Other}\}). \tag{9}$$

The edit that someone who is not married cannot be the spouse of the head of household can be written as

$$(\{\text{Unmarried, Divorced, Widowed}\}, \{\text{less than 16 years, 16 years or older}\}, \{\text{Spouse}\}). \tag{10}$$

Note that whereas for numerical data, an edit is *satisfied* if a certain condition is fulfilled, for categorical data, an edit is *violated* if a certain condition is fulfilled. We assume that all edits are written in the form (8). We call a categorical variable $i$ involved in an edit given by (8) if $F_i^j \neq D_i$.

Now, if we eliminate a variable $v_r$, we start by determining all index sets $S$ such that

$$\bigcup_{j \in S} F_r^j = D_r \tag{11}$$

and

$$\bigcap_{j \in S} F_i^j \neq \emptyset \qquad \text{for all } i = 1, \ldots, r-1, r+1, \ldots, n. \tag{12}$$

From these index sets, we select the *minimal* ones, that is, the index sets $S$ that obey (11) and (12), but none of their proper subsets obey (11).

Given such a minimal index set $S$, we construct the implied edit

$$\left(\bigcap_{j \in S} F_1^j, \ldots, \bigcap_{j \in S} F_{r-1}^j, D_r, \bigcap_{j \in S} F_{r+1}^j, \ldots, \bigcap_{j \in S} F_n^j\right). \tag{13}$$

For example, if we eliminate variable *Marital status* from the edits (9) and (10), we obtain the implied edit

({Married, Unmarried, Divorced, Widowed}, {less than 16 years}, {Spouse}),

which expresses that someone who is less than 16 years of age cannot be the spouse of the head of household.

Note that variable $v_r$ is not involved in the edit (13). By adding the implied edits resulting from all minimal sets $S$ to the set of edits and removing all edits involving the eliminated variable, one obtains the updated set of current edits.

The algorithms for numerical and categorical data can be combined into a single algorithm for numerical and categorical data (see De Waal and Quere, 2003). That algorithm can be further extended to deal with integer-valued data in a heuristic manner (see De Waal, 2005).

### *6.4. Illustration of the Fellegi–Holt-based algorithm*

In this subsection, we illustrate the algorithm for numerical data described in Subsection 6.3 by means of an example. Suppose the explicit edits are given by

$$T = P + C \tag{14}$$

$$P \leq 0.5T \tag{15}$$

$$-0.1T \leq P \tag{16}$$

$$T \geq 0 \tag{17}$$

$$T \leq 550N, \tag{18}$$

where $T$ denotes the turnover of an enterprise, $P$ its profit, $C$ its costs, and $N$ the number of employees. Let us consider a specific erroneous record with values $T = 100$, $P = 40{,}000$, $C = 60{,}000$, and $N = 5$. Edits (16)–(18) are satisfied, whereas edits (14) and (15) are violated. The reliability weights of the variables $T$, $P$, and $C$ equal 1, and the reliability weight of variable $N$ equals 2. As edits (14) and (15) are violated, the record contains errors.

We select a variable, say $T$, and construct two branches: one where $T$ is eliminated and one where $T$ is fixed to its original value. We consider the first branch and eliminate $T$ from the set of edits. We obtain the following edits.

$$P \leq 0.5(P + C) \tag{19}$$

$$-0.1(P + C) \leq P \tag{20}$$

$$P + C \geq 0 \tag{21}$$

$$P + C \leq 550N. \tag{22}$$

Edits (19)–(21) are satisfied, edit (22) is violated. Because edit (22) is violated, changing $T$ is not a solution to the error localization problem. If we were to continue examining the branch where $T$ is eliminated by eliminating and fixing more variables, we would find that the best solution in this branch has an objective value (5) equal to 3. We now consider the other branch where $T$ is fixed to its original value. We fill in the original value of $T$ in edits (14)–(18) and obtain (after removing any tautology that might arise) the following edits:

$$100 = P + C \tag{23}$$

$$P \leq 50 \tag{24}$$

$$-10 \leq P \tag{25}$$

$$100 \leq 550N. \tag{26}$$

Edits (25) and (26) are satisfied, and edits (23) and (24) are violated. We select another variable, say $P$, and again construct two branches: one where $P$ is eliminated and one where $P$ is fixed to its original value. Here, we only examine the former branch and obtain the following edits (again after removing any tautology that might arise):

$$100 - C \leq 50$$

$$-10 \leq 100 - C \tag{27}$$

$$100 \leq 550N \tag{28}$$

Only edit (27) is violated. We select variable $C$ and again construct two branches: one where $C$ is eliminated and another where $C$ is fixed to its original value. We only examine the former branch and obtain edit (28) as the only implied edit. As this edit is satisfied by the original value of $N$, changing $P$ and $C$ is a solution to the error localization problem. By examining all branches of the tree, including the ones that we have skipped here, we find that this is the only optimal solution to this record.

## 7. Macro-editing

### 7.1. *Introduction to macro-editing*

Thus far, we have examined micro-editing methods, that is, methods that use the data of a single record and related auxiliary information to check and correct it. In this section, we examine macro-editing methods. Macro-editing techniques often examine the potential impact on survey estimates to identify suspicious data in individual records. Macro-editing can lead to the detection of errors that would go unnoticed with selective editing or automatic editing. Micro-editing and macro-editing are complementary. Errors that are apparent from one point of view may not be apparent from the other. For instance, micro-editing may reveal more errors than macro-editing, but macro-editing may trace bigger, more influential errors.

Macro-editing can be seen as a form of selective editing. A major difference between macro-editing and selective editing is the moment at which they are applied in the SDE process. Whereas selective editing can be used early in the SDE process while a substantial part of the data to be edited may still be collected, macro-editing is used at the end of the SDE process when (most of) the data have already been collected. This allows a different approach. Whereas selective editing basically treats each record to be edited separately, macro-editing treats the data set to be edited as a whole. Selective editing checks whether each record to be edited is plausible; macro-editing checks whether the data set as a whole is plausible.

We distinguish between two forms of macro-editing. The first form is called the aggregation method (see, e.g., Granquist, 1990, 1995). It formalizes and systematizes what every statistical agency does before publication: verifying whether figures to be published seem plausible. This is accomplished by comparing quantities in publication tables with the same quantities in previous publications, with quantities based on register data, or with related quantities from other sources. Examples of this form of macro-editing are the foreign trade surveys of the Netherlands (see Van de Pol and Diederen, 1996) and Canada (see Laflamme et al., 1996). Only if an unusual quantity is observed, a micro-editing procedure is applied to the individual records and fields contributing to this quantity. An unusual quantity may, for instance, be detected by checking whether

$$\left| \frac{Y - \hat{Y}}{\hat{Y}} \right| > p/100, \tag{29}$$

where $Y$ denotes a publication figure to be checked, $\hat{Y}$ an "anticipated" value for this publication figure, and $p$ a certain percentage. If (29) holds true, that is, if $Y$ deviates more than $p\%$ from its anticipated value, the microdata underlying the publication figure $Y$ are subjected to a micro-editing procedure.

Generally, in software packages for macro-editing (see also Chapter 13), the influence of individual observations on population figures is estimated. Starting from the most influential observation, individual data can be interactively checked and corrected, raising weights can be adjusted, or records can be removed all together. The interactive editing process terminates when further corrections have a negligible effect on the estimated population figures. The impact of such changes to the data on estimates of publication figures can be monitored by re-estimating the publication figures each time a change has been made.

A second form of macro-editing is the distribution method. Here, the available data, either the data set to be edited or a reference data set, are used to characterize the distribution of the variables. Next, all individual values are compared with this distribution. Typically, measures of location and spread are computed. Records containing values that could be considered uncommon (given the distribution) are candidates for further inspection and possibly for correction. The distribution method is examined in more detail in Subsection 7.2.

### *7.2. Exploratory data analysis and related techniques*

There is an area in statistics providing all kinds of techniques for analyzing the distribution of variables, namely exploratory data analysis (EDA) (see, e.g., Tukey, 1977). Many EDA techniques can be applied in macro-editing. Advocates of EDA stress the importance of the use of graphical techniques (see, e.g. Chambers et al., 1983). These techniques can provide much more insight in the behavior of variables than numerical techniques do. Graphs of the distribution of the data show a lot of information and often are capable of showing unexpected properties that would not have been discovered if just numerical quantities were computed.

The application of EDA techniques for data editing has been the subject of a number of papers. DesJardins (1997) gives a description of how several EDA techniques can be used during the data editing stage. The techniques range from traditional EDA techniques, such as boxplots and scatterplots, to more advanced techniques, such as so-called 6D-plots and "industry plots." Industry plots have been devised by DesJardins to present a comprehensive overview of an entire survey in a single graph. An industry plot attempts to depict the multivariate relation between the key variables of each individual company. In such an industry plot, the "normal" companies are clustered around the center point of the plot, whereas outlying companies lie far from the center point. Once an industry plot has been designed, it is a powerful tool to quickly detect outliers. However, designing an industry plot seems a nontrivial task. Another drawback of industry plots is that they have to be redesigned for each type of survey.

Bienas et al. (1997) describe the application of graphical EDA techniques to identify potentially incorrect data in two different surveys. The EDA techniques that were applied were boxplots, scatterplots, and bivariate fitting. Transformations, such as taking logarithms, were applied to the data to discern patterns more easily. The fitting methods that were applied were ordinary least squares and resistant regression, which reduces the influence of outlying cases on the fit of the regression model. Ordinary least squares fitting proved very useful when there are only a few unusual records that can be easily distinguished from the usual records. In the case that there are relatively many outlying records, resistant fitting proved to be more useful. Bienas et al. (1997) mention that the EDA approach can be combined with batch-type micro-editing.

Frequently used techniques in software packages for macro-editing (see also Chapter 13) are so-called anomaly plots, time series analysis, outlier detection methods, and the already mentioned EDA techniques such as boxplots and scatterplots. Anomaly plots are graphical overviews of the important estimates, where unusual estimates are highlighted. Once suspicious data have been detected on a macrolevel, in such an anomaly plot, one can usually drill-down to subpopulations and individual records. Outlying records can often be identified by means of graphical EDA techniques. In particular, scatterplots comparing the data in the current period to the corresponding data in a previous period can often be used. Also, the concept of linked plots, where an outlier in one plot is automatically also highlighted as outlier in other plots, helps the analyst to study the behavior of an outlier in one plot in other plots. Besides graphical EDA techniques, software packages for macro-editing sometimes also offer a mathematical (multivariate) outlier detection algorithm for identifying outlying records.

## *7.3. Possibilities and pitfalls of macro-editing*

One may wonder whether the application of macro-editing approaches will result in microdata of less statistical quality than would have been obtained after exhaustive micro-editing. Data users who consider their applications more "micro" than the usual publication figures of totals and means often have to be convinced that macro-editing approaches are not harmful, especially for multivariate micro analysis. A reassuring point for these users is that so-called microlevel analysis does not actually involve the inspection of individual records. Multivariate analysis brings along the estimation of parameters that are always some sort of an aggregate. For instance, the estimation of output elasticities for energy, labor, and material from annual construction survey data turned out to differ less than one standard deviation when comparing results after no data editing, selective data editing, and exhaustive data editing (see Van de Pol and Bethlehem, 1997).

Another point to convince skeptic data users that application of macro-editing techniques does not lead to a loss of quality is that all methods of data editing, including the traditional exhaustive micro-editing approach, will leave some errors unnoticed and not corrected because not all errors are apparent. In case of overediting, some other fields will be changed without good justification. Data editors are human, which means that they make errors and miss errors from time to time. This will occur less often when they have good tools to navigate in the data and to distinguish between important and unimportant errors. Because multivariate methods often are sensitive to outliers, data editing methods that trace these outliers, such as macro-editing techniques, should be welcomed.

Despite these points in favor of macro-editing, changing from an exhaustive micro-editing approach to a much less exhaustive macro-editing approach is a big step for many potential users. They have to be convinced that the application of a macro-editing approach can result in data of sufficiently high-quality. In the words of DesJardins (1997): "Introducing graphical EDA can be a tough nut."

Graphical macro-editing certainly offers a lot of possibilities, but unfortunately there are some problems and pitfalls one should be aware of when applying this approach and before deciding to develop a software tool for macro-editing.

A limitation of macro-editing, at least in the applications known to us, is that it is much more suited for editing of economic data than of social data. A drawback of macroediting is that the time and resources required for editing are hard to predict. A further drawback is that one needs to wait with the macro-editing process until all or most of the data have arrived and are ready for processing.

Persons can interpret data that are depicted in several scatterplots simultaneously, so graphical macro-editing allows one to edit relatively large amounts of data simultaneously. There is also a limit, however. It is impossible for (most) human beings to interpret, say, 10 scatterplots at the same time. For a data set with many important key variables, graphical macro-editing is usually not the most suitable editing method, unless applied in combination with other SDE methods.

A very important methodological drawback of the aggregation method is that this approach involves the risk that records contributing to publication figures that are considered nonsuspicious still do contain influential errors, errors that were not detected and corrected. This will lead to biased publication figures. Relying fully on macro-editing may also prevent publication of unexpected but true changes in trend. Outliers in one direction may be removed until outliers in the opposite direction cancel out the unexpected trend.

For more on the possibilities and pitfalls of macro-editing, we refer to De Waal et al. (2000).

### *7.4. Macro-editing versus micro-editing*

An advantage of macro-editing in comparison to micro-editing is that micro-editing, either automatically or interactively, requires edits. Specifying edits, for instance the bounds of ratio-edits, can be difficult and time-consuming. Of course, one does not want to specify edits that are too lenient in the sense that influential incorrect data are not detected. On the other hand, one also does not want to specify edits that are too severe in the sense that many correct, or only slightly incorrect, records are considered suspicious because this would result in overediting. So, not having to specify edits clearly has its benefits.

Although not having to specify edits is one of the advantages of macro-editing, it is at the same time also a bit dangerous. When edits are specified and the micro-editing approach is used, it is clear when a record will be considered suspicious. When the macro-editing approach is used and edits are not specified, it is for a substantial part left to the subject-matter specialists who do the editing to decide which records are suspicious and which are not. That is, it will depend on the partly subjective judgment of the subject-matter specialists how the records are divided into suspicious and nonsuspicious records.

Automatic editing and imputation can be implemented in such a way that the results can be reproduced, that is, if the same data set is edited and imputed again, the same results are obtained. This is not the case for interactive editing and macro-editing. The results of interactive editing and macro-editing are partly subjective, that is, they partly depend on the specific subject-matter specialist who edits the data. Different subject-matter specialists, or even the same subject-matter specialist at different moments in time, may obtain different results.

When the incoming raw data contain many errors, that is, when almost every record needs correction, micro-editing is more efficient than macro-editing. In that case, the extra effort to trace erroneous records from a macro point of view should be postponed until the data set has a reasonably good quality due to micro-editing.

An argument for maintaining some sort of micro-editing is that this is the only way to make sure that records are internally consistent, that is they satisfy the edits. Also, automatic correction of "obvious" systematic errors should always be done before macro-editing in our opinion. This form of micro-editing is not costly and can improve the estimates of aggregates and distributions used in the macro-editing phase.

## 8. A strategy for statistical data editing

In this section, we propose a strategy for SDE in which we combine the editing techniques described in the previous sections. We assume that a data set to be edited has already been collected. Our proposed strategy depends on whether the data are numerical or categorical. We start with our strategy for numerical data. For these data, we advocate an SDE approach that consists of the following phases:

(1) correction of "obvious" (systematic) errors, such as thousand-errors, sign errors, interchanged returns and costs, and rounding errors;
(2) application of selective editing to split the records in a critical stream and a noncritical stream;

(3) editing of the data: the records in the critical stream are edited interactively, and the records in the noncritical stream are edited automatically;
(4) validation of the publication figures by means of macro-editing.

The above steps are used at Statistics Netherlands in the production process for structural business statistics (see De Jong, 2002). The goal of the first phase is to treat errors that are obviously errors and that once detected are also easy to correct. Typically, "obvious" systematic errors, such as thousand-errors, are dealt within this phase. The main goal of the second phase is to select the influential errors. In the third phase, these errors are treated interactively. Most influential errors will be resolved by the subject-matter specialists; in some cases, the respondents will be re-contacted. In the third phase also noninfluential errors are treated. As these errors often occur in a high number of records, they have to be detected and corrected as efficiently as possible, both in terms of budget and time. Automatic editing is hence the most often used way to handle noninfluential errors. The fourth phase, the validation phase, is performed by subject-matter specialists who use macro-editing to compare the publication figures based on the edited data to publication figures from a previous year, for instance. In this final step, the focus is more on the overall results than on the correctness of individual records. An additional goal of macro-editing is to check whether the SDE process itself has functioned well.

One could argue that with selective editing, the automatic editing step is superfluous. Personally, we advocate the use of automatic editing, even when selective editing is used. We mention three reasons. First, the sum of the errors of the records in the noncritical stream may have an influential effect on the publication figures, even though each error itself may be noninfluential. This can in particular be the case if the data contain systematic errors as then a substantial part of the data may be biased in the same direction. The correction of "obvious" systematic errors evidently leads to data of higher statistical quality. In addition, provided that the set of edits used is sufficiently powerful, application of the Fellegi-Holt paradigm also generally results in data of higher statistical quality. Second, many noncritical records will be internally inconsistent, that is, they will fail specified edits, if they are not edited, which may lead to problems when publication figures are calculated or when microdata are released to external researchers. Finally, automatic editing provides a mechanism to check the quality of the selective editing procedures. If selective editing is well-designed and well-implemented, the records that are not selected for interactive editing need no or only slight adjustments. Records that are substantially changed during the automatic editing step, therefore, possibly point to an incorrect design or implementation of the selective editing step.

Phases 2 and 4 of our strategy for numerical data do not, or hardly, apply to categorical data. For those data, our proposed strategy simply consists of checking and correcting errors, first obvious ones as in phase 1 and later more complex ones as in phase 3, as much as possible automatically.

We feel that a combined approach using, if applicable, selective editing, interactive editing, automatic editing, and macro-editing can improve the efficiency of the traditional interactive SDE process while at the same time maintaining or even enhancing the statistical quality of the produced data.

## 9. Discussion

In this chapter, we have focused on identifying errors in the data as this has traditionally been considered the most important aim of SDE in practice. In fact, however, this is only one of the goals of SDE. Granquist (1995) identifies the following main goals of SDE:

(1) identify error sources to provide feedback on the entire survey process
(2) provide information about the quality of the incoming and outgoing data
(3) identify and treat influential errors and outliers in individual data
(4) when needed, provide complete and consistent individual data.

During the last few years, the first two goals—providing feedback on the other survey phases, such as the data collection phase, and providing information on the quality of the collected data and the final results—have gained in importance. The feedback on other survey phases can be used to improve those phases and reduce the amount of errors arising in these phases. The SDE forms part of the entire statistical process at NSIs. A direction for potential future research is hence the relation between SDE and other steps of the statistical process, such as data collection (see, e.g., Børke, 2008) and statistical disclosure control (see Shlomo and De Waal, 2008). In the next few years, the first two goals of SDE are likely to become even more important.

From our discussion of SDE, the reader may have gotten the feeling that the basic problems of SDE are fixed and will never change. This is definitely not the case! The world is rapidly changing and this certainly holds true for SDE. The traditional way of producing data, by sending out questionnaires to selected respondents or interviewing selected respondents, and subsequently processing and analyzing the observed data, is for a substantial part being replaced by making use of already available register data. This presents us with new problems related to SDE.

First, differences in definitions of the variables and the population units between the available register data and the desired information have to be resolved before register data can be used. This can be seen as a special form of SDE. Second, the external register data may have to be edited themselves. Major differences between editing self-collected survey data and external register data are that in the former case, one knows, in principle, all the details regarding the data collection process, whereas in the latter case one does not, and that in the former case one can recontact respondents as a last resort, whereas in the latter case this is generally impossible. Another difference is that the use of register data requires co-operation with other agencies, for instance tax offices. An increased use of register data seems to be the way of the future for most NSIs. The main challenge for the near future for SDE is to adapt itself, so we can handle these data efficiently and effectively.

SDE and imputation are more closely related than space restrictions allow us to describe in this book. In practice, one often needs a well-balanced selected mix of SDE and imputation techniques to edit and impute a data set (see, e.g., Pannekoek and De Waal, 2005). Often in survey and census practice, imputation is carried out to deal with edit failures. Apart from briefly sketching the basic idea underlying NIM, where imputations are used to identify erroneous fields, in Section 6, we have not examined the relation between SDE and imputation any further.

In this chapter, we have also not examined methods that deal simultaneously with outliers and missing data in multivariate settings. There is a growing literature on these methods; we refer the interested reader to Béguin and Hulliger (2004, 2008), Ghosh-Dastidar and Schafer (2006), Elliott and Stettler (2007), and in particular to the paper by Little and Smith (1987). For some recent work on outliers in the context of data editing, we refer to Di Zio et al. (2008). For how to deal with outliers, in general, we refer to Chapter 11.

In this book, we also do not examine how to impute missing data in such a way that all specified edits are satisfied. Some recent work has been carried out in this area. For imputation of categorical data subject to edits, we refer to Winkler (2003) and for imputation of numerical data subject to edits to Drechsler and Raghunathan (2008), Pannekoek et al. (2008), and, in particular, Tempelman (2007).

*Sample Surveys: Design, Methods and Applications, Vol. 29A*
ISSN: 0169-7161

DOI: 10.1016/S0169-7161(08)00010-2

# 10

# Imputation and Inference in the Presence of Missing Data

*David Haziza*

## 1. Introduction

Nonresponse inevitably occurs in most, if not all, surveys. Essentially, survey statisticians distinguish between two types of nonresponse, total or unit nonresponse and partial or item nonresponse. Unit nonresponse occurs when all the survey variables are missing or not enough usable information is available. For example, a sample unit may refuse to participate in the survey or it may prematurely terminate an interview. In the latter case, the sample unit is identified as a total nonrespondent even if some information has been collected because it is judged to be insufficient. Item nonresponse occurs when some but not all the survey variables have missing values. For example, a sample unit may refuse to respond to sensitive items or may not know the answer to some items, or missing values can be the result of edit failures. A comprehensive discussion of statistical data editing is given in Chapter 9. Unit nonresponse is usually treated by a weight adjustment procedure. With a weight adjustment procedure, the nonrespondents are deleted and the survey weights of respondents are adjusted to compensate for the deletions. These procedures are described in Chapter 8. Imputation is a process where an artificial value is produced to replace a missing value. Although imputation is sometimes used to handle unit nonresponse, it is mostly used to compensate for item nonresponse.

The main effects of (unit or item) nonresponse include as follows: (i) bias of point estimators, (ii) increase of the variance of point estimators (since the observed sample size is smaller than the sample size initially planned), and (iii) bias of the complete data variance estimators. The main objective when treating (unit or item) nonresponse is the reduction of the nonresponse bias, which occurs if respondents and nonrespondents are different with respect to the survey variables.

Although multiple imputation is gaining in popularity in national statistical institutes, the vast majority of surveys use some form of single imputation. For this reason, we mainly focus on single imputation that consists of creating a single imputed value to replace a missing value resulting in a single complete data file. Multiple imputation (Rubin, 1987), discussed in Section 10.7, consists of creating $M \geq 2$ imputed values to fill in a missing value resulting in $M$ complete data files.

Single imputation is widely used in surveys for treating item nonresponse because it presents the following advantages: (i) it leads to the creation of a complete data file, so the results of different analyses are consistent with each other and (ii) unlike weighting adjustment for each item, imputation allows for the use of a single survey weight for all items.

However, imputation presents certain risks, for instance: (i) even though imputation leads to the creation of a complete data file, inferences are valid only if the underlying assumptions about the response mechanism and/or the imputation model are satisfied; (ii) some imputation methods tend to distort the distribution of the variables of interest (i.e., the variables being imputed); (iii) treating the imputed values as if they were observed may lead to a substantial underestimation of the variance of the estimator, especially if the item nonresponse rate is appreciable; (iv) imputing for each item separately has the effect of distorting relationships between variables.

In the absence of nonresponse, survey samplers usually try to avoid using estimation procedures whose validity depends on the validity of a given model. To avoid assumptions on the distribution of the data, the properties of estimators are generally based on the sampling design. This approach is the so-called design-based approach or randomization approach to survey sampling. This does not mean that models are useless under the design-based approach. In fact, they play an important role in the determination of efficient sampling and estimation procedures. The use of models is unavoidable in the presence of nonresponse, and the properties of (point and variance) estimators (e.g., bias and variance) will depend on the validity of the assumed models. Consequently, imputation is essentially a modeling exercise. The quality of the estimates will thus depend on the availability (at the imputation stage) of good auxiliary information and on its judicious use in the construction of imputed values and/or imputation classes.

Auxiliary information plays an important role in surveys because it allows the survey statistician to use more efficient sampling and estimation procedures. Also, it can be used to reduce nonsampling errors such as nonresponse errors, coverage errors, and measurement errors. In our discussion, the problem of coverage errors, and measurement errors is not addressed. We distinguish between three sets of auxiliary variables. The first is the set of design variables we assume to be available for all the units in the population at the design stage. The design variables are typically used to stratify the population or use some form of probability proportional-to-size sampling. The second set of auxiliary variables is used to construct imputed values and/or imputation classes and is typically related to the variable being imputed and/or to the response probability to this variable. In other words, these variables will be useful in reducing the nonresponse bias and possibly reduce the nonresponse variance. Finally, we assume that, at the estimation stage, a set of auxiliary variables (often called calibration or benchmark variables) is available for all the sample units and that the population total for each variable in this set is known. The calibration variables are usually specified by the data users to ensure consistency with known totals. Note that the three sets of auxiliary variables are not necessarily disjoint, so a given auxiliary variable can be used at different stages in a survey.

## 2. Context and defnitions

In this section, we begin by introducing the expansion estimator and the generalized regression (GREG) estimator in the context of complete data and their corresponding imputed estimators in Section 2.1. In Section 2.2, we present several important imputation methods used in practice. The concept of nonresponse mechanism is discussed in Section 2.3. Finally, two approaches for inference are presented in Section 2.4.

## 2.1. An imputed estimator

Let $U = \{1, 2, \ldots, N\}$ be a population of $N$ identifiable elements. In this section, we consider the problem of estimating a population total $Y = \sum_{i \in U} y_i$, where $y_i$ denotes the $i$th value of the variable of interest $y$, $i = 1, \ldots, N$. To that end, we select a random sample, $s$, of size $n$, according to a given sampling design $p\,(s)$. Let $\pi_i$ denote the first-order inclusion probability of unit $i$ in the sample and let $d_i = 1/\pi_i$ denote its design weight. We assume that the sampling design is noninformative in the sense that the probability of inclusion in the sample does not depend on the variable of interest after accounting for the design variables in the estimation procedure (e.g., see Pfefferman 1993).

In the absence of nonresponse, we consider two complete data estimators of the population total $Y$. The first estimator is the well-known expansion estimator given by

$$\hat{Y}_\pi = \sum_{i \in s} d_i y_i, \tag{1}$$

for example, see Särndal et al. (1992, Chapter 2). The estimator $\hat{Y}_\pi$ is a $p$-unbiased estimator of $Y$, that is, $E_p\left(\hat{Y}_\pi\right) = Y$, where $E_p$ denotes the expectation with respect to the sampling design $p(.)$. Note that $\hat{Y}_\pi$ does not use any auxiliary information apart from the one used in the sampling procedure. To denote a variance estimator of $\hat{Y}_\pi$, it is convenient to use the operator notation. Using this notation, we write $v\left(\hat{Y}_\pi\right) = v\,(y)$. For example, an $p$-unbiased variance estimator of the design variance, $V_p\left(\hat{Y}_\pi\right)$, is given by

$$v\left(\hat{Y}_\pi\right) \equiv v\,(y) = \sum_{i \in s} \sum_{j \in s} \frac{\left(\pi_{ij} - \pi_i \pi_j\right)}{\pi_{ij}} \frac{y_i}{\pi_i} \frac{y_j}{\pi_j},$$

where $\pi_{ij}$ denotes the second-order inclusion probability for units $i$ and $j$ and $V_p$ denotes the variance with respect to the sampling design.

Often, some auxiliary information is available at the estimation stage. In this case, an alternative estimator that incorporates auxiliary information is the GREG estimator given by

$$\hat{Y}_{\mathrm{G}} = \sum_{i \in s} w_i y_i, \tag{2}$$

where $w_i = d_i g_i$ and

$$g_i = 1 + c_i^{-1}\left(\mathbf{X} - \hat{\mathbf{X}}_\pi\right)' \left(\sum_{i \in s} d_i c_i^{-1} \mathbf{x}_i \mathbf{x}_i'\right)^{-1} \mathbf{x}_i \tag{3}$$

with $\mathbf{X} = \sum_{i \in U} \mathbf{x}_i$, $\hat{\mathbf{X}}_\pi = \sum_{i \in s} d_i \mathbf{x}_i$, $\mathbf{x}_i = \left(x_{1i}, \ldots, x_{pi}\right)'$ is a vector of $p$ auxiliary (calibration) variables and $c_i$ denotes a known constant attached to unit $i$. The GREG estimator $\hat{Y}_{\mathrm{G}}$ is asymptotically $p$-unbiased for $Y$. Also, $\hat{Y}_{\mathrm{G}}$ belongs to the class of calibration estimators since $\hat{\mathbf{X}}_{\mathrm{G}} = \sum_{i \in s} w_i \mathbf{x}_i = \mathbf{X}$ (Deville and Särndal, 1992). The ratio estimator and the poststratified estimator are special cases of $\hat{Y}_{\mathrm{G}}$. Depending on the context, we use either $\hat{Y}_\pi$ or $\hat{Y}_{\mathrm{G}}$ as prototype estimators (i.e., estimators that we would have used in the ideal situation of complete response to item $y$). An asymptotically $p$-unbiased estimator of $V_P\left(\hat{Y}_{\mathrm{G}}\right)$ is given by $v\left(\hat{Y}_{\mathrm{G}}\right) = v\,(ge)$, where $e_i = y_i - \mathbf{x}_i'\hat{\boldsymbol{\gamma}}$ and $\hat{\boldsymbol{\gamma}} = \left(\sum_{i \in s} d_i \mathbf{x}_i \mathbf{x}_i'/c_i\right)^{-1} \sum_{i \in s} d_i \mathbf{x}_i y_i / c_i$ (e.g., see Särndal, et al., 1992; Chapters 6 and 7).

In the presence of nonresponse to item $Y$, it is not possible to compute either (1) or (2) since some $y$-values are missing. In this case, using $\hat{Y}_{\mathrm{G}}$ as the prototype estimator, we define an imputed estimator of the population total $Y$ as

$$\hat{Y}_{\mathrm{IG}} = \sum_{i \in s} w_i r_i y_i + \sum_{i \in s} w_i \left(1 - r_i\right) y_i^*, \tag{4}$$

where $r_i$ is a response indicator for unit $i$ such that $r_i = 1$ if unit $i$ responded to item $y$ and $r_i = 0$, otherwise, and $y_i^*$ denotes the imputed value for missing $y_i$ whose value depends on the imputation method used. Note that the imputed estimator (4) is simply the weighted sum of observed and imputed values in the sample. Thus, its computation does not require the response indicators $r_i$ in the imputed data file. Similarly, if we use $\hat{Y}_\pi$ as the prototype estimator, an imputed estimator, denoted by $\hat{Y}_{\mathrm{I}\pi}$, is obtained from (4) by replacing $w_i$ with $d_i$. Finally, let $s_r$ and $s_m$ denote the random sets of respondents and nonrespondents, respectively. We have $s = s_r \cup s_m$. Under complete response to item $y$ (i.e., $s_r = s$), note that both imputed estimators $\hat{Y}_{\mathrm{I}\pi}$ and $\hat{Y}_{\mathrm{IG}}$ reduce to the prototype estimators, $\hat{Y}_\pi$ and $\hat{Y}_{\mathrm{G}}$, respectively.

## 2.2. *Imputation methods*

Imputation methods may be classified into two broad classes: deterministic and random (or stochastic). Deterministic methods are those that yield a fixed imputed value given the sample if the imputation process is repeated as opposed to random methods that do not necessarily yield the same imputed value. Most of the imputation methods (deterministic and random) used in practice can be represented as a special case of the following model (Kalton and Kasprzyk, 1986):

$$\begin{aligned} &y_i = f\left(\mathbf{z}_i\right) + \epsilon_i, \\ &E_m\left(\epsilon_i\right) = 0, \mathrm{Cov}_m\left(\epsilon_i, \epsilon_j\right) = 0 \quad \text{if } i \neq j,\ V_m\left(\epsilon_i\right) = \sigma_i^2 = \sigma^2 \nu\left(\mathbf{z}_i\right) \end{aligned} \tag{5}$$

where $\mathbf{z} = \left(z_1, \ldots, z_q\right)'$ is a vector of auxiliary variables available at the imputation stage for all the sampled units, $f(.)$ is a given function, $\sigma^2$ is an unknown parameter, $\nu(.)$ is a known function, and $E_m$, $V_m$, and $\mathrm{Cov}_m$ denote, respectively, the expectation, the variance, and the covariance operators with respect to the model (5). Note that the imputation model (5) is used to motivate the particular imputation used. In the case of deterministic imputation, the imputed value $y_i^*$ is obtained by estimating $f(\mathbf{z})$ by $\hat{f}_r(\mathbf{z})$ based on the responding units, $i \in s_r$; that is, $y_i^* = \hat{f}_r(\mathbf{z}_i)$ for $i \in s_m$. Random imputation can be seen as a deterministic imputation plus a random noise $\epsilon^*$; that is, $y_i^* = \hat{f}_r(\mathbf{z}_i) + \hat{\sigma}_i \epsilon_i^*$ for $i \in s_m$, where $\hat{\sigma}_i$ is an estimator of $\sigma_i$. It is natural to select (usually with replacement) the random component $\epsilon_i^*$ from the set, $E_r = \left\{e_j;\ j \in s_r\right\}$, of standardized residuals observed from the responding units, where $e_j = \frac{1}{\hat{\sigma}_j}\left[y_j - \hat{f}\left(\mathbf{z}_j\right)\right] - \bar{e}_r$ and $\bar{e}_r$ is the weighted mean of the $e_j$'s for the respondents. Commonly used deterministic methods include regression imputation (REGI), ratio imputation (RAI), mean imputation (MI), auxiliary value imputation (AVI), and nearest neighbor imputation (NNI). These methods are described in Section 2.2.1. Except for NNI, deterministic methods tend to distort the distribution of the variables being imputed. Commonly used random imputation methods include random REGI and random hot-deck imputation (RHDI). These methods are described in Section 2.2.2. Random imputation methods tend to preserve the distribution of the variable being imputed (Chen et al., 2000) but they suffer from an additional component of variance due to the use of a random imputation mechanism.

An alternative classification of the imputation methods consists of distinguishing the donor imputation methods and the predicted value imputation methods. In the case of donor imputation methods, the nonrespondent (recipient) missing values are replaced by the values of a respondent (donor). Therefore, the imputed values are actual, observed data. Donor imputation is convenient when it is desired to impute more than one variable at a time since a unique donor can be used to impute all missing values of a given nonrespondent while satisfying postimputation edit constraints specified by subject-matter specialists. This feature help preserve relationships between survey variables, contrary to independently imputing each variable. Satisfying edit constraints is a desirable goal when a public-use microdata file is produced since it ensures that no gross error will remain in the imputed survey data file. The NNI and the RHDI are examples of donor imputation methods. Predicted value imputation uses the value obtained from fitting a model using the respondent values. These methods do not yield observed values, in general, which could lead to awkward imputed values, for example, when the variable being imputed is binary. Both deterministic and random REGI are examples of predicted value imputation. The reader is referred to Kovar and Whitridge (1995) for a discussion of imputation methods in business surveys.

### 2.2.1. *Some deterministic imputation methods*

Commonly used deterministic imputation methods include the following:

(i) REGI: It consists of using a regression model to predict the missing values. In this case, we have $f(\mathbf{z}_i) = \mathbf{z}_i'\boldsymbol{\beta}$ and $v(\mathbf{z}_i) = \boldsymbol{\lambda}'\mathbf{z}_i$ for a specified vector of constants $\boldsymbol{\lambda}$. It follows that the imputed values are given by

$$y_i^* = \mathbf{z}_i'\hat{\mathrm{B}}_r,\ i \in s_m, \tag{6}$$

where

$$\hat{\mathrm{B}}_r = \left(\sum_{i \in s} \omega_i r_i \mathbf{z}_i \mathbf{z}_i' / \left(\boldsymbol{\lambda}'\mathbf{z}_i\right)\right)^{-1} \sum_{i \in s} \omega_i r_i \mathbf{z}_i y_i / \left(\boldsymbol{\lambda}'\mathbf{z}_i\right) \tag{7}$$

is the weighted least square estimator of $\boldsymbol{\beta}$ based on the responding units and $\omega_i$ is a weight attached to unit $i$. A special case of REGI is simple linear regression imputation (SLRI), which is obtained when $\mathbf{z}_i = (1, z_i)$ and $v(\mathbf{z}_i) = 1$. As we argue in Section 3.3.1, several options for $\omega_i$ are available. When $\omega_i = d_i$, we are in presence of survey weighted deterministic REGI, whereas the choice

$\omega_i = 1$ leads to unweighted deterministic REGI. Other weighting alternatives are possible depending on the approach used for inference (see section 3.3.1).

(ii) RAI: It is a special case of REGI that uses a single auxiliary variable. In this case, we have $f(\mathbf{z}_i) = \beta z_i$ and $\nu(\mathbf{z}_i) = z_i$. It follows that the imputed values are given by

$$y_i^* = \frac{\bar{y}_r}{\bar{z}_r} z_i, \; i \in s_m, \tag{8}$$

where $(\bar{y}_r, \bar{z}_r) = \frac{1}{\sum_{i \in s} \omega_i r_i} \sum_{i \in s} \omega_i r_i (y_i, z_i)$ are the means of the respondents for variables $y$ and $z$ respectively. Note that RAI assumes that the relationship between the variable of interest $y$ and the auxiliary variable $z$ goes through the origin.

(iii) MI: It is another special case of REGI for which $\mathbf{z}_i = 1$ for all $i \in s$, $f(\mathbf{z_i}) = \beta$, and $\nu(\mathbf{z}_i) = 1$. It follows that the imputed values are given by

$$y_i^* = \bar{y}_r, \; i \in s_m. \tag{9}$$

(iv) NNI: It is a nonparametric imputation method. Hence, we do not attempt to specify the form of $f(\mathbf{z}_i)$ nor the function $\nu(\mathbf{z}_i)$. We have

$$y_i^* = y_j \text{ such that dist}\left(\mathbf{z}_i, \mathbf{z}_j\right) \text{ is minimum, } j \in s_r, \tag{10}$$

where dist (., .) is a distance measure to be determined (e.g., the Euclidean distance). It is assumed that the auxiliary variables in the vector $\mathbf{z}$ are all quantitative and have been standardized so that they are on the same scale. For example, this could be done either by subtracting the mean and dividing by the standard deviation for each auxiliary variable or by replacing the values of each auxiliary variable by their ranks.

(v) AVI: For a given nonresponding unit $i \in s_m$, AVI consists of replacing the missing value of a variable of interest $y$ using only reported values coming from this unit $i$ but using other auxiliary variables. Therefore, a unit with a missing $y$-value is never imputed using reported $y$-values of other units when AVI is used. A special case of this imputation method is historical imputation (sometimes called carry-forward imputation), which is particularly useful in repeated economic surveys for variables that tend to be stable over time (e.g., number of employees). Under AVI, the imputed values are given by

$$y_i^* = z_i. \tag{11}$$

AVI can be seen as a special case of the imputation model (5) with $f(\mathbf{z}_i) = z_i$.

### 2.2.2. *Some random imputation methods*

Commonly used stochastic imputation methods include the following:

(i) Random REGI: It is closely related to deterministic REGI (6) except that a random noise is added to the prediction. It follows that the imputed values are given by

$$y_i^* = \mathbf{z}_i'\hat{\mathbf{B}}_r + \left(\boldsymbol{\lambda}'\mathbf{z}_i\right)^{1/2} \epsilon_i^*, \tag{12}$$

where $\epsilon_i^* = e_j$ with probability $\omega_j \Big/ \sum\limits_{l\in s} \omega_l r_l$ and $e_j = \left(\boldsymbol{\lambda}'\mathbf{z}_j\right)^{-1/2}\left(y_j - \mathbf{z}_j'\hat{\mathbf{B}}_r\right) - \bar{e}_r$ with $\bar{e}_r = \frac{\sum\limits_{j\in s}\omega_j r_j e_j}{\sum\limits_{j\in s}\omega_j r_j}$.

(ii) RHDI: It is a special case of random REGI with $\mathbf{z}_i = 1$ for all $i \in s$ and $\nu\left(\mathbf{z}_i\right) = 1$. RHDI consists in selecting donor values with replacement from the set of respondents $s_r$ with probabilities $\omega_j / \sum_{l\in s_r} \omega_l$ to replace missing values. That is,

$$y_i^* = y_j \text{ with } P\left(y_i^* = y_j\right) = \frac{\omega_j}{\sum_{l\in s}\omega_l r_l},\ j \in s_r. \tag{13}$$

RHDI can be seen as weighted MI plus a random noise; that is, $y_i^* = \bar{y}_r + \epsilon_i^*$, where $\epsilon_i^* = e_j$ with probability $\frac{\omega_j}{\sum\limits_{l\in s}\omega_l r_l}$ and $e_j = \left(y_j - \bar{y}_r\right) - \bar{e}_r$.

### 2.3. *The nonresponse mechanism*

The situation in the presence of nonresponse is similar to the one prevailing in the presence of two-phase sampling, which is often used in surveys when the sampling frame contains little or no auxiliary information. In the context of two-phase sampling, the survey statistician knows the inclusion probabilities to both phases, which makes it possible to construct $p$-unbiased estimators of population totals. In the presence of nonresponse to item $y$, the set of respondents can be viewed as a second phase sample, except that the inclusion probabilities in the set of respondents are unknown. In this context, these inclusion probabilities are called response probabilities and will be denoted by $p_i$. Note that $p_i$ may depend on the realized sample $s$. Since the $p_i$'s are typically unknown, we must make some assumptions about the nonresponse mechanism, which we define next.

Let $\mathbf{I} = (I_1, \ldots, I_N)'$ be the vector of sample selection indicators, where $I_i = 1$ if unit $i$ is selected in the sample and $I_i = 0$, otherwise, and $\mathbf{r} = (r_1, \ldots, r_N)'$ be the vector of response indicators to item $y$. The distribution of $\mathbf{r}$, $q\left(\mathbf{r}|\mathbf{I}\right)$, is called the nonresponse mechanism and is generally unknown (except in the case of planned nonresponse). Let $p_i = P\left(r_i = 1|s, i \in s\right)$ be the response probability of unit $i$ to item $y$. We assume that $p_i > 0$ for all $i$, which may not be realistic in most surveys because a fraction of sampled units are hard core nonrespondents (Kott, 1994). Also, we assume that the units respond independently of one another; that is, $p_{ij} = P\left(r_i = 1, r_j = 1|s, i \in s, j \in s, i \neq j\right) = p_i p_j$. The assumption of independence is usually satisfied in practice although it is easy to come up with situations where it is not satisfied. For example, in cluster sampling, the units within clusters (e.g., households) may not respond independently of one another.

The causes leading to missing values are numerous. For example, a value could be missing because of edit failure or because the unit refused to respond. In this case, we clearly have two distinct nonresponse mechanisms. However, trying to describe all the possible reasons that lead to missing values is practically unrealistic (Schafer and Graham, 2002). In the remainder of this chapter, we will thus refer to *the* nonresponse mechanism.

The simplest type of nonresponse mechanism is the uniform nonresponse mechanism for which the response probability is constant for all the units in the population. That is, $p_i = p$ for all $i$. In this case, the probability of response is independent of all the variables available (auxiliary variables and variables of interest). When the nonresponse mechanism is uniform, we say that the data are *missing completely at random* (MCAR) (Rubin, 1976). This mechanism is, in general, not realistic in most practical applications. However, it is customary to assume uniform response within imputation classes (see Section 5).

We now discuss the notion of ignorability of the nonresponse mechanism, which is always defined with respect to an imputation model (Rubin, 1976). Let $\mathbf{z}$ be the vector of auxiliary variables selected in the imputation model. The nonresponse mechanism is ignorable if the probability of response, $p_i$, is independent of the error term $\varepsilon_i$ in the imputation model. That is, it is ignorable if after accounting for $\mathbf{z}$ in the imputation procedure, the response probability does not depend on the error term. Note that the response probability may depend on the error term when both the response probability and the error term are related to $\mathbf{z}$ but there must be no residual relationship between the probability of response and the error term after accounting for $\mathbf{z}$. Otherwise, the nonresponse mechanism is nonignorable. When the nonresponse mechanism is ignorable, the data are said to be *missing at random* (MAR), whereas when it is nonignorable, the data are said to be *not missing at random* (NMAR). Note that a uniform nonresponse mechanism is automatically ignorable. It is possible to eliminate the nonresponse bias when the nonresponse mechanism is ignorable. When the probability of response depends on the variable of interest (and so the nonresponse mechanism is automatically nonignorable), the estimators will remain biased even after accounting for the appropriate auxiliary information but we expect to achieve a good bias reduction if the auxiliary variables are highly related to the variable being imputed. In practice, the ignorability of the nonresponse mechanism is assumed because it is generally impossible to test whether we are in presence of ignorable or nonignorable response except in the context of planned nonresponse. In the majority of surveys, we can expect that the nonresponse is nonignorable, and so a nonresponse bias is generally unavoidable. In this case, it is important to make a serious modeling exercise to build a reasonable model that will help reduce the nonresponse bias. Estimation in the presence of a nonignorable nonresponse mechanism has been considered by Greenlees et al. (1982), Beaumont (2000), and Qin et al. (2002) among others.

To illustrate the concept of ignorability, consider the case of a scalar $\mathbf{z}$ and suppose that the probability of response depends on the variable $z$. If the variable of interest $y$ is related to $z$ (so the error term depends on $z$), then the nonresponse mechanism is ignorable if $z$ is used in the imputation procedure (by using, e.g., SLRI or RAI); otherwise, the nonresponse mechanism is nonignorable. If the variable $z$ is not related to $y$ (so the error term does not depend on $z$), then there is no need to include $z$ in the imputation model.

### 2.4. Approaches to inference

Different approaches may be used for evaluating the quality (e.g., bias and variance) of the imputed estimator and to derive corresponding variance estimators. To understand the nature of these approaches, we first identify three sources of randomness: (i) the imputation model $m$, which generates the vector of $y$-values, $\mathbf{y} = (y_1, \ldots, y_N)'$; (ii) the sampling design $p(s)$, which generates the vector of sample selection indicators, $\mathbf{I} = (I_1, \ldots, I_N)'$, and (iii) the nonresponse mechanism $q(\mathbf{r}|\mathbf{I})$, which generates the vector of response indicators, $\mathbf{r} = (r_1, \ldots, r_N)'$. Different combinations of these distributions may be used to assess the properties of an estimator. Next, we describe two such combinations that will lead to the nonresponse model (NM) approach and the imputation model (IM) approach.

### *2.4.1. The nonresponse model approach*

In the NM approach, explicit assumptions, called the nonresponse model, about the nonresponse mechanism are made. We assume that the probability of response $p_i$, for unit $i$, is linked to an $l$-vector of auxiliary variables $\mathbf{u}_i$ according to a model $p_i = f\left(\mathbf{u}_i'\boldsymbol{\eta}\right)$, where $\boldsymbol{\eta}$ is the $l$-vector of model parameters. A frequently used model is the logistic regression model given by

$$p_i = \exp\left(\mathbf{u}_i'\boldsymbol{\eta}\right)/\exp\left(1 + \mathbf{u}_i'\boldsymbol{\eta}\right). \tag{14}$$

Letting $\mathbf{u}_i = 1$ for all $i$ in (14) leads to the uniform nonresponse model (UNM), under which the response probability is assumed to be constant for all $i$. In the NM approach, inference is made with respect to the joint distribution induced by the sampling design and the assumed nonresponse model, whereas the vector of $y$-values, $\mathbf{y}$, is treated as fixed. The NM approach has been studied by Beaumont (2005), Haziza and Rao (2006), Kim and Park (2006), Rao (1990, 1996), Rao and Sitter (1995), and Shao and Steel (1999) among others.

### *2.4.2. The imputation model approach*

In the IM approach, explicit assumptions about the distributions of the values of the variables of interest are made. Here, inference is with respect to the joint distribution induced by the imputation model, the sampling design, and the nonresponse model. Unlike the NM approach, the underlying nonresponse mechanism is not explicitly specified, except for the MAR assumption. The IM approach has been studied by Brick et al. (2004), Deville and Särndal (1994), Särndal (1992), and Shao and Steel (1999), among others. Under both deterministic and random REGI, the model (5) with $f\left(\mathbf{z}_i\right) = \mathbf{z}_i'\boldsymbol{\beta}$ and $v\left(\mathbf{z}_i\right) = \boldsymbol{\lambda}'\mathbf{z}_i$ is assumed.

### *2.4.3. Which approach to use?*

Recall that imputation is primarily used to reduce the nonresponse bias, assuming that some auxiliary information can explain the item to be imputed and/or the response probability. Hence, the choice between modeling the response probability and modeling the item of interest should be dictated by the quality of the nonresponse and imputation models. Although it may seem intuitively more appealing to model the variable of interest (IM approach), there are some cases encountered in practice for which it may be easier to model the response probability to item $y$ (NM approach). For example, Haziza and Rao (2006) reported the case of the Capital Expenditure Survey conducted at Statistics Canada that produces data on investment made in Canada. For this survey, two important variables of interest are capital expenditures on new construction (CC) and capital expenditures on new machinery and new equipment (CM). In a given year, a large number of businesses have not invested any amount of money on new construction or new machinery. As a result, the sample data file contains a large number of zeros for the two variables CC and CM. Modeling these two variables may thus prove to be difficult, whereas modeling the response probabilities may be simpler if auxiliary information related to the response probability to CC an CM is available. For example, a logistic regression model could be fitted with the response indicator to CC (CM) as the dependent variable.

## 3. Bias of the imputed estimator

In Section 2.1, we noted that the complete data estimator $\hat{Y}_{\mathrm{G}}$ given by (2) is asymptotically $p$-unbiased for the population total $Y$. What can we say about the imputed estimator $\hat{Y}_{\mathrm{IG}}$ given by (4)? To study its properties, we use the standard decomposition of the total error, $\hat{Y}_{\mathrm{IG}} - Y$, as a starting point:

$$\hat{Y}_{\mathrm{IG}} - Y = \left(\hat{Y}_{\mathrm{G}} - Y\right) + \left(\hat{Y}_{\mathrm{IG}} - \hat{Y}_{\mathrm{G}}\right). \tag{15}$$

The term $\hat{Y}_{\mathrm{G}} - Y$ in (15) is called the sampling error, whereas the term $\hat{Y}_{\mathrm{IG}} - \hat{Y}_{\mathrm{G}}$ is called the nonresponse error. In practice, it is impossible to measure the magnitude of the nonresponse bias (before and after imputation). It is customary to assume that, after the data has been imputed, the nonresponse bias is small and can be neglected. This assumption is only tenable if the nonresponse mechanism is ignorable with respect to the imputation model.

### *3.1. Nonresponse bias under the NM approach*

Using (15), the bias of the imputed estimator $\hat{Y}_{\mathrm{I}}$ under deterministic imputation can be expressed as

$$\mathrm{Bias}\left(\hat{Y}_{\mathrm{IG}}\right) = E_{pq}\left(\hat{Y}_{\mathrm{IG}} - Y|\mathbf{I}\right) \approx E_p B_q\left(\hat{Y}_{\mathrm{IG}}|\mathbf{I}\right), \tag{16}$$

where $B_q\left(\hat{Y}_{\mathrm{IG}}|\mathbf{I}\right) = E_q\left(\hat{Y}_{\mathrm{IG}} - \hat{Y}_G|\mathbf{I}\right)$ is the conditional nonresponse bias under the NM approach. Hence, the imputed estimator $\hat{Y}_{\mathrm{IG}}$ is asymptotically $pq$-unbiased if $B_q\left(\hat{Y}_{\mathrm{IG}}|\mathbf{I}\right)$ is asymptotically equal to zero for any sample $s$. This condition is satisfied if the nonresponse mechanism is ignorable with respect to the assumed imputation model. In the case of random imputation, we need to take the imputation mechanism (which consists of randomly selecting the residuals) into account. In this case, the bias of the imputed estimator $\hat{Y}_{\mathrm{IG}}$ can be expressed as

$$\mathrm{Bias}\left(\hat{Y}_{\mathrm{IG}}\right) = E_{pqI}\left(\hat{Y}_{\mathrm{IG}} - Y|\mathbf{I}\right) \approx E_p B_{qI}\left(\hat{Y}_{\mathrm{IG}}|\mathbf{I}\right), \tag{17}$$

where the subscript $I$ denotes the imputation mechanism and $B_{qI}\left(\hat{Y}_{\mathrm{IG}}|\mathbf{I}\right) = E_{qI}\left(\hat{Y}_{\mathrm{IG}} - \hat{Y}_{\mathrm{G}}|\mathbf{I}\right)$.

### *3.2. Nonresponse bias under the IM approach*

Using (15), the bias of the imputed estimator $\hat{Y}_{\mathrm{IG}}$ under deterministic imputation can be expressed as

$$\mathrm{Bias}\left(\hat{Y}_{\mathrm{IG}}\right) = E_{mpq}\left(\hat{Y}_{\mathrm{IG}} - Y\right) = E_{pqm}\left(\hat{Y}_{\mathrm{IG}} - Y|\mathbf{I}, \mathbf{r}\right) = E_{pq} B_m\left(\hat{Y}_{\mathrm{IG}}|\mathbf{I}, \mathbf{r}\right), \tag{18}$$

where $B_m\left(\hat{Y}_{\mathrm{IG}}|\mathbf{I},\mathbf{r}\right) = E_m\left(\hat{Y}_{\mathrm{IG}} - \hat{Y}_{\mathrm{G}}|\mathbf{I},\mathbf{r}\right)$ is the conditional nonresponse bias under the IM approach. Hence, the imputed estimator $\hat{Y}_{\mathrm{IG}}$ is asymptotically $mpq$-unbiased if $B_m\left(\hat{Y}_{\mathrm{IG}}|\mathbf{I},\mathbf{r}\right)$ is equal to zero for any sample $s$. Note that the second equality in (18) follows from the fact that the sampling design is assumed to be noninformative and the nonresponse mechanism ignorable. In the case of random imputation, the bias of the imputed estimator $\hat{Y}_{\mathrm{IG}}$ can be expressed as

$$\mathrm{Bias}\left(\hat{Y}_{\mathrm{IG}}\right) = E_{mpqI}\left(\hat{Y}_{\mathrm{IG}} - Y|\mathbf{I},\mathbf{r}\right) \approx E_{pq}B_{mI}\left(\hat{Y}_{\mathrm{IG}}|\mathbf{I},\mathbf{r}\right), \tag{19}$$

where $B_{mI}\left(\hat{Y}_{\mathrm{IG}}|\mathbf{I},\mathbf{r}\right) = E_{mI}\left(\hat{Y}_{\mathrm{IG}} - \hat{Y}_{\mathrm{G}}|\mathbf{I},\mathbf{r}\right)$.

### *3.3. The bias in some special cases*

The bias of imputed estimator under REGI, AVI, and NNI is addressed in this section.

#### *3.3.1. Regression imputation*

As we mentioned in Section 2.2.1, there are several valid choices of the weights $\omega_i$. Here, we consider three options: (i) the option $\omega_i = 1$, which lead to unweighted imputation; (ii) the option $\omega_i = d_i$, which leads to the customary survey weighted imputation, and (iii) the option $\omega_i = d_i\frac{1-\hat{p}_i}{\hat{p}_i} \equiv \tilde{d}_i$, where $\hat{p}_i$ denotes the estimated response probability to item $y$ for unit $i$. Note that the estimated probabilities may be obtained by fitting a parametric (e.g., logistic) regression model or by using a nonparametric nonresponse model, which is typically weakly dependent on modeling assumptions (Da Silva and Opsomer, 2006; Little and An, 2004). The third option for $\omega_i$ was studied by Beaumont (2005), Kim and Park (2006), and Haziza and Rao (2006). A question at this point is: what choice of weight $\omega_i$ is more adequate? The answer to this question is far from obvious and partly depends on the approach (NM or IM) used for inference.

Consider the imputed estimator $\hat{Y}_{\mathrm{IG}}$ given by (4) under deterministic REGI for which the imputed values are given by (6). Under the choice $\omega_i = 1$, it is easy to show that under the IM approach, we have $B_m\left(\hat{Y}_{\mathrm{IG}}|\mathbf{I},\mathbf{r}\right) = 0$ if the imputation model is correctly specified. As a result, the imputed estimator $\hat{Y}_{\mathrm{IG}}$ is asymptotically $mpq$-unbiased for $Y$. In other words, if we are willing to put complete reliance on the imputation model, the use of the design weights $d_i$ in the construction of the imputed values is not justified. However, under the NM approach, the imputed estimator $\hat{Y}_{\mathrm{IG}}$ is generally asymptotically $pq$-biased under the choice $\omega_i = 1$. In this case, the asymptotic conditional bias of $\hat{Y}_{\mathrm{IG}}$ is given by

$$B_q\left(\hat{Y}_{\mathrm{IG}}|\mathbf{I}\right) \approx -\sum_{i\in s} w_i\,(1-p_i)\left(y_i - \mathbf{z}_i'\hat{\mathbf{B}}_p^{(1)}\right), \tag{20}$$

which is not equal to zero, in general, where $\hat{\mathbf{B}}_p^{(1)} = \left(\sum_{i\in s} p_i\mathbf{z}_i\mathbf{z}_i'/\left(\boldsymbol{\lambda}'\mathbf{z}_i\right)\right)^{-1} \sum_{i\in s} p_i\mathbf{z}_i y_i/\left(\boldsymbol{\lambda}'\mathbf{z}_i\right)$. Note that even when $p_i = p$ (UNM approach), the bias (20) does not vanish. Therefore, although the choice $\omega_i = 1$ leads to a valid imputed estimator under the IM approach, it cannot generally be justified under the NM approach.

We now consider the option $\omega_i = d_i$. It can be shown that under the IM approach, the conditional nonresponse bias, $B_m\left(\hat{Y}_{\mathrm{IG}}|\mathbf{I}, \mathbf{r}\right)$, is equal to zero if the imputation is correctly specified. On the other hand, under the UNM approach, the conditional nonresponse bias, $B_q\left(\hat{Y}_{\mathrm{IG}}|\mathbf{I}\right)$, is also asymptotically equal to zero if the units actually have equal response probabilities. As a result, the imputed estimator $\hat{Y}_{\mathrm{IG}}$ is asymptotically *mpq*-unbiased and *pq*-unbiased (under the UNM approach) for $Y$. However, under a general NM approach (i.e., the $p_i$'s may vary from one unit to another), the imputed estimator $\hat{Y}_{\mathrm{IG}}$ is asymptotically *pq*-biased. Thus, the imputed values (6) using $\omega_i = d_i$ are generally inadequate under the NM approach. In fact, the asymptotic conditional nonresponse bias of $\hat{Y}_{\mathrm{IG}}$ is given by

$$B_q\left(\hat{Y}_{\mathrm{IG}}|\mathbf{I}\right) \approx -\sum_{i \in s} w_i(1-p_i)\left(y_i - \mathbf{z}_{\mathbf{i}}'\hat{\mathbf{B}}_p^{(d)}\right), \tag{21}$$

where $\hat{\mathbf{B}}_p^{(d)} = \left(\sum_{i\in s} d_i p_i \mathbf{z}_i \mathbf{z}_i'/\left(\boldsymbol{\lambda}'\mathbf{z}_i\right)\right)^{-1} \sum_{i\in s} d_i p_i \mathbf{z}_i y_i/\left(\boldsymbol{\lambda}'\mathbf{z}_i\right)$. The bias (21) vanishes when $p_i = p$ (UNM approach), as expected. What choice of weights $\omega_i$ will lead to an imputed estimator that is valid under either the IM approach or the NM approach? One such option is $\omega_i = \tilde{d}_i$, where it can be shown that the imputed estimator $\hat{Y}_{\mathrm{IG}}$ is asymptotically unbiased under either the IM approach if the imputation is correctly specified or the NM approach if the nonresponse model is correctly specified. In this case, the imputed estimator is said to be doubly robust in the sense that it can be justified from either approach if at least one of the models (imputation or nonresponse) is correctly specified. When $\hat{p}_i = \hat{p}$, note that the imputed values obtained using the option $\omega_i = d_i$ are identical to those obtained using the option $\omega_i = \tilde{d}_i$.

The double robustness property is attractive in practice because it provides some protection against the misspecification of one model or the other. However, if we put complete reliance on the imputation model (IM approach), the option $\omega_i = 1$ is generally more efficient than the other two options, especially if the design weights $d_i$ are not significantly correlated with the variable being imputed and are widely dispersed. This situation occurs frequently in household surveys. This point is illustrated in a simulation study in Section 3.4.1. Doubly robust inference is discussed in Haziza and Rao (2006), Kang and Schafer (2008), Kott (1994), Little and An (2004), Robins et al. (2008), among others.

Finally, note that from a bias perspective, the results under random REGI, for which the imputed values are given by (12), are identical to those obtained under deterministic REGI since $E_I\left(\epsilon_i^*|\mathbf{I}, \mathbf{r}\right) = 0$.

### 3.3.2. *Auxiliary value imputation*

The AVI is motivated by the model (5) with $f(\mathbf{z}_i) = z_i$. This imputation model is somewhat restrictive because it assumes that the intercept goes through the origin and that the slope is equal to 1. Under the IM approach, the imputed estimator $\hat{Y}_{\mathrm{IG}}$ is *mpq*-unbiased for $Y$. However, under the NM approach, the conditional nonresponse bias is given by $B_q\left(\hat{Y}_{\mathrm{IG}}|\mathbf{I}\right) = -\sum_{i\in s} w_i(1-p_i)(y_i - z_i)$, which is not equal to zero, in general.

Therefore, the imputed estimator $\hat{Y}_{\text{IG}}$ under AVI is $pq$-biased. For more details on AVI, the reader is referred to Shao (2000) and Beaumont et al. (2007).

### *3.3.3. Nearest neighbor imputation*

The NNI is motivated by the model (5). Chen and Shao, 2000 considered the special case of a scalar $z$. They showed that, under some mild regularity conditions, the imputed estimator $\hat{Y}_{\text{I}\pi}$ is asymptotically $mpq$-unbiased for $Y$. The main advantage of NNI is that the functions $f(.)$ and $\nu(.)$ do not need to be specified explicitly in order for the imputed estimator to be asymptotically unbiased. The reader is also referred to Rancourt et al. (1994).

## *3.4. Some numerical examples*

In this section, we perform two simulation studies. The first investigates on the performance of imputed estimators (in terms of relative bias and mean square error) under both unweighted and weighted RHDI, whereas the second illustrates the importance of performing a complete modeling exercise before choosing an imputation method.

### *3.4.1. Simulation study 1*

We generated a finite population of size $N = 1000$ with three variables: two variables of interest $y_1$ and $y_2$ and an auxiliary variable $\psi$. To do so, we first generated $\psi$ from a gamma distribution with shape parameter $\alpha_0 = 1$ and scale parameter $\alpha_1 = 50$. Then, the $y_1$-values were generated according to the model, $y_{1i} = 2\psi_i + \epsilon_i$, where the $\epsilon_i$'s are generated from a normal distribution with mean 0 and variance $\sigma^2$. The variance $\sigma^2$ was chosen to lead to a model $R^2$-value approximately equal to 0.64. Finally, the variable $y_2$ was generated independently of $y_1$ and $\psi$ from a gamma distribution with shape parameter $\alpha_0 = 2$ and scale parameter $\alpha_1 = 50$. The objective is to estimate the population totals $Y_j = \sum_{i \in U} y_{ji}$, $j = 1, 2$.

From $U$, we generated $R = 25,000$ samples of size $n = 50$ according to the Rao–Sampford proportional-to-size sampling procedure (Rao, 1965; Sampford, 1967), using $\psi$ as the measure of size. In this case, the inclusion probability of unit $i$ in the sample is defined as $\pi_i = n\frac{\psi_i}{\sum_{i \in U}\psi_i}$. Note that the coefficient of variation of the $\psi$-values, CV$(\psi)$, was set to 1.2, which may be considered as high. Under the Rao–Sampford design, we have CV$(\psi) =$ CV$(\pi)$ and so the sampling weights $d_i$ are widely dispersed. Also, note that the variable $y_1$ is highly related to the size variable $\psi$, whereas the variable $y_2$ is unrelated to $\psi$. This situation is frequent in surveys with multiple characteristics (e.g., Rao, 1966). In other words, unlike for the variable $y_2$, the variable $y_1$ is highly related to the sampling weight $d_i$. In each simulated sample, nonresponse to items $y_1$ and $y_2$ was independently generated according to a uniform response mechanism with probability 0.6. To compensate for nonresponse to items $y_1$ and $y_2$, we used weighted RHDI for which the imputed values are given by (13) with $\omega_i = d_i$ and unweighted RHDI which consists of setting $\omega_i = 1$.

From each simulated sample, we calculated the imputed estimator $\hat{Y}_{\text{I}\pi}$. We define the Monte–Carlo expectation of an estimator $\hat{\theta}$ as

$$E_{\text{MC}}\left(\hat{\theta}\right) = \frac{1}{R}\sum_{r=1}^{R}\hat{\theta}^{(r)}, \tag{22}$$

where $\hat{\theta}^{(r)}$ denotes the estimator $\hat{\theta}$ for the $r$th simulated sample, $r = 1, \ldots, R$. As a measure of the bias of $\hat{Y}_{\mathrm{I}\pi}$, we used the Monte–Carlo percent relative bias (RB) given by

$$\mathrm{RB}_{\mathrm{MC}}\left(\hat{Y}_{\mathrm{I}\pi}\right) = 100 \times \frac{E_{\mathrm{MC}}\left(\hat{Y}_{\mathrm{I}\pi}\right) - Y}{Y}, \tag{23}$$

where $E_{\mathrm{MC}}\left(\hat{Y}_{\mathrm{I}\pi}\right)$ is obtained from (22) by replacing $\hat{\theta}$ with $\hat{Y}_{\mathrm{I}\pi}$. As a measure of variability of $\hat{Y}_{\mathrm{I}\pi}$, we used the Monte–Carlo mean square error (MSE) given by

$$\mathrm{MSE}_{\mathrm{MC}}\left(\hat{Y}_{\mathrm{I}\pi}\right) = E_{\mathrm{MC}}\left(\hat{Y}_{\mathrm{I}\pi} - Y\right)^2, \tag{24}$$

which is obtained from (22) by replacing $\hat{\theta}$ with $\left(\hat{Y}_{\mathrm{I}\pi} - Y\right)^2$. Table 1 reports the Monte–Carlo percent relative bias given by (23) as well as the relative efficiency defined as $\mathrm{RE} = \frac{\mathrm{MSE}_{\mathrm{MC}}^{(\mathrm{un})}\left(\hat{Y}_{\mathrm{I}\pi}\right)}{\mathrm{MSE}_{\mathrm{MC}}^{(\mathrm{w})}\left(\hat{Y}_{\mathrm{I}\pi}\right)}$, where $\mathrm{MSE}_{\mathrm{MC}}^{(\mathrm{un})}\left(\hat{Y}_{\mathrm{I}\pi}\right)$ and $\mathrm{MSE}_{\mathrm{MC}}^{(\mathrm{w})}\left(\hat{Y}_{\mathrm{I}\pi}\right)$ denote the Monte–Carlo MSE of $\hat{Y}_{\mathrm{I}\pi}$ under unweighted RHDI and weighted RHDI, respectively. We use a similar notation for the Monte–Carlo percent relative bias in Table 1.

For the variable $y_1$, the RB of the imputed estimator under unweighted RHDI is large (approximately 28.5%), whereas it is small under weighted RHDI (see Table 1). This result is not surprising since unweighted RHDI does not account for the sampling weight $d_i$ despite the fact that the sampling weight and the variable $y_1$ are strongly related. In other words, the imputation model is misspecified because it failed to include the sampling weight. For the variable $y_2$, which is poorly related to the sampling weight, we note that the imputed estimator under both unweighted RHDI and weighted RHDI shows a small bias. However, the imputed estimator under unweighted RHDI is more efficient than the corresponding estimator under weighted RHDI with a value of RE equal to 0.83. This can be explained by the fact that, since the sampling weights are widely dispersed and are not related to $y_2$, their use in the construction of imputed values is essentially equivalent to adding random noise.

### 3.4.2. *Simulation study 2*

We generated three finite populations of size $N = 1000$ with two variables: a variable of interest $y$ and an auxiliary variable $z$. To do so, we first generated $z$ from a gamma distribution with shape parameter $\alpha_0 = 2$ and scale parameter $\alpha_1 = 25$. For population 1, the $y$-values were generated according to the model $y_i = 2z_i + \epsilon_i$. For population 2, we used the model $y_i = 50 + 2z_i + \epsilon_i$. For population 3, we used the model $y_i = e^{0.05z_i} + \epsilon_i$. In the three population, the $\epsilon_i$'s are generated from a normal distribution with mean 0

Table 1
Weighted versus unweighted RHDI

| Variable | $\mathrm{RB}_{\mathrm{MC}}^{(\mathrm{w})}\left(\hat{Y}_{\mathrm{I}\pi}\right)$ | $\mathrm{RB}_{\mathrm{MC}}^{(\mathrm{un})}\left(\hat{Y}_{\mathrm{I}\pi}\right)$ | RE |
|---|---|---|---|
| $y_1$ | 1.6 | 28.4 | 4.2 |
| $y_2$ | 0.5 | 0.9 | 0.83 |

and variance $\sigma^2$. For populations 1 and 2, the variance $\sigma^2$ was chosen to lead to a model $R^2$-value approximately equal to 0.64. The objective is to estimate the total $Y = \sum_{i \in U} y_i$ in each population. Note that for both populations 1 and 2, the model used to generate the data is linear, whereas it is nonlinear for population 3.

From each population, we generated $R = 10,000$ random samples of size $n = 100$ according to simple random sampling without replacement. In each sample, nonresponse to item $y$ was generated such that the response probability $p_i$ for unit $i$ is given by $\log \frac{p_i}{1-p_i} = \lambda_0 + \lambda_1 z_i$. The values of $\lambda_0$ and $\lambda_1$ were chosen to give a response rate approximately equal to 70%. The response indicators $r_i$ were then generated independently from a Bernoulli distribution with parameter $p_i$.

To compensate for the nonresponse to item $y$, we used four imputation methods: MI, RAI, SLRI, and NNI. The last three imputation methods used $z$ as the auxiliary variable. From each simulated sample, we calculated the imputed estimator $\hat{Y}_{I\pi}$. As a measure of the bias of $\hat{Y}_{I\pi}$, we used the Monte–Carlo percent relative bias given by (23). As a measure of variability of $\hat{Y}_{I\pi}$, we used the percent Monte–Carlo relative root mean square error (RRMSE) given by $\mathrm{RRMSE}_{\mathrm{MC}}\left(\hat{Y}_{I\pi}\right) = \frac{\sqrt{\mathrm{MSE}_{\mathrm{MC}}\left(\hat{Y}_{I\pi}\right)}}{Y}$, where $\mathrm{MSE}_{\mathrm{MC}}\left(\hat{Y}_{I\pi}\right)$ is given by (24).

Table 2 reports the Monte–Carlo RB and RRMSE of the resulting estimators. For population 1, we note that the imputed estimator under MI is heavily biased (24.5%), which is not surprising since the response probability and the variable of interest are both related to the variable $z$ and MI does not account for $z$. This also applies to populations 2 and 3. For population 1, it suffices to include $z$ in the imputation model (RAI, SLRI, and NNI) to reduce the bias considerably. Note that both RAI and SLRI give almost identical results in terms of RRMSE, which can be explained by the fact that, for population 1, the intercept is not significant. The NNI leads to a slightly higher RRMSE than RAI and SLRI.

For population 2, it is interesting to note that the imputed estimator under RAI is heavily biased (−16.8 %) despite the high correlation between the variables $y$ and $z$. In fact, the RB (in absolute value) is even larger than the one obtained under MI (4.2%). This result can be explained by the fact that RAI forces the intercept to go through the origin, whereas the intercept is highly significant for population 2. It suffices to add the intercept in the imputation model (SLRI) to eliminate the bias. This example illustrates that taking only the coefficient of correlation into account for choosing an imputation method can be a risky strategy.

For population 3, for which the model of $y$ given $z$ is nonlinear, the three methods MI, RAI, and SLRI lead to heavily biased estimators since, in this case, the imputation model is clearly misspecified. On the other hand, the imputed estimator under NNI is nearly

Table 2
Comparison of imputation methods

| | MI | | RAI | | SLRI | | NNI | |
|---|---|---|---|---|---|---|---|---|
| | RB | RRMSE | RB | RRMSE | RB | RRMSE | RB | RRMSE |
| Population 1 | 24.5 | 26.3 | −0.0 | 7.3 | 0.3 | 7.5 | 1.3 | 8.1 |
| Population 2 | 4.2 | 4.5 | −16.8 | 16.9 | −0.0 | 1.3 | 0.2 | 1.4 |
| Population 3 | 41.1 | 123.1 | 13.0 | 93.6 | −118.3 | 128.5 | 0.1 | 82.1 |

unbiased, which demonstrates the advantage of NNI over the parametric imputation methods such as REGI.

### *3.5. Choosing an imputation method in practice*

The results in the preceding sections clearly show that imputation is essentially a modeling exercise. Hence, the choice of an appropriate set of auxiliary variables related to the variable being imputed and/or the response propensity is a crucial step in the imputation process. Also, it is important that the imputation model accounts for sampling design features such as stratification and clustering if appropriate. In the case of stratified sampling, the strata identifiers should be included in the imputation model if they are related to the variable being imputed. In the case of cluster sampling, the use of random effect models should be considered if the intracluster correlations are appreciable. This aspect was studied by Haziza and Rao (2003), Yuan and Little (2007), and Shao (2007).

Model validation is thus an important step of the imputation process. It includes the detection of outliers or the examination of plots such as plots of residuals versus the predicted values, plots of residuals versus the auxiliary variables selected in the model, and plots of residuals versus variables not selected in the model. The choice of imputation method should be dictated by the shape of the data at hand. If the relationship between a variable of interest and a set of auxiliary variable is not linear, then NNI or a nonparametric imputation method such as nonparametric regression imputation should be considered.

In practice, the imputation method should also be chosen with respect to the type of parameter we are trying to estimate as well as the nature of the variable being imputed (continuous or categorical). For example, if we are interested in estimating a quantile, some deterministic methods such as REGI should be avoided because they tend to distort the distribution of the variables being imputed. As a result, estimators of quantiles could be heavily biased. Random imputation methods or NNI could prove useful in this case. Also, if the variable being imputed is categorical, donor imputation methods (e.g., NNI and RHDI) are preferable to avoid the possibility of impossible values in the imputed data file.

## 4. Variance of the imputed estimator

In this section, we give the variance expressions for the imputed estimator $\hat{Y}_{\text{IG}}$ under both the NM approach and the IM approach. We assume that $\hat{Y}_{\text{IG}}$ is asymptotically unbiased for $Y$.

### *4.1. Variance under the NM approach*

Using the decomposition (15), the variance of the imputed estimator $\hat{Y}_{\text{IG}}$ under deterministic imputation can be expressed as

$$E_{pq}\left(\hat{Y}_{\text{IG}} - Y\right)^2 = V_{\text{SAM}}^q + V_{\text{NR}}^q, \tag{25}$$

where $V_{\text{SAM}}^{q} = V_p\left(\hat{Y}_{\text{G}}\right)$ is the sampling variance of the prototype estimator $\hat{\text{Y}}_{\text{G}}$ and $V_{\text{NR}}^{q} = E_p V_q\left(\hat{Y}_{\text{IG}}|\mathbf{I}\right)$ is the nonresponse variance. The term $V_q\left(\hat{Y}_{\text{IG}}|\mathbf{I}\right)$ is the conditional nonresponse variance under the NM approach. In the case of random imputation, one needs to account of the variance due to the random selection of the residuals. This variance is called the imputation variance. The total variance of the imputed estimator $\hat{Y}_{\text{IG}}$ can be thus expressed as

$$E_{pq\text{I}}\left(\hat{Y}_{\text{IG}} - Y\right)^2 = V_{\text{SAM}}^{q} + V_{\text{NR}}^{q} + V_{\text{I}}^{q}, \tag{26}$$

where $V_{\text{NR}}^{q} = E_p V_q E_{\text{I}}\left(\hat{Y}_{\text{IG}}|\mathbf{I}\right)$ is the nonresponse variance and $V_{\text{I}}^{q} = E_{pq} V_{\text{I}}\left(\hat{Y}_{\text{IG}}|\mathbf{I}\right)$ is the imputation variance. The conditional nonresponse/imputation variance under the NM approach is given by $V_{q\text{I}}\left(\hat{Y}_{\text{IG}}|\mathbf{I}\right) = V_q E_{\text{I}}\left(\hat{Y}_{\text{IG}}|\mathbf{I}\right) + E_q V_{\text{I}}\left(\hat{Y}_{\text{IG}}|\mathbf{I}\right)$.

### *4.2. Variance under the IM approach*

Using the decomposition (15), the variance of the imputed estimator $\hat{Y}_{\text{IG}}$ under deterministic imputation can be expressed as

$$E_{mpq}\left(\hat{Y}_{\text{IG}} - Y\right)^2 = V_{\text{SAM}}^{m} + V_{\text{NR}}^{m} + V_{\text{MIX}}^{m}, \tag{27}$$

where $V_{\text{SAM}}^{m} = E_m V_p\left(\hat{Y}_{\text{G}}\right)$ is the anticipated sampling variance of the prototype estimator $\hat{Y}_{\text{G}}$, $V_{\text{NR}}^{m} = E_{pq} V_m\left(\hat{Y}_{\text{IG}} - \hat{Y}_{\text{G}}|\mathbf{I}, \mathbf{r}\right)$ is the nonresponse variance, and $V_{\text{MIX}}^{m} = 2E_{pq}\text{Cov}_m\left(\hat{Y}_{\text{G}} - Y, \hat{Y}_{\text{IG}} - \hat{Y}_{\text{G}}|\mathbf{I}, \mathbf{r}\right)$ is a mixed component. The term $V_m\left(\hat{Y}_{\text{IG}} - \hat{Y}_{\text{G}}|\mathbf{I}, \mathbf{r}\right)$ is the conditional nonresponse variance under the IM approach. In the case of random imputation, the variance of the imputed estimator $\hat{Y}_{\text{IG}}$ can be expressed as

$$E_{mpq\text{I}}\left(\hat{Y}_{\text{IG}} - Y\right)^2 = V_{\text{SAM}}^{m} + V_{\text{NR}}^{m} + V_{\text{MIX}}^{m} + V_{\text{I}}^{m}, \tag{28}$$

where $V_{\text{NR}}^{m} = E_{pq} V_m E_{\text{I}}\left(\hat{Y}_{\text{IG}} - Y|\mathbf{I}, \mathbf{r}\right)$ is the nonresponse variance, $V_{\text{MIX}}^{m} = 2E_{pq}\text{Cov}_m E_{\text{I}}\left(\hat{Y}_{\text{G}} - Y, \hat{Y}_{\text{IG}} - \hat{Y}|\mathbf{I}, \mathbf{r}\right)$, and $V_{\text{I}}^{m} = E_{mpq} V_{\text{I}}\left(\hat{Y}_{\text{IG}}|\mathbf{I}, \mathbf{r}\right)$ is the imputation variance. The term $V_{m\text{I}}\left(\hat{Y}_{\text{IG}} - \hat{Y}_{\text{G}}|\mathbf{I}, \mathbf{r}\right) = V_m E_{\text{I}}\left(\hat{Y}_{\text{IG}} - \hat{Y}_{\text{G}}|\mathbf{I}, \mathbf{r}\right) + E_m V_{\text{I}}\left(\hat{Y}_{\text{IG}} - \hat{Y}_{\text{G}}|\mathbf{I}, \mathbf{r}\right)$ is the conditional nonresponse/imputation variance under the IM approach.

## 5. Imputation classes

In practice, imputation is rarely done at the overall sample level. Instead, it is customary to first divide respondents and nonrespondents into classes before imputing missing values. These imputation classes are formed on the basis of auxiliary information recorded for all units in the sample. The objective in forming the classes is first to reduce the nonresponse bias and also to reduce the nonresponse variance as well as the imputation variance in the case of random imputation. There are at least two reasons motivating the formation of imputation classes instead of directly imputing the value resulting from the use of a regression model: (i) it is more convenient when it is desired to impute more than one variable at a time and (ii) it is more robust to model miss-pecification.

Suppose that the sample $s$ is partitioned into $C$ classes, $s_1, \ldots, s_c$ of sizes $n_1, \ldots, n_c$. We have $s = \bigcup_{c=1}^{C} s_c$ and $\sum_{c=1}^{C} n_c = n$. In the remainder of this section, the subscript $(ci)$ will denote unit $i$ in class $c$. We restrict ourself to the prototype estimator $\hat{Y}_{\mathrm{I}\pi}$. An imputed estimator of the population total $Y$ based on $C$ classes can be expressed as

$$\hat{Y}_{\mathrm{I}\pi,C} = \sum_{c=1}^{C} \hat{Y}_{\mathrm{I}\pi c}, \tag{29}$$

where $\hat{Y}_{\mathrm{I}\pi c} = \sum_{i \in s_c} d_{ci} r_{ci} y_{ci} + \sum_{i \in s_c} d_{ci} (1 - r_{ci})\, y_{ci}^*$ denotes the imputed estimator in class $c$, $c = 1, \ldots, C$. We consider the case of survey weighted RHDI within classes for which a missing $y$-value in class $c$ is replaced by the $y$-value of a donor selected (with replacement) from the set of respondents in class $c$ and with probability proportional to its design weight $d_{ci}$. That is, $y_{ci}^* = y_{cj}$ for $j \in s_{r_c}$ such that $P\left(y_{ci}^* = y_{cj}\right) = \frac{d_{cj}}{\sum_{l \in s_c} d_{cl} r_{cl}}$, where $s_{r_c}$ denotes the random set of respondents in class $c$.

### 5.1. Properties under the NM approach

Under the NM approach, it can be shown that the conditional bias of $\hat{Y}_{\mathrm{I}\pi,C}$ in (29) can be approximated by

$$B_{q\mathrm{I}}\left(\hat{Y}_{\mathrm{I}\pi,C} | \mathbf{I}\right) \approx \sum_{c=1}^{C} \bar{p}_c^{-1} \sum_{i \in s_c} d_{ci} \left(p_{ci} - \bar{p}_c\right) \left(y_{ci} - \bar{y}_c\right), \tag{30}$$

where, $\bar{p}_c = \sum_{i \in s_c} d_{ci} p_{ci} / \sum_{i \in s_c} d_{ci}$, and $\bar{y}_c = \sum_{i \in s_c} d_{ci} y_{ci} / \sum_{i \in s_c} d_{ci}$. From (30), it follows that the bias is approximately equal to zero if the sample covariance between the response probability and the variable of interest is approximately equal to zero in each class. This is satisfied, for example, when the classes are homogeneous with respect to the response probabilities and/or the variable of interest. In practice, classes will be formed with respect to estimated response probabilities or with respect to substitute variable closely related to $y$ (see Section 5.3).

Next, we turn to the conditional nonresponse/imputation variance of $\hat{Y}_{\mathrm{I}\pi,C}$ under the NM approach that can be approximated by

$$\begin{aligned} V_{q\mathrm{I}}\left(\hat{Y}_{\mathrm{I}\pi,C} | \mathbf{I}\right) \approx{} & \sum_{c=1}^{C} \bar{p}_c^{-2} \sum_{i \in s_c} d_{ci}^2 p_{ci} (1 - p_{ci}) \left(y_{ci} - \bar{y}_c^{(p)}\right)^2 \\ & + \sum_{c=1}^{C} \sum_{i \in s_c} d_{ci}^2 (1 - r_{ci})\, s_{rc}^2, \end{aligned} \tag{31}$$

where $\bar{y}_c^{(p)} = \sum_{i\in s_c} d_{ci}p_{ci}y_{ci}/\sum_{i\in s_c} d_{ci}p_{ci}$ and $s_{rc}^2 = \frac{1}{\sum_{i\in s_c} d_{ci}r_{ci}} \sum_{i\in s_c} d_{ci}r_{ci}\,(y_{ci} - \bar{y}_{rc})^2$ with $\bar{y}_{rc} = \sum_{i\in s_c} d_{ci}r_{ci}y_{ci}/\sum_{i\in s_c} d_{ci}r_{ci}$. Note that the first term on the right-hand side of (31) is the conditional nonresponse variance, whereas the second term is the imputation variance due to RHDI. It is clear from (31) that the nonresponse variance and the imputation variance can be reduced by forming imputation classes that are homogeneous with respect to the variable of interest since the terms $\left(y_{ci} - \bar{y}_c^{(p)}\right)$ and $s_{rc}^2$ are small in this case. It is not clear, however, that forming imputation classes homogeneous with respect to the response probabilities will help in reducing the nonresponse or the imputation variance.

### *5.2. Properties under the IM approach*

We assume the following general imputation model:

$$\begin{aligned} y_i &= \mu_i + \epsilon_i \\ E_m(\epsilon_i) = 0,\ E_m&\left(\epsilon_i\epsilon_j\right) = 0 \quad \text{if} \quad i \neq j \text{ and } V_m(\epsilon_i) = \sigma_i^2. \end{aligned} \tag{32}$$

Under the IM approach and model (32), it can be shown that the conditional bias of $\hat{Y}_{\mathrm{I}\pi,C}$ in (29) is given by

$$B_{m\mathrm{I}}\left(\hat{Y}_{\mathrm{I}\pi,C}|\mathbf{I}\right) = -\sum_{c=1}^{C}\sum_{i\in s_c} d_{ci}\left(\mu_{ci} - \bar{\mu}_{r_c}\right), \tag{33}$$

where $\bar{\mu}_{r_c} = \sum_{i\in s_c} d_{ci}r_{ci}\mu_{ci}/\sum_{i\in s_c} d_{ci}r_{ci}$. The bias in (33) vanishes if $\mu_{ci}$ is constant within each imputation class, which corresponds to the model underlying RHDI within classes (or MI within classes). Hence, the objective will be to form classes that are homogeneous with respect to $\mu_i$. Since $\mu_i$ is typically unknown, classes will be made homogeneous with respect to the substitute variable $\hat{\mu}_i$.

Under the IM approach, the conditional nonresponse/imputation variance of $\hat{Y}_{\mathrm{I}\pi,C}$ is given by

$$\begin{aligned} V_{m\mathrm{I}}\left(\hat{Y}_{\mathrm{I}\pi,C} - \hat{Y}_\pi|\mathbf{I},\mathbf{r}\right) &= \sum_{c=1}^{C}\hat{p}_c^{-2}\left[\hat{p}_c^2\sum_{i\in s_c} d_{ci}^2(1 - r_{ci})\,\sigma_{ci}^2 + \left(1 - \hat{p}_c^2\right)\sum_{i\in s_c} d_{ci}^2 r_{ci}\sigma_{ci}^2\right] \\ &\quad + \sum_{c=1}^{C}\sum_{i\in s_c} d_{ci}^2(1 - r_{ci})\,s_{rc}^2, \end{aligned} \tag{34}$$

where $\hat{p}_c = \sum_{i\in s_c} d_{ci}r_{ci}/\sum_{i\in s_c} d_{ci}$ is the weighted response rate in class $c$. The variance in (34) will be small if the model variances $\sigma_{ci}^2$ are small, which occurs when the imputation model is highly predictive.

### *5.3. Construction of classes*

In practice, several methods are used to form imputation classes. Here, we consider two methods: (i) the cross-classification method and (ii) the score method (sometimes called predictive mean matching or response propensity stratification, depending on the context).

The cross-classification method consists of cross-classifying categorical variables and is widely used in practice. If the auxiliary variables chosen to form the classes are related to the response probability and/or to the variable of interest, then the imputation classes will likely help in reducing the nonresponse bias. However, this method can lead to a huge number of classes. For example, cross-classifying 8 variables, each with 5 categories, leads to the creation of 390,625 classes. As a result, a large number of classes may contain few or no observations, which could potentially lead to unstable estimators. In practice, it is customary to specify certain constraints to ensure the stability of the resulting estimators. For example, we can specify that the number of respondents within a class must be greater than or equal to a certain level. On the other hand, we can also specify, that, within a class, the proportion of respondents must be greater than or equal to a certain level. If the constraints are not met, classes are generally collapsed by, for example, eliminating one of the auxiliary variables and cross-classifying the remaining variables. If the constraints are too severe, a large number of auxiliary variables may have to be dropped to satisfy the constraints, which in turn may result in a relatively poor (nonresponse or imputation) model. As a result, the nonresponse bias may not have been reduced to the maximum extent. Also, the cross-classification requires a proper ordering of the auxiliary variables that will determine which variable will be dropped first, which variable will be dropped second, and so on. Finally, since a respondent may be used as a donor in several stages of the process, the resulting classes are not disjoint. As a result, application of the available variance estimation methods (see Section 6) is not straightforward because of the collapsing of the classes at each stage of the process. In practice, it is customary to treat the classes as if they were disjoint. This method was studied by Haziza and Beaumont (2007).

The score method consists of first estimating the response probabilities $p_i$ by $\hat{p}_i, i \in s$ using the assumed nonresponse model, or estimating the conditional means $\mu_i$ by $\hat{\mu}_i$, $i \in s$ using the assumed imputation model. The scores $\hat{p}$ and $\hat{\mu}$ may be seen as a summary of the information contained in the auxiliary variables related to the response probability and the variable of interest, respectively. Using one of the two scores $\hat{p}$ or $\hat{\mu}$, partition the sample according to an equal quantile method (which consists of first ordering the observations according to the selected score and partitioning the resulting sample into $C$ classes of approximately equal size) or a classification algorithm. The resulting classes are then homogeneous with respect to the chosen score $\hat{p}$ or $\hat{\mu}$. The score method has been studied in the context of weighting for unit nonresponse by Eltinge and Yansaneh (1997), Little (1986), and in the context of imputation, by Haziza and Beaumont (2007). Results from numerous simulation studies show that, unlike the cross-classification method, the score method requires a relatively small number of classes (typically between 5 and 50) to achieve a significant nonresponse bias reduction. Therefore, the number of respondents per class is typically large, which ensures the stability of the resulting estimators, and no ordering of the auxiliary variables is needed. Predictive mean matching (Little, 1988) can be seen as the limit case of the score method when the number of classes is equal to the overall number of respondents in the sample. Finally, note that it is possible to form the classes so that they are simultaneously homogeneous with respect to both scores $\hat{p}$ and $\hat{\mu}$. The resulting estimators are then doubly robust in the sense that they are still valid even if one model or the other is misspecified. In the context of weighting for unit nonresponse, the simultaneous use of both scores was studied by Smith et al. (2004) and Vartivarian and Little (2002). In the context of imputation, it was studied by Haziza and Beaumont (2007).

## 6. Variance estimation

Variance estimation in the presence of single imputation has been widely treated in the literature. The reader is referred to Lee et al. (2002), Rao (1996), and Shao (2002) for an overview on the topic. Before the 1990's, it was customary to treat the imputed values as if they were observed values. Nowadays, surveys are increasingly using variance estimation methods designed to handle nonresponse and imputation. Failing to account for the nonresponse and imputation will result in variances (or coefficients of variation) typically too small and inferences (e.g., confidence intervals or tests of hypothesis) will be potentially misleading, especially if the response rates are low. For example, a 95% confidence interval for the population total $Y$ is given by

$$\hat{Y}_{\mathrm{IG}} \pm 1.96\sqrt{v\left(\hat{Y}_{\mathrm{IG}}\right)}, \tag{35}$$

where $v\left(\hat{Y}_{\mathrm{IG}}\right)$ denotes an estimator of the variance of $\hat{Y}_{\mathrm{IG}}$. It is well known that the confidence interval (35) is valid if the following criteria are met as follows: (i) the asymptotic distribution of $\hat{Y}_{\mathrm{IG}}$ is normal; (ii) the estimator $\hat{Y}_{\mathrm{IG}}$ is unbiased (or asymptotically unbiased) for $Y$, and (iii) the variance estimator $v\left(\hat{Y}_{\mathrm{IG}}\right)$ is consistent for the true variance of $\hat{Y}_{\mathrm{IG}}$. If one of the three criteria is not satisfied, then the coverage probability of the confidence interval (35) may be considerably different than 95%. In the presence of imputed data, the criterion (i) is often satisfied (see e.g., Rao and Shao, 1992). As we discussed in Section 2.3, the criterion (ii) is only met if the assumed (nonresponse or imputation) model is valid. Finally, the criterion (iii) is clearly not met if standard variance estimation methods (i.e., methods that treat the imputed values as observed values) are used. In this case, the coverage probability of the confidence interval (35) may be considerably smaller than 95% if the nonresponse rate is appreciable.

We distinguish between two frameworks for variance estimation: (i) the customary two-phase framework (TPF) and (ii) the reverse framework (RF). In the TPF, nonresponse is viewed as a second phase of selection. First, a random sample is selected from the population according to a given sampling design. Then, the set of respondents is generated according to the nonresponse mechanism. In the RF, the order of sampling and response is reversed. First, the population is randomly divided into a population of respondents and a population of nonrespondents according to the nonresponse mechanism. Then, a random sample is selected from the population (containing respondents and nonrespondents) according to the sampling design. The RF usually facilitates the derivation of variance estimators, but unlike the TPF, it requires the additional assumption that the nonresponse mechanism does not depend on which sample is selected. This assumption is satisfied in many situations encountered in practice. On the other hand, the TPF leads to a natural decomposition of the total variance. That is, the total variance can be expressed as the sum of the sampling variance and the nonresponse variance which allows the survey statistician to get an idea of the relative magnitude of each component. Under the RF, there is no easy interpretation of the variance components.

For each framework, inference can be based either on an IM approach or a NM approach. The IM approach requires the validity of an imputation model, whereas the NM approach requires the validity of a nonresponse model. We assume that the imputed estimator $\hat{Y}_{\mathrm{IG}}$ is asymptotically unbiased for $Y$, so bias is not an issue here. Note that the response indicators $r_i$ must be included in the imputed data file for variance estimation purposes. We consider the case of a single imputation class but the extension to multiple classes is relatively straightforward.

### *6.1. The two-phase framework under the IM approach*

Särndal (1992) proposed a variance estimation method under the IM approach and illustrated it under RAI and simple random sampling without replacement. Deville and Särndal (1994) extended the method to the case of an arbitrary design and deterministic REGI. In both papers, the authors considered the expansion estimator, $\hat{Y}_\pi$, as the prototype estimator. Under the IM approach and deterministic imputation, the total variance of the imputed estimator is given by (27). In this section, we consider the case of the imputed estimator $\hat{Y}_{\mathrm{IG}}$. The estimation of $V^m_{\mathrm{SAM}}$, $V^m_{\mathrm{NR}}$, and $V^m_{\mathrm{MIX}}$ may be performed as follows:

(i) To estimate $V^m_{\mathrm{SAM}}$, it suffices to estimate $V_p\left(\hat{Y}_{\mathrm{G}}\right)$. Let $\hat{V}_{\mathrm{SAM}}$ be an asymptotically $p$-unbiased complete data variance estimator of $V_p\left(\hat{Y}_{\mathrm{G}}\right)$. Also, let $\hat{V}_{\mathrm{ORD}}$ be the "naive" variance estimator of $\hat{Y}_{\mathrm{IG}}$, that is, the variance estimator obtained by treating the imputed values as if they were observed. It is well known that for several imputation methods (in particular, the deterministic methods), $\hat{V}_{\mathrm{ORD}}$ is a biased estimator of $V^m_{\mathrm{SAM}}$. In most cases, $\hat{V}_{\mathrm{ORD}}$ underestimates $V^m_{\mathrm{SAM}}$. To compensate for this underestimation, Särndal (1992) proposed to evaluate the following expectation, $E_m\left(\hat{V}_{\mathrm{SAM}} - \hat{V}_{\mathrm{ORD}}|\mathbf{I}, \mathbf{r}\right) \equiv V_{\mathrm{DIF}}$. Then, determine a $m$-unbiased estimator, denoted by $\hat{V}^m_{\mathrm{DIF}}$, of $V_{\mathrm{DIF}}$. This will usually require the estimation of certain parameters of the assumed imputation model. Finally, a $mpq$-unbiased estimator of $V^m_{\mathrm{SAM}}$ is given by $\hat{V}^m_{\mathrm{SAM}} = \hat{V}_{\mathrm{ORD}} + \hat{V}^m_{\mathrm{DIF}}$.

(ii) To estimate $V^m_{\mathrm{NR}}$, it suffices to find a $m$-unbiased estimator, denoted by $\hat{V}^m_{\mathrm{NR}}$, of $V_m\left(\hat{Y}_{\mathrm{IG}} - \hat{Y}_G|\mathbf{I}, \mathbf{r}\right)$. Again, this will require the estimation of unknown parameters of the imputation model $m$. We have $E_m\left(\hat{V}^m_{\mathrm{NR}}|\mathbf{I}, \mathbf{r}\right) = V_m\left(\hat{Y}_{\mathrm{IG}} - \hat{Y}_{\mathrm{G}}|\mathbf{I}, \mathbf{r}\right)$. It follows that $\hat{V}^m_{\mathrm{NR}}$ is an asymptotically $mpq$-unbiased estimator of $V^m_{\mathrm{NR}}$.

(iii) To estimate $V^m_{\mathrm{MIX}}$, it suffices to find a $m$-unbiased estimator, denoted by $\hat{V}^m_{\mathrm{MIX}}$, of $\mathrm{Cov}_m\left(\hat{Y}_{\mathrm{G}} - Y, \hat{Y}_{\mathrm{IG}} - \hat{Y}_{\mathrm{G}}|\mathbf{I}, \mathbf{r}\right)$. Again, this will require the estimation of unknown parameters of the imputation model $m$. As a result, the estimator $\hat{V}_{\mathrm{MIX}}$ is asymptotically $mpq$-unbiased for $V^m_{\mathrm{MIX}}$.

Finally, an asymptotically $mpq$-unbiased estimator of the total variance, $V\left(\hat{Y}_{\mathrm{IG}}\right)$, denoted by $\hat{V}^m_{\mathrm{TP}}$, is given by

$$\hat{V}^m_{\mathrm{TP}} = \hat{V}_{\mathrm{ORD}} + \hat{V}^m_{\mathrm{DIF}} + \hat{V}^m_{\mathrm{NR}} + \hat{V}^m_{\mathrm{MIX}}.$$

We now make some remarks on Särndal's method, which are as follows:

(a) Unlike most deterministic methods, the naive variance estimator $\hat{V}_{\mathrm{ORD}}$ is asymptotically unbiased for $V^m_{\mathrm{SAM}}$ for random imputation methods (e.g., random REGI) and the derivation of $\hat{V}_{\text{DIF}}^{m}$ may be omitted in this case. For deterministic methods, the derivation of $\hat{V}_{\text{DIF}}^{m}$ for an arbitrary design involves tedious algebra. An alternative that does not require any derivation involves the construction of a new set of imputed values. It consists of adding a randomly selected residual to the imputed values obtained under the deterministic method. Then, use a standard variance estimator valid in the complete response case using the new imputed values. Let $\hat{V}_{\text{ORD}}^{*}$ denote the resulting variance estimator. It can be shown that $\hat{V}_{\text{ORD}}^{*}$ is an asymptotically $mpqI$-unbiased estimator of $V_{\text{SAM}}^{m}$. Note that this new set of imputed values is used only to obtain a valid estimator of the sampling variance and is not used to estimate the parameter of interest $Y$. In practice, one could, for example, create a variance estimation file containing the new set of imputed values and use standard variance estimation systems (used in the complete data case) to obtain an estimate of the sampling variance.

(b) In the case of self-weighting unistage designs and REGI, the component $\hat{V}_{\text{MIX}}^{m}$ is exactly equal to 0 when the prototype estimator is the expansion estimator $\hat{Y}_{\pi}$. Even when it is not exactly zero, Deville and Särndal (1994) argue that this component is typically much smaller than the terms $\hat{V}_{\text{SAM}}^{m}$ and $\hat{V}_{\text{NR}}^{m}$, so it may be omitted in the computation of the total variance. However, Brick et al. (2004) showed that in the case of unequal probability designs, the contribution (positive or negative) of $\hat{V}_{\text{MIX}}^{m}$ to the total variance may be important. Also, Beaumont et al. (2007) show that under AVI, the component $\hat{V}_{\text{MIX}}^{m}$ is always negative and its contribution to the total variance may be considerable. Thus, the computation of this component should not be omitted, in general.

(c) The variance components $\hat{V}_{\text{DIF}}^{m}$, $\hat{V}_{\text{NR}}^{m}$, and $\hat{V}_{\text{MIX}}^{m}$ are derived under the selected imputation model. Hence, their validity depends on the validity of the assumed model. For example, under REGI, one must correctly specify the vector of auxiliary variable $\mathbf{z}$ as well as the variance structure $\sigma_i^2$. In other words, both the first and the second moments of the imputation model must be correctly specified to ensure that the resulting variance estimators are asymptotically valid. Modeling the variance structure may prove to be difficult in practice. To overcome this problem, it could be estimated nonparametrically by using, for example, the respondents $y$-values and penalized least squares estimation (Beaumont et al., 2007).

(d) Unlike replication methods (see Section 6.4), the method is not computer intensive.

(e) The method can be applied for more complex parameters such as the ratio of two totals, where both variables involved may be missing (Haziza, 2007). The application of the method for nonsmooth parameters (e.g., median) has not been yet studied.

We now discuss variance estimation for $\hat{Y}_{\text{IG}}$ under weighted deterministic REGI for which the imputed values are given by (6) with $\omega_i = d_i$. An estimator of $V_{\text{SAM}}^{m}$ is obtained by first creating a new set of imputed values under weighted random REGI. That is, the missing values are replaced by the imputed values given in (12). Then, an asymptotically unbiased estimator of $V_{\text{SAM}}^{m}$ is given by $\hat{V}_{\text{SAM}}^{m} = \hat{V}_{\text{ORD}}^{*}$, which is a complete data variance estimator (i.e., the variance estimator that treats the new imputed values as if they were observed). To obtain the variance components $\hat{V}_{\text{NR}}^{m}$ and $\hat{V}_{\text{MIX}}^{m}$, we first express the imputed estimator $\hat{Y}_{\text{IG}}$ as

$$\hat{Y}_{\text{IG}} = \sum_{i \in s} w_i^* r_i y_i,$$

where $w_i^* = d_i \left[ g_i + \left( \hat{\mathbf{Z}}_{\text{G}} - \hat{\mathbf{Z}}_{r\text{G}} \right)' \hat{\mathbf{T}}_r^{-1} \frac{\mathbf{z}_i}{(\boldsymbol{\lambda}'\mathbf{z}_i)} \right]$ with $\hat{\mathbf{Z}}_{\text{G}} = \sum_{i \in s} w_i \mathbf{z}_i$, $\hat{\mathbf{Z}}_{r\text{G}} = \sum_{i \in s} w_i r_i \mathbf{z}_i$, $\hat{\mathbf{T}}_r = \sum_{i \in s} d_i r_i \mathbf{z}_i \mathbf{z}_i' / \left( \lambda' \mathbf{z}_i \right)$, and $g_i$ is given by (3). Expressing the imputed estimator $\hat{Y}_{\text{IG}}$ as a weighted sum of the $y$-values simplifies the derivation of $\hat{V}_{\text{NR}}^m$ and $\hat{V}_{\text{MIX}}^m$ considerably. We have

$$\hat{V}_{\text{NR}}^m = \hat{\sigma}^2 \left[ \sum_{i \in s} \left( w_i^* - w_i \right)^2 r_i \left( \boldsymbol{\lambda}'\mathbf{z}_i \right) + \sum_{i \in s} w_i^2 (1 - r_i) \left( \boldsymbol{\lambda}'\mathbf{z}_i \right) \right]$$

and

$$\hat{V}_{\text{MIX}}^m = \hat{\sigma}^2 \sum_{i \in s} (w_i - 1) \left( w_i^* r_i - w_i \right) r_i \left( \boldsymbol{\lambda}'\mathbf{z}_i \right),$$

where $\hat{\sigma}^2$ is an estimator of $\sigma^2$. A simple but slightly biased estimator of $\sigma^2$ is given by $\hat{\sigma}^2 = \frac{\sum_{i \in s} r_i \left( y_i - \mathbf{z}_i' \hat{\mathbf{B}}_r \right)}{\sum_{i \in s} r i \left( \boldsymbol{\lambda}'\mathbf{z}_i \right)}$ (Deville and Särndal 1994). Note that the component $\hat{V}_{\text{MIX}}^m$ is not equal to zero, in general, even for self-weighting designs. Finally, under mild regularity conditions, note that the three variance components $\hat{V}_{\text{SAM}}^m$, $\hat{V}_{\text{NR}}^m$, and $\hat{V}_{\text{MIX}}^m$, are all of order $O_p \left( N^2/n \right)$, so they all need to be computed to obtain a valid estimator of the total variance.

### *6.2. The two-phase framework under the NM approach*

The NM approach was studied by Rao (1990) and Rao and Sitter (1995) in the context of simple random sampling without replacement and by Beaumont (2005) in the case of arbitrary designs. Under AVI, it was studied by Beaumont et al. (2007). Under deterministic imputation, the total variance of the imputed estimator is given by (25). Both components $V_{\text{SAM}}^q$ and $V_{\text{NR}}^q$ can be estimated unbiasedly by using, for example, a Taylor linearization procedure. Note that to estimate $V_{\text{NR}}^q$, it suffices to estimate $V_q \left( \hat{Y}_{\text{IG}} | \mathbf{I} \right)$. An estimator of the total variance $V \left( \hat{Y}_{\text{IG}} \right)$ is given by

$$\hat{V}_{\text{TP}}^q = \hat{V}_{\text{SAM}}^q + \hat{V}_{\text{NR}}^q.$$

Both $\hat{V}_{\text{SAM}}^q$ and $\hat{V}_{\text{NR}}^q$ are of order $O_p \left( N^2/n \right)$, so both terms need to be computed to obtain a valid estimator of the total variance.

### *6.3. The reverse framework*

The RF was proposed by Fay (1991) and the variance estimation method under this framework was developed by Shao and Steel (1999). Recall that we assume that the response probability does not depend on the realized sample $s$.

### *6.3.1. The NM approach*

Under the NM approach and deterministic imputation, the total variance of the imputed estimator $\hat{Y}_{\mathrm{IG}}$ can be expressed as

$$V\left(\hat{Y}_{\mathrm{IG}}\right) = E_q V_p\left(\hat{Y}_{\mathrm{IG}}|\mathbf{r}\right) + V_q E_p\left(\hat{Y}_{\mathrm{IG}}|\mathbf{r}\right). \tag{36}$$

A variance estimator is obtained by separately estimating the two terms on the right-hand side of (36).

(i) To estimate the component $E_q V_p\left(\hat{Y}_{\mathrm{IG}}|\mathbf{r}\right)$, it suffices to estimate $V_p\left(\hat{Y}_{\mathrm{IG}}|\mathbf{r}\right)$, which is the variance due to sampling conditional on the vector of response indicators $\mathbf{r}$. This component, denoted by $\hat{V}_1$, is readily obtained by using any standard variance estimation technique available in the complete data case since the response indicator $r_i$ can now be seen as a characteristic of unit $i$. For example, Taylor linearization or replication methods such as the jackknife or the bootstrap can be used.

(ii) The estimation of the component $V_q E_p\left(\hat{Y}_{\mathrm{IG}}|\mathbf{r}\right)$ will require the estimation of the response probabilities $p_i$. The estimator, denoted by $\hat{V}_2^q$, can be obtained using Taylor linearization.

An estimator of the total variance under the NM approach is thus given by

$$\hat{V}_R^q = \hat{V}_1 + \hat{V}_2^q.$$

### *6.3.2. The IM approach*

Under the IM approach and deterministic imputation, the total variance of the imputed estimator $\hat{Y}_{\mathrm{IG}}$ can be expressed as

$$V\left(\hat{Y}_{\mathrm{IG}} - Y\right) = E_{mq} V_p\left(\hat{Y}_{\mathrm{IG}} - Y|\mathbf{r}\right) + E_q V_m E_p\left(\hat{Y}_{\mathrm{IG}} - Y|\mathbf{r}\right) + V_q E_{mp}\left(\hat{Y}_{\mathrm{IG}} - Y|\mathbf{r}\right). \tag{37}$$

Noting that $E_{mp}\left(\hat{Y}_{\mathrm{IG}} - Y|\mathbf{r}\right) \approx 0$, the third term on the right-hand side of (37) is much smaller than the other two, so we omit it from the calculations. To estimate $E_{mq} V_p\left(\hat{Y}_{\mathrm{IG}} - Y|\mathbf{r}\right)$, it suffices to estimate $V_p\left(\hat{Y}_{\mathrm{IG}} - Y|\mathbf{r}\right)$ as in the case of the NM approach, which leads to $\hat{V}_1$. To estimate $E_q V_m E_p\left(\hat{Y}_{\mathrm{IG}} - Y|\mathbf{r}\right)$, it suffices to estimate $V_m E_p\left(\hat{Y}_{\mathrm{IG}}|\mathbf{r}\right)$, which will require the estimation of certain parameters of the imputation model.

An estimator of the total variance under the IM approach is thus given by

$$\hat{V}_R^m = \hat{V}_1 + \hat{V}_2^m.$$

### *6.3.3. Some remarks on the reverse framework*

We make the following remarks:

(a) The variance component $\hat{V}_1$ is identical for both the IM and the NM approaches and its validity does not depend on the validity of the assumed (nonresponse or imputation) model. As a result, the component $\hat{V}_1$ is doubly robust in the sense that it is valid under either approach. The validity of the components $\hat{V}_2^q$ and $\hat{V}_2^m$ depend on the validity of the assumed models.

(b) Under mild regularity conditions, the component $\hat{V}_1$ is of order $O_p\big(N^2/n\big)$, whereas the components $\hat{V}_2^q$ and $\hat{V}_2^m$ are of order $O_p(N)$. The contribution of $\hat{V}_2$ to the total variance, $\frac{\hat{V}_2}{\hat{V}_1+\hat{V}_2}$ is thus of order $O_p\big(\frac{n}{N}\big)$, where $\hat{V}_2$ denotes either $\hat{V}_2^q$ or $\hat{V}_2^m$. As a result, the contribution of $\hat{V}_2$ is negligible when the sampling fraction, $n/N$, is negligible, in which case its computation may be omitted. The total variance of $\hat{Y}_{\mathrm{IG}}$ can then be estimated by $\hat{V}_1$.

(c) As we argue in Section 6.4, both the Rao–Shao jackknife (Rao and Shao, 1992) and the Shao–Sitter bootstrap (Shao and Sitter, 1996) find their justification under the RF since they both attempt to estimate the component $V_p\Big(\hat{Y}_{\mathrm{IG}}|\mathbf{r}\Big)$.

### *6.4. Replication methods*

In the preceding sections, we considered the linearization technique to obtain asymptotically valid variance estimators. In this section, we consider the use of replication variance estimation methods in the context of imputation for item nonresponse. Unlike Taylor linearization procedures, replication methods neither require separate derivation for each particular estimator nor require joint inclusion probabilities that may be difficult to obtain for complex designs. For a overview on replication methods in the absence of nonresponse, the reader is referred to Wolter (2007). In the presence of imputed data, several replication methods have been studied in the literature (Davison and Sardy, 2007; Shao, 2002). In this section, we focus on two methods: (i) the jackknife and (ii) the bootstrap.

We consider stratified multistage sampling designs. The population under consideration is stratified into $L$ strata with $N_h$ primary sampling units (PSU's) or clusters in the $h$th stratum. Within each stratum, $n_h \geq 2$, clusters are selected from stratum $h$, independently across strata. The first-stage clusters are usually selected without replacement to avoid the selection of the same cluster more than once. Within the $(hi)$th sampled first-stage cluster, $m_{hi}$ ultimate units (elements) are sampled according to some probability sampling method, $i = 1, \ldots, n_h$; $h = 1, \ldots, L$. Note that we do not need to specify the number of stages or the sampling methods beyond the first stage. We simply assume that subsampling within sampled clusters is performed to ensure unbiased estimation of cluster totals, $Y_{hi}$. We denote the $k$th sampled element in the $i$th sampled cluster of the $h$th stratum as $(hik)$, $k = 1, \ldots, m_{hi}$; $i = 1, \ldots, n_h$; $h = 1, \ldots, L$. Let $d_{hik}$ denote the design weights attached to the sample element $(hik)$. Using the design weights, an estimator of the population total $Y$ is $\hat{Y} = \sum_{(hik)\in s} d_{hik} y_{hik}$, where $s$ denote the total sample of elements $(hik)$.

At the variance estimation stage, it is a common practice to treat the sample as if the first-stage clusters are drawn with replacement to simplify the derivation of variance estimators. Typically, this approximation leads to overestimation of the true variance when the first-stage clusters are selected without replacement, but the bias will be small if the overall first-stage sampling fraction $\sum_{h=1}^{L} n_h / \sum_{h=1}^{L} N_h$ is small (Shao, 2002).

### *6.4.1. The Jackknife*

The jackknife method consists in calculating a set of replicate estimates, derived from a subset of the sample data, and in estimating the variance using the replicate estimates. In the absence of nonresponse, we proceed according to the following steps:

(i) remove one cluster, say (gj);

(ii) adjust the design weights $d_{hik}$ to obtain the so-called jackknife weights $d_{hik}b_{(gj)}$, where $b_{(gj)}$ is an adjustment factor we apply when cluster $(gj)$ has been deleted such that

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

(iii) compute the estimator $\hat{Y}_{\mathrm{G}}^{(gj)}$, which is calculated the same way as $\hat{Y}_{\mathrm{G}}$ but using the adjusted weights $d_{hik}b_{(gj)}$ instead of the design weights $d_{hik}$;

(iv) insert back the cluster deleted in step (i) and delete the next cluster;

(v) repeat the steps (i)–(iv) until all the clusters have been deleted.

A jackknife variance estimator of $\hat{Y}_{\mathrm{G}}$ is then given by

$$v_{\mathrm{J}}\left(\hat{Y}_{\mathrm{G}}\right) = \sum_{g=1}^{L}\left(\frac{n_g - 1}{n_g}\right)\sum_{j=1}^{n_h}\left(\hat{Y}_{\mathrm{G}}^{(gj)} - \hat{Y}_{\mathrm{G}}\right)^2. \tag{38}$$

This method is called delete-cluster jackknife. In the presence of nonresponse to item $y$, the use of (38) may lead to serious underestimation of the variance of the estimator, especially, if the nonresponse rate is appreciable. Rao and Shao (1992) proposed an *adjusted jackknife* variance estimator that may be applied in the case of deterministic or random REGI. Under weighted random REGI, the Rao–Shao-adjusted jackknife is calculated in the usual way except that whenever the $(gj)$th cluster is deleted, the imputed values, $y^*_{hik}$, are adjusted to $y^{*(gj)}_{hik} = y^*_{hik} + \left[E_{\mathrm{I}(gj)}\left(y^*_{hik}|\mathbf{I}, \mathbf{r}\right) - E_{\mathrm{I}}\left(y^*_{hik}|\mathbf{I}, \mathbf{r}\right)\right]$, where $E_{\mathrm{I}(gj)}(.|\mathbf{I}, \mathbf{r})$ denotes the expectation with respect to the imputation mechanism when the $(gj)$th cluster is deleted. The adjusted imputed values $y^{*(gj)}_{hik}$ reflect the fact that the donor set is changed when a cluster is deleted from the sample. In the case of weighted deterministic REGI, $y^{*(gj)}_{hik}$ reduces to $\mathbf{z}'_{hik}\hat{\mathbf{B}}_r^{(gj)}$, where $\hat{\mathbf{B}}_r^{(gj)}$ is computed the same way as $\hat{\mathbf{B}}_r$ given by (7) with $d_{hik}b_{(gj)}$ instead of $d_{hik}$. In this case, applying the adjustment is equivalent to reimputing missing values in the replicates obtained by deleting the $(gj)$th cluster, using the donors in that replicate. Using the adjusted imputed values, the Rao–Shao jackknife variance estimator is given by

$$v_{\mathrm{JRS}}\left(\hat{Y}_{\mathrm{IG}}\right) = \sum_{g=1}^{L}\left(\frac{n_g - 1}{n_g}\right)\sum_{j=1}^{n_h}\left(\hat{Y}_{\mathrm{IG}}^{a(\mathrm{gj})} - \hat{Y}_{\mathrm{IG}}\right)^2, \tag{39}$$

where $\hat{Y}_{\mathrm{IG}}^{a(gj)}$ is computed the same way as $\hat{Y}_{\mathrm{IG}}^{(gj)}$ but with the adjusted imputed values $y^{*(gj)}_{hik}$ instead of the imputed values $y^*_{hik}$ (Yung and Rao, 2000).

### *6.4.2. The bootstrap*

The bootstrap method proposed by Efron (1979) is a useful replication method for obtaining variance estimators for complex parameters. Lahiri (2003) and Shao

(2003) give comprehensive overviews of the boostrap in the context of survey sampling.

In the case of complete response, Rao and Wu (1988) proposed a rescaling bootstrap procedure for stratified multistage designs. Applying the Rao–Wu bootstrap in the presence of missing responses and treating the missing values as true values may lead to serious underestimation of the variance of the estimator. Under imputation for missing data, Shao and Sitter (1996) proposed a rescaling bootstrap procedure for imputed survey data that may be described as follows:

(i) Draw a simple random sample with replacement, $s^*$ of size $n^* = n - 1$ from $s$ after imputation. The sample $s^*$ is often called a bootstrap sample.

(ii) Let $r^*_{hik}$ be the response indicator for element $(hik)$ in $s^*$. Let $s^*_r = \{(hik) \in s^* : r^*_{hik} = 1\}$ and let $s^*_m = \{(hik) \in s^* : r^*_{hik} = 0\}$. Apply the same imputation procedure used to obtain the imputed estimator $\hat{Y}_{\mathrm{IG}}$ for reimputing the nonrespondent values in $s^*$ (i.e., the missing values in $s^*_m$, using the bootstrap donor set $s^*_r$).

(iii) Compute the imputed estimator, $\hat{Y}^*_{\mathrm{IG}}$, from the imputed data in (ii).

(iv) Repeat (i)–(iii) $B$ times to get $\hat{Y}^{*(1)}_{\mathrm{IG}}, \ldots, \hat{Y}^{*(B)}_{\mathrm{IG}}$.

A bootstrap variance estimator is then given by

$$v_{\mathrm{BSS}}\left(\hat{Y}_{\mathrm{IG}}\right) = \frac{1}{(B-1)} \sum_{b=1}^{B} \left(\hat{Y}^{*(b)}_{\mathrm{IG}} - \bar{\hat{Y}}_{\mathrm{IG}}\right)^2, \tag{40}$$

where $\bar{\hat{Y}}_{\mathrm{IG}} = \frac{1}{B}\sum_{b=1}^{B} \hat{Y}^{*(b)}_{\mathrm{IG}}$. When the within-stratum cluster sample sizes, $n_h$, are small (say, $n_h = 2$), $v_{\mathrm{BSS}}\left(\hat{Y}_{\mathrm{IG}}\right)$ may be heavily biased. Saigo et al. (2001) modified the Shao–Sitter procedure to overcome this difficulty. Shao and Sitter (1996) also discussed the without replacement bootstrap (e.g., Sitter, 1992a) and the mirror-match bootstrap (Sitter, 1992b).

### *6.4.3. Some remarks on the replication methods*

We make the following remarks:

(a) Both $v_{\mathrm{JRS}}\left(\hat{Y}_{\mathrm{IG}}\right)$ and $v_{\mathrm{BSS}}\left(\hat{Y}_{\mathrm{IG}}\right)$ estimate the first term on the right-hand side of (36), $E_q V_p\left(\hat{Y}_{\mathrm{IG}}|\mathbf{r}\right)$, but not the second term $V_q E_p\left(\hat{Y}_{\mathrm{IG}}|\mathbf{r}\right)$, and both estimators are asymptotically equivalent to the variance estimator we would have obtained if the clusters were selected with replacement. Therefore, if the overall sampling fraction $\sum_{h=1}^{L} n_h / \sum_{h=1}^{L} N_h$ is negligible, both variance estimators are asymptotically unbiased for $E_q V_p\left(\hat{Y}_{\mathrm{IG}}|\mathbf{r}\right)$. Also, noting that, in this case, we can omit the derivation of the second term in (36) (see remark (b) in Section 6.3.3), it follows that the Rao–Shao jackknife and the Shao–Sitter bootstrap variance estimators are asymptotically unbiased for the total variance when the overall sampling fraction $\sum_{h=1}^{L} n_h / \sum_{h=1}^{L} N_h$ is negligible. In other words, provided the overall sampling fraction is negligible, $v_{\mathrm{JRS}}\left(\hat{Y}_{\mathrm{IG}}\right)$ and $v_{\mathrm{BSS}}\left(\hat{Y}_{\mathrm{IG}}\right)$ are valid regardless of the validity of the underlying imputation or nonresponse model. If the overall sampling fraction $\sum_{h=1}^{L} n_h / \sum_{h=1}^{L} N_h$ is not negligible, then both $v_{\mathrm{JRS}}\left(\hat{Y}_{\mathrm{IG}}\right)$ and

$v_{\mathrm{BSS}}\left(\hat{Y}_{\mathrm{IG}}\right)$ will tend to overestimate $E_q V_p\left(\hat{Y}_{\mathrm{IG}}|\mathbf{r}\right)$. Also, in this case, the contribution of the second component $V_q E_p\left(\hat{Y}_{\mathrm{IG}}|\mathbf{r}\right)$, which is not accounted for, may be important, especially if the nonresponse rate is appreciable.

(b) Unlike the jackknife, the bootstrap can be used for nonsmooth parameters.

(c) In the case of NNI, Chen and Shao (2001) showed that the Rao–Shao jackknife overestimates the variance of the imputed estimator $\hat{Y}_{\mathrm{I}\pi}$. Therefore, they proposed to use a partially adjusted jackknife that works in the same way as the Rao–Shao jackknife except that, when an unit is deleted, the imputed values are partially adjusted (see also Shao and Wang, 2008). The application of the bootstrap to NNI has not yet been studied.

(d) Unlike the Taylor linearization procedure, the jackknife is computer intensive, whereas the bootstrap is highly computer intensive.

## 7. Multiple imputation

Multiple imputation was originally proposed by to handle missing data (Rubin, 1987; Little and Rubin, 2002). For each nonrespondent, $M$ imputations are generated, resulting in $M$ completed data sets, which allows the analyst to use standard techniques of analysis designed for complete data. Analyses are performed separately on each of the $M$ completed data sets, and the results are then combined according to Rubin's rule to get point estimates as well as variance estimates. Let $\theta$ be a parameter of interest and $\hat{\theta}$ be an estimator of $\theta$ under complete response. Also, let $\hat{V}$ be an estimator of the variance of $\hat{\theta}$, $V_p\left(\hat{\theta}\right)$. We assume that $\hat{\theta}$ and $\hat{V}$ are (asymptotically) $p$-unbiased for $\theta$ and $V_p\left(\hat{\theta}\right)$, respectively.

Let $\hat{\theta}_{\mathrm{I}}^{(t)}$ and $\hat{V}_{\mathrm{I}}^{(t)}$ be the point estimator and variance estimator using the $t$th imputed data set, treating the imputed values as true values, $t = 1, \ldots, M$. The multiple imputed estimator of $\theta$ is defined as

$$\hat{\theta}_{\mathrm{I},M} = \frac{1}{M}\sum_{t=1}^{M}\hat{\theta}_{\mathrm{I}}^{(t)}. \tag{41}$$

The variance estimator associated with $\hat{\theta}_{\mathrm{I},M}$ is given by

$$T_M = \bar{V}_M + \left(1 + \frac{1}{M}\right)B_M \tag{42}$$

where $\bar{V}_M = \frac{1}{M}\sum_{t=1}^{M}\hat{V}_{\mathrm{I}}^{(t)}$ and $B_M = \frac{1}{M-1}\sum_{t=1}^{M}\left(\hat{\theta}_{\mathrm{I}}^{(t)} - \hat{\theta}_{\mathrm{I},M}\right)^2$. The term $\bar{V}_M$ is an estimator of the sampling variance, whereas the term $\left(1 + \frac{1}{M}\right)B_M$ is an estimator of the nonresponse/imputation variance. Inferences are then based on the approximation

$$T_M^{-1/2}\left(\hat{\theta} - \theta\right) \sim t_\nu,$$

where the degrees of freedom, $\nu$, is given by $\nu = (M-1)\left[1 + \frac{1}{\rho}\right]^2$, with $\rho = \frac{(M-1)B_M}{\bar{V}_M}$. The quantity $\rho$ can be interpreted as the relative increase in variance due to nonresponse.

An improved version for the degrees of freedom from small sample has been proposed by Barnard and Rubin (1999). Clearly, multiple imputation methods are required to be random. As for the case of single imputation, the imputation model should account for sampling design features such as stratification and clustering when appropriate. Reiter et al. (2006) illustrated empirically that failing to do so can lead to severe biases.

In the context of complex designs, there has been some controversy about the validity of multiple imputation in the past two decades. Next, we attempt to clarify and summarize some of the issues raised in the literature. To that end, we distinguish between two approaches for studying the properties of point and variance estimators in the context of multiple imputation: the NM approach described in Section 2.4.1 and a variant of the IM approach called Bayesian imputation model (BIM) approach. In the BIM approach, inferences require specification of a prior distribution $p(\mathbf{y})$ of the vector of the population $y$-values. Inferences are then based on the posterior predictive distribution $p(\mathbf{y}_{\text{mis}}|\mathbf{y}_{\text{obs}})$, where $\mathbf{y}_{\text{obs}}$ and $\mathbf{y}_{\text{mis}}$ denote the observed part and missing part, respectively.

### *7.1. Multiple imputation under the NM approach*

For inferences to be valid under the NM approach, the multiple imputation procedure must be proper. Rubin (1987, pp. 118–119) gives three sets of conditions required for an imputation procedure to be proper. Loosely, speaking, an imputation method is proper if it displays the appropriate amount of variability. All the imputation methods described in Section 2.2.2 are not proper in the sense of Rubin. An example of a proper imputation method closely related to RHDI is the approximate bayesian bootstrap (ABB) described in Rubin (1987, p. 124). In the context of simple random sampling without replacement, it is easily seen that ABB is proper in the sense of Rubin. Proper random REGI, that can be viewed as an extension of ABB, is described in Rubin (1987, p. 166–168). Another imputation procedure consists of generating the imputed values from the conditional distribution $P(\mathbf{y}_{\text{mis}}|\mathbf{y}_{\text{obs}})$, using iterative simulation methods such as Gibbs sampling (e.g., Little and Rubin, 2002; Schafer, 1997). Schafer (1997) call this *Bayesianly proper imputation*. However, Bayesianly proper imputation is not sufficient for proper imputation (Nielsen, 2003). In the context of complex sampling designs, Binder and Sun (1996) argue that Rubin's condition for proper imputation is difficult to verify and may not hold generally. Also, Fay (1992, 1996) showed that an imputation that is proper for a given parameter of interest may not be proper for another.

In practice, most surveys use some form of nonproper random imputation such as those presented as described in Section 2.2.2, in which case multiple imputation is known to lead to invalid inferences. To overcome this problem, Bjørnstad (2007) suggested a simple modification to the variance estimator (42). He proposed to use $T_M^* = \bar{V}_M + \left(k + \frac{1}{M}\right) B_M$ instead of (42), where $k$ is such that $E_{pq\text{I}}\left(T_M^*\right) = V\left(\hat{\theta}_{\text{I},M}\right)$. Note that $T_M^* = T_M$ when $k = 1$. Bjørnstad derived the value $k$ in several special cases. For example, in the case of simple random sampling without replacement and a uniform response mechanism, he showed that $k \approx 1/p_r$ was appropriate for RHDI, where $p_r$ denotes the expected response rate. Although this approach looks promising, extensions to more complex cases are needed.

### 7.2. *Multiple imputation under the BIM approach*

Kim et al. (2006) studied the properties of the variance estimator $T_M$ under the BIM approach, see also Kott (1995). They considered prototype estimators of the form

$$\hat{\theta} = \sum_{i \in s} b_i y_i, \tag{43}$$

where $b_i$ is a coefficient attached to unit $i$ that does not depend on $y_i$. Population totals and domain totals are special cases of (43). As in the case of deterministic imputation (see Section 3), we express the total error of $\hat{\theta}_{\mathrm{I},M}$, $\hat{\theta}_{\mathrm{I},M} - \theta$ as

$$\hat{\theta}_{\mathrm{I},M} - \theta = \left(\hat{\theta} - \theta\right) + \left(\hat{\theta}_{\mathrm{I},\infty} - \hat{\theta}\right) + \left(\hat{\theta}_{\mathrm{I},M} - \hat{\theta}_{\mathrm{I},\infty}\right), \tag{44}$$

where $\hat{\theta}_{\mathrm{I},\infty} = lim_{M \to \infty} \hat{\theta}_{\mathrm{I},M}$. The terms $\hat{\theta} - \theta$, $\hat{\theta}_{\mathrm{I},\infty} - \hat{\theta}$, and $\hat{\theta}_{\mathrm{I},M} - \hat{\theta}_{\mathrm{I},\infty}$ in (44) represent the sampling, nonresponse, and imputation errors, respectively. The total variance of $\hat{\theta}_{\mathrm{I},M}$ is obtained from (28) by replacing $Y$, $\hat{Y}_{\mathrm{G}}$, and $\hat{Y}_{\mathrm{IG}}$ with $\theta$, $\hat{\theta}$, and $\hat{\theta}_{\mathrm{I},M}$, respectively. Under mild regularity conditions, Kim et al. (2006) showed that $\bar{V}_M$ is asymptotically $mpqI$-unbiased for the anticipated sampling variance, $V^m_{\mathrm{SAM}}$, whereas $\left(1 + \frac{1}{M}\right) B_M$ is asymptotically $mpqI$-unbiased for the nonresponse/imputation variance, $V^m_{\mathrm{NR}} + V^m_{\mathrm{I}}$. However, the variance estimator $T_M$ does not track the mixed component, $V^m_{\mathrm{MIX}}$. As a result, the bias of $\hat{T}_{\mathrm{M}}$ is given by $B\left(T_M\right) = -V^m_{\mathrm{MIX}}$. Note that the bias can be positive or negative as $V^m_{\mathrm{MIX}}$ can take positive or negative values. At this point, a question naturally arises: when do we have $V^m_{\mathrm{MIX}} = 0$? A sufficient condition for $V^m_{\mathrm{MIX}}$ to be equal to 0 is that the prototype estimator $\hat{\theta}$ must be self-efficient (e.g., Meng and Romero, 2003). A prototype estimator is self-efficient if and only if

$$V\left(\hat{\theta}_{\mathrm{I},\infty}\right) = V_p\left(\hat{\theta}\right) + V\left(\hat{\theta}_{\mathrm{I},\infty} - \hat{\theta}\right).$$

In other words, a prototype estimator is self efficient if and only if the variance of the multiple imputed estimator $\hat{\theta}_{\mathrm{I},\infty}$ is larger than the variance we would have obtained had complete response been possible. The fact that a prototype estimator is self-efficient or not depends on the sampling design used to select the sample, the parameter we are trying to estimate, and the imputation method used to compensate for nonresponse.

In conclusion, if the prototype variance estimator is not self-efficient, the multiple variance estimator $T_M$ can be considerably biased, in which case the use of a bias-adjusted variance estimator similar to those proposed by Kim et al. (2006) should be considered.

### 7.3. *Fractional imputation*

Fractional imputation (FI) was originally proposed by Kalton and Kish (1984). It was studied by Fay (1996), Kim and Fuller (2004), and Fuller and Kim (2005) for donor random imputation methods such as RHDI. The FI replaces each missing value with $M \geq 2$ imputed values and assign a weight to each imputed value. For example, each imputed value may receive $1/M$ times the original weight. Kim and Fuller (2004) studied the properties of FI under the IM approach, whereas Fuller and Kim (2005) studied its properties under the NM approach. One advantage of FI over single imputation is that the imputation variance can be reduced or eliminated when FI is used. In the latter case, the imputed estimator is said to be fully efficient.

FI is similar to multiple imputation in the sense that each missing value is replaced by $M \geq 2$ imputed values but may be distinguished as follows: (a) under multiple imputation and the NM approach, the imputation method should be proper for the estimators to be valid, whereas the validity of FI does not require the imputation method to be proper. (b) Under multiple imputation, the variance estimator is given by (42), whereas any variance estimation method such as those presented in Section 6 can be used.

## 8. Conclusions

In this chapter, we have not examined how to impute missing data in such a way that all specified edits are satisfied. Some work has been carried out in this area. For imputation of categorical data subject to edits, we refer to Winkler (2003) and for imputation of numerical data subject to edits, refer to Pannekoek et al. (2008) and, in particular, Tempelman (2007).

In practice, parameters measuring relationships such as domain means (domain totals), regression coefficients, and coefficients of correlation are often of interest. The case of domain means is particularly important in practice since estimates for various subpopulations are commonly required in surveys. Ideally, the imputation model (in the context of both single and multiple imputation) should contain the appropriate set of domain indicators. However, domains are generally not specified at the imputation stage and are only known at the analysis stage. As a result, the imputer's model is often different from the analyst's model. For example, the imputation model may not contain a set of domain indicators that are highly related to the variable being imputed but that is of interest to the analyst. In this case, the resulting imputed estimators may be considerably biased. To overcome this difficulty, Haziza and Rao (2005) proposed a bias-adjusted estimator that is asymptotically unbiased under either the NM approach or the IM approach. Skinner and Rao (2002) considered the problem of estimating bivariate parameters such as coefficient of correlations when marginal imputation (i.e., imputing one variable at the time) is used. Because marginal tends to distort the relationships between variables, the resulting imputed estimators are generally biased. Skinner and Rao (2002) proposed a bias-adjusted estimator under simple random sampling without replacement and the NM approach. Finally, Shao and Wang (2002) considered the problem of estimating a coefficient of correlation and proposed a joint imputation procedure that leads to asymptotically unbiased estimators under the IM approach. The problem of relationship is an important one and should receive more attention in the future.

Finally, it is not always possible to impute for the nonrespondents values by using a single imputation method. Such situations often occur in business surveys for which composite imputation within classes is typically used to compensate for nonresponse to a given item $y$, depending on the availability of auxiliary information. For example, if the value of a given business is recorded at a previous occasion, AVI is used. If the historical value is not available but other auxiliary variables are available, then RAI or REGI may be used. If no auxiliary information is available, then MI is used. Inferences in the case of composite imputation has been studied by Shao and Steel (1999). This topic requires further research.

*Sample Surveys: Design, Methods and Applications, Vol. 29A*
ISSN: 0169-7161

DOI: 10.1016/S0169-7161(08)00011-4

11

# Dealing with Outliers in Survey Data

*Jean-François Beaumont and Louis-Paul Rivest*

## 1. Introduction

In survey sampling, the characteristics of the population of interest can often be expressed in terms of means and totals. These statistics are sensitive to the presence of outlying units which have unusual values for some of the corresponding variables. It may occur that two or three units account for an important percentage, say 5−10%, of the population total of a survey variable. The estimation of the population mean of such a variable raises challenging statistical issues. The problem is amplified when some large units are associated with large survey weights. Such units may have a huge influence on the estimates of the population characteristics.

The occurrence of large units is common in population of businesses. If auxiliary information is available to identify these units, their impact on the survey estimates can be minimized when constructing the survey design. This is typically accomplished by stratifying according to a size measure and by putting all the large units in a take-all stratum in which selection is done with certainty; see Chapter 17 of this volume by Hidiroglou and Lavallée. However, some large units may still be unexpectedly selected in the sample due to imperfect auxiliary information at the time of stratification. This chapter is concerned with the occurrence of these large units in a survey sample. In many instances, their effect cannot be accounted for by a simple poststratification. They occur in a haphazard way.

Using the terminology of Chambers (1986), outliers can be classified as being either representative or nonrepresentative. Nonrepresentative outliers are most likely caused by reporting errors, but they may also be units that are deemed unique in the population although this may not be easy to determine. Representative outliers are other large observations that are representative of the nonsampled part of the population and that are dealt with at the estimation stage of a survey. Nonrepresentative outliers are managed at the data collection and editing stages of a survey using outlier detection techniques.

There are numerous outlier detection techniques in the sample survey literature. Lee (1995) provided a review of some of these methods. Hidiroglou and Lavallée (Chapter 17 of this volume) described the most common methods used in business surveys, including the well-known technique developed by Hidiroglou and Berthelot (1986). These methods determine a threshold above which units are flagged as being potential outliers. Although Winsorization or M-estimation, discussed in Sections 2, 3, and 4 of this chapter, have not been specifically developed for the identification of outliers, they could also be used as alternative methods of outlier detection since these methods also involve a threshold that separates outliers (or influential units) from the other units. All the above methods deal with outliers with respect to a single variable. More recently, techniques have been developed to identify multivariate outliers (Béguin and Hulliger, 2004, 2008; Chambers et al., 2004; Franklin et al., 2000).

The goal of outlier detection techniques is to identify suspicious values and to either confirm the reported values with the respondent or to correct them. If a reporting error is found later, at the editing and processing stages of the survey (see Chapter 9 of this volume by De Waal), it may not be possible to fix the error since the respondent may not be available anymore. In this case, the erroneous value is often set to missing and imputed (see Chapter 10 of this volume by Haziza). In sample surveys conducted by national statistical agencies, it is thus less likely for the remaining outlying values to be created by reporting errors than in other areas of Statistics due to this careful data validation and editing.

Although it may usually not be possible to identify all the reporting errors in practice, we make the assumption that the impact of the remaining errors is negligible and that the outstanding large observed values are representative outliers; that is, they are not caused by reporting errors and are not deemed unique in the population. Note that in the rare event where a unit is deemed unique in the population and not selected with certainty, its survey weight is usually simply set equal to 1. This situation is rare as these very large units can often be detected before sampling so that they can be selected with certainty yielding a sampling weight of 1.

Representative outliers may greatly influence survey estimates. Including or excluding an outlier in the calculation of the sample mean can have a dramatic impact on its magnitude. While standard design-based estimators of totals are approximately unbiased, representative outliers can dramatically increase their variances. Estimation methods that curb the influence of large values produce more stable estimates, but are biased. The art of outlier treatment in survey sampling lies in the management of this bias-variance trade-off.

There is a large body of literature on the detection and the treatment of outliers in classical statistics; see for instance, Barnett and Lewis (1994). This includes formal methods to ascertain whether a sample value is an outlier given that the data come from a known distribution and techniques to detect the presence of outliers when fitting a complex statistical model. The field of Robust Statistics (see Hampel et al., 1986) develops estimation methods that are insensitive to the presence of outliers. These methods may lead to highly biased estimators of population totals and population means in survey sampling because these population parameters are themselves quite sensitive to population outliers. Thus, the Robust Statistics literature offers little assistance for the understanding of the bias-variance relationship that is central to the treatment of outliers in survey data.

At the estimation stage, one can adopt a two-step strategy for dealing with representative outliers in a sample. First outliers are identified using a detection rule such as that of Hidiroglou and Berthelot (1986) or those mentioned earlier. Then, the sampling weights of the identified outliers are reduced and the population characteristics are estimated using the reduced weight. Hidiroglou and Srinath (1981) suggested three methods of weight reduction. They all give an important reduction in the conditional mean squared error (MSE) of the estimators, given a fixed number of outliers in the sample. Several extensions of this approach are discussed by Lee (1995). In many instances, the two-step approach of Hidiroglou and Srinath is useful. However, in samples drawn from skewed distributions, such as those presented in Table 1, establishing a threshold for outliers is somewhat arbitrary. An alternative approach is to perform simultaneously outlier detection and treatment so that the specification of the threshold becomes part of the estimation procedure. This is the approach on which we focus in this chapter.

The next section deals with the estimation of the mean of a skewed distribution in an infinite population. Section 3 discusses the estimation of the total of a finite population when the largest data values are Winsorized. Coping with outliers in a calibration estimator, such as the generalized regression estimator, is studied in Section 4

Table 1
Data for 10 rainbow smelt inventories in Lake St-Jean, Quebec

| Year | 1996 | 1997 | 1998 | 1999 | 2000 | 2001 | 2002 | 2003 | 2004 | 2005 |
|---|---|---|---|---|---|---|---|---|---|---|
| | 167.82 | 76.77 | 61.79 | 19.03 | 92.61 | 7.04 | 94.91 | 17.05 | 77.80 | 1506.24 |
| | 67.73 | 6.69 | 44.75 | 10.77 | 83.83 | 5.77 | 68.77 | 4.95 | 45.77 | 686.71 |
| | 63.39 | 5.90 | 42.32 | 4.23 | 81.86 | 3.92 | 37.07 | 3.85 | 36.25 | 518.29 |
| | 60.89 | 5.51 | 36.39 | 0.80 | 30.03 | 1.49 | 8.90 | 0.70 | 26.60 | 330.73 |
| | 40.54 | 2.03 | 28.22 | 0.47 | 28.05 | 0.85 | 5.23 | 0.65 | 23.05 | 208.82 |
| | 26.09 | 1.58 | 14.91 | 0.13 | 14.57 | 0.73 | 4.37 | 0.50 | 12.75 | 153.42 |
| | 22.60 | 0.87 | 11.10 | 0.10 | 10.50 | 0.67 | 2.86 | 0.40 | 6.60 | 18.14 |
| | 15.59 | 0.82 | 10.88 | 0.07 | 9.33 | 0.67 | 1.77 | 0.40 | 2.37 | 6.52 |
| | 12.35 | 0.57 | 9.00 | 0.07 | 7.68 | 0.62 | 1.59 | 0.40 | 1.67 | 3.92 |
| | 9.21 | 0.52 | 7.96 | 0.07 | 7.16 | 0.61 | 0.92 | 0.30 | 0.97 | 3.19 |
| | 5.86 | 0.29 | 6.31 | 0.07 | 5.07 | 0.55 | 0.85 | 0.30 | 0.45 | 2.88 |
| | 5.02 | 0.15 | 5.46 | 0.03 | 1.58 | 0.22 | 0.45 | 0.27 | 0.43 | 2.34 |
| | 4.63 | 0.11 | 5.01 | 0.03 | 0.91 | 0.15 | 0.39 | 0.25 | 0.35 | 1.76 |
| | 3.90 | 0.10 | 4.68 | 0.03 | 0.68 | 0.14 | 0.35 | 0.23 | 0.23 | 1.33 |
| | | | 4.56 | 0.03 | 0.58 | | 0.34 | 0.20 | 0.20 | 0.94 |
| | | | 3.87 | 0.02 | 0.56 | | 0.27 | 0.20 | 0.17 | 0.81 |
| | | | 3.46 | | 0.55 | | 0.19 | 0.20 | 0.10 | 0.66 |
| | | | 2.52 | | 0.53 | | 0.19 | 0.13 | 0.10 | 0.57 |
| | | | 2.13 | | 0.50 | | 0.17 | 0.13 | 0.07 | 0.46 |
| | | | 2.12 | | 0.49 | | | 0.08 | 0.07 | 0.27 |
| | | | 1.11 | | 0.44 | | | 0.07 | 0.07 | 0.24 |
| | | | 0.87 | | 0.32 | | | 0.07 | 0.04 | 0.15 |
| | | | 0.71 | | 0.31 | | | 0.07 | 0.03 | 0.15 |
| | | | 0.44 | | 0.26 | | | 0.03 | 0.03 | 0.15 |
| | | | 0.43 | | 0.25 | | | 0.03 | 0.03 | 0.14 |
| | | | 0.42 | | 0.23 | | | 0.03 | | |
| | | | 0.15 | | 0.19 | | | 0.03 | | |
| | | | 0.14 | | 0.18 | | | | | |
| | | | 0.08 | | 0.14 | | | | | |
| # 0 values | 0 | 2 | 3 | 12 | 9 | 30 | 14 | 7 | 9 | 9 |
| Sample size, $n$ | 14 | 16 | 32 | 28 | 38 | 44 | 33 | 34 | 34 | 34 |

where M-estimation is discussed. Section 5 investigates the treatment of stratum jumpers and extreme design weights via Winsorization of survey weights and weight smoothing. Finally, some practical issues are discussed in the last section as well as potential future work.

## 2. Estimation of the mean of an asymmetric distribution in an infinite population

### 2.1. Some asymmetric distributions

Several distributions have been considered in the literature to model asymmetric positive variables; a selection of models is presented in Table 2. It gives the density, the expectation $\mu$, the variance $\sigma^2$, the coefficient of variation ($\text{CV} = \sigma/\mu$), and the skewness coefficient ($\gamma_1 = E\{(Y-\mu)^3\}/\sigma^3$) for three asymmetric densities.

In Table 2, $f(\alpha) = \frac{\Gamma(1+3/\alpha)-3\Gamma(1+1/\alpha)\Gamma(1+2/\alpha)+2\{\Gamma(1+1/\alpha)\}^3}{[\Gamma(1+2/\alpha)-\{\Gamma(1+1/\alpha)\}^2]^{3/2}}$. The densities of Table 2 are all written in terms of a positive shape parameter $\alpha$. The Weibull distribution with shape parameter $\alpha = 1$ is the exponential distribution. Fig. 1 shows that both the CV and the skewness of the density increase with the shape parameter. The Weibull family is considered by Fuller (1991), and the lognormal distribution is investigated by Myers and Pepin (1990). The Pareto distribution gives an instance of an extreme skewness since some of its moments are not finite. Its variance is defined only if $\alpha > 2$.

The underlying distribution for an asymmetric sample with large positive values is typically unknown. One can attempt to select a parametric model for such a data set by using some goodness of fit tests; an estimator of $\mu$, optimal for the model selected, can be calculated. Myers and Pepin (1990) investigated this technique for distributions that are close to the lognormal distribution. They showed that the parametric estimator for $\mu$ is very sensitive to a misspecification of the underlying distribution, undetectable by goodness of fit tests. They concluded that the sample mean is a better estimator for $\mu$ than an optimal lognormal estimator when the underlying distribution is close to a lognormal model. Thus, the estimation of $\mu$ via a parametric model is not pursued

Table 2
Three asymmetric densities defined on $(0, \infty)$ for modeling skewed variables

| Model | Weibull | Lognormal | Pareto |
|---|---|---|---|
| Density | $\alpha y^{\alpha-1}\exp(-y^{\alpha})$ | $\frac{1}{y\sqrt{2\pi}\alpha}\exp\left(\frac{-\{\log(y)\}^2}{2\alpha^2}\right)$ | $\frac{\alpha}{(1+y)^{\alpha+1}}$ |
| $\text{E(Y)} = \mu$ | $\Gamma(1+1/\alpha)$ | $\exp(\alpha^2/2)$ | $1/(\alpha-1)$ |
| $\text{Var(Y)} = \sigma^2$ | $\Gamma(1+2/\alpha)-\{\Gamma(1+1/\alpha)\}^2$ | $\exp(\alpha^2)\left\{\exp(\alpha^2)-1\right\}$ | $\frac{\alpha}{(\alpha-1)^2(\alpha-2)}$ |
| CV | $\sqrt{\frac{\Gamma(1+2/\alpha)}{\{\Gamma(1+1/\alpha)\}^2}-1}$ | $\sqrt{\exp(\alpha^2)-1}$ | $\sqrt{\frac{\alpha}{(\alpha-2)}}$ |
| $\gamma_1$ | $f(\alpha)$ | $\left\{2+\exp(\alpha^2)\right\}\sqrt{\exp(\alpha^2)-1}$ | $\sqrt{\frac{\alpha-2}{\alpha}}\frac{2(\alpha+1)}{\alpha-3}$ |

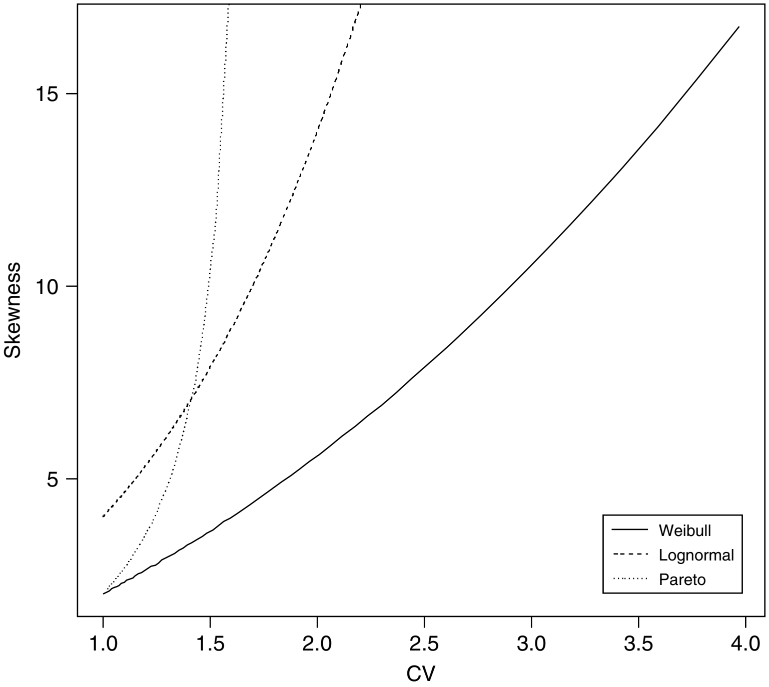


Fig. 1. The CV-skewness relationship for the three families of Table 2.

in the remaining of this chapter. Nonparametric alternatives to the sample mean are presented in the next sections. The distributions of Table 2 are regarded as test cases for the estimation procedures discussed.

When a simple random sample of size $n$ is obtained from a skewed distribution $F(y)$ with skewness coefficient $\gamma_1$, the skewness of the sample mean is $\gamma_1/\sqrt{n}$. Thus, the distribution of the sample mean has a heavy right tail when $\gamma_1$ is large and $n$ is moderate. This implies that the sample mean can, at some occasions, take relatively large values. The methods presented next limit the occurrence of these large sample means by curtailing the effect of the largest values in the sample.

Sections 2.2, 2.3, and 2.4 consider a sample of size $n$, $y_1, y_2, \ldots, y_n$, which is drawn from an infinite population with a skewed distribution $F(y)$. The order statistics for this sample are denoted $y_{(1)} < y_{(2)} < \ldots < y_{(n)}$.

### 2.2. *Searls' Winsorized mean*

This section presents a nonparametric alternative to the sample mean as an estimator of the population mean $\mu$. Searls (1966) suggested to Winsorize the largest values to estimate the population mean. If $R$ stands for the Winsorization cutoff, the Winsorized mean is given by

$$\bar{y}_R = \sum_{i=1}^{n} \min(y_i, R)/n. \tag{1}$$

When the distribution $F$ and the sample size $n$ are fixed, the cutoff $R_n$ that minimizes the MSE of $\overline{y}_R$ is the solution of

$$\frac{R-\mu}{n-1} = E_F\{\max(Y-R,0)\}, \tag{2}$$

where $E_F(\cdot)$ denotes an expectation taken with respect to the distribution $F$. When historical data are available to estimate the underlying distribution $F$, solving (2) can yield a good Winsorization cutoff if the training sample size is much larger than $n$. However, estimating $F$ with $F_n$, the sample empirical distribution function, and solving (1) with $F = F_n$ may result in an unstable estimator as the solution $\hat{R}_n$ depends heavily on the largest data values; see the simulations in Rivest and Hurtubise (1995) and the example of Section 2.5. Pooling several samples to estimate $F$ can sometimes be envisioned; this is illustrated in Section 2.5. If there is no additional information to estimate the cutoff, alternative estimators are presented in Sections 2.3 and 2.4.

When $R$ is fixed, MSE$(\overline{y}_R)$ is easily estimated by

$$\text{mse}(\overline{y}_R) = \frac{\sum_{i=1}^{n}\left[\min(y_i,R)-\overline{y}_R\right]^2}{n(n-1)} + \left[\frac{\sum_{i=1}^{n}\max(y_i-R,0)}{n}\right]^2.$$

When $R_n$ is obtained from (2), the expected number of Winsorized data points, $n\{1-F(R_n)\}$, decreases with an increase in skewness. This number is larger for a Weibull distribution than for a Pareto distribution. For the distributions of Table 2, it is less than 2 for samples as large as $n = 300$. Thus, estimating the cutoff $R_n$ with a large order statistic looks promising. This is considered in the next section.

### 2.3. *The once-Winsorized mean*

The once-Winsorized mean is obtained by taking $y_{(n-1)}$, the second largest order statistic, as a Winsorization cutoff $R$ in (1). It is given by $\overline{y}_1 = \overline{y} - (y_{(n)} - y_{(n-1)})/n$. Rivest (1994) shows that taking $R = y_{(n-1)}$ gives a smaller MSE than $R = y_{(n-2)}$, provided that the underlying distribution $F$ has a finite variance. Thus, $\overline{y}_1$ is the best possible Winsorized mean when the cutoff is selected among extreme order statistics. A MSE estimator proposed in Rivest (1994) is given by

$$\text{mse}(\overline{y}_1) = \frac{s^2}{n} - \frac{(y_{(n)}+y_{(n-1)}-2\overline{y}_1)(y_{(n)}-3y_{(n-1)}+2y_{(n-2)})}{n^2},$$

where $s^2$ is the sample variance. This estimator is consistent even in instances where $F$ has an infinite variance, such as the Pareto with parameter $\alpha = 2$. For the exponential distribution, $\overline{y}$ and $\overline{y}_1$ have the same MSE. Winsorizing improves the precision of the sample mean when the underlying distribution is more skewed than the exponential distribution. This holds true if the exponential Q-Q plot has a convex shape. Since, for an exponential sample, $\mathrm{E}(y_{(i)})$ is proportional to $\sum_{j=1}^{i} 1/(n-j+1)$ this Q-Q plot consists of the points

$$\left\{\left(\sum_{j=1}^{i} 1/(n-j+1),\, y_{(i)}\right), i=1,\ldots,n\right\}.$$

The MSEs of the optimal Searls' Winsorized mean and of the once-Winsorized mean can be expressed as $\sigma^2/n - \psi(n)$ with $n\psi(n)$ goes to 0 as $n$ goes to infinity. The form of $\psi(n)$ and the rate of convergence of $n\psi(n)$ to 0 depend of the max domain of attraction of the underlying distribution $F(y)$. This highlights that the optimal Winsorized mean and the once-Winsorized mean have the same asymptotic distribution as the sample mean; thus standard asymptotic calculations cannot bring out the gains in precision associated with these estimators.

### 2.4. *Fuller's preliminary test estimator*

To determine whether the tails of the distribution $F$ are heavier than exponential tails, a simple statistic proposed in Fuller (1991) is

$$F_{Tj} = \frac{\sum\limits_{i=n-j+1}^{n} Z_i/j}{\sum\limits_{i=n-T_j}^{n-j} Z_i/\left(T_j - j + 1\right)} = \frac{N_j}{D_j},$$

where $Z_i = (n-i+1)(y_{(i)} - y_{(i-1)})$ is a normalized spacing, and $j$ and $T_j$ are integers to be determined. When the underlying distribution $F$ is exponential, $F_{Tj}$ has an $F$ distribution with $2j$ and $2(T_j - j)$ degrees of freedom. One can reject the hypothesis that the upper tail of the distribution is exponential if $F_{Tj}$ is large. The preliminary test estimator is

$$\overline{y}_P = \begin{cases} \overline{y} & \text{if} \quad F_{Tj} < K_j \\ \dfrac{1}{n}\left\{\sum\limits_{i=1}^{n-j} y_{(i)} + j\left[y_{(n-j)} + K_j D_j\right]\right\} & \text{if} \quad F_{Tj} \geq K_j, \end{cases}$$

where $D_j$ is the denominator for $F_{Tj}$ defined above and $K_j$ is a predetermined cutoff. The largest observations are Winsorized only if the null hypothesis of an exponential tail is rejected. This estimator depends on three tuning parameters, $j$, $T_j$, and $K_j$ that are defined by the user. Fuller (1991) showed, through simulations, that $j = 3$, $T_j = 4n^{1/2} - 10$, and $K_j = 3.5$ yield an estimator with good sampling properties. These values are used for the numerical illustrations presented in the next section. Unfortunately, no estimator for the MSE of $\overline{y}_p$ is available.

### 2.5. *An example: Analysis of 10 years of inventories for the rainbow smelt in Lake St-Jean, Quebec*

The data set in Table 1 gives the outcomes of 10 years of trawl survey, from 1996 to 2005, for the rainbow smelt in Lake St-Jean, Quebec. We are grateful to Michel Legault

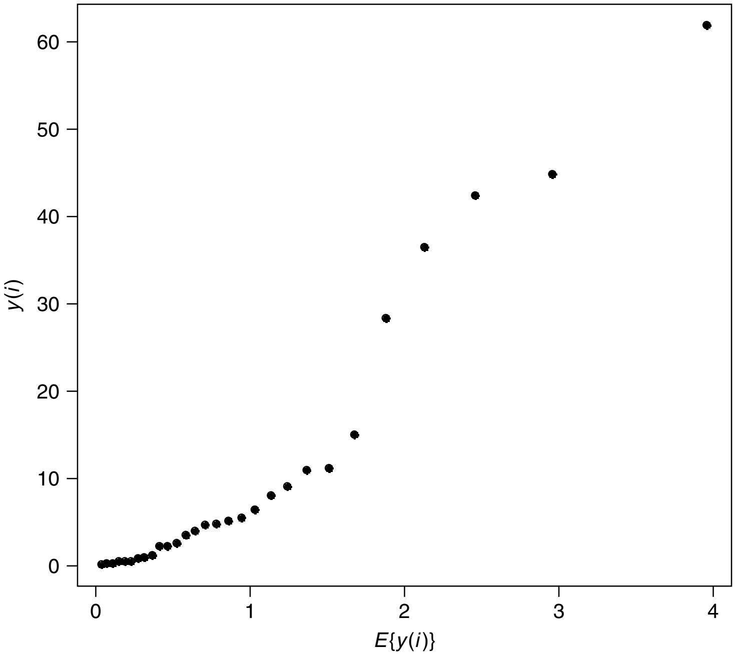


Fig. 2. Exponential Q-Q plot for the non-null densities of 1998.

from the Ministry of Natural Resources and Fauna of Quebec for his permission to use this unpublished data set. Each data value gives the density in number of fish per $1000^3$ meters of water at a random sampling point on the lake. The original data were stratified by depth and each data point had an associated area measure. These two variables are omitted since they did not account for much of the variability. The area of lake St-Jean is more than $10^4$ km$^2$; this makes the sampling fractions negligible. We consider these samples as coming from infinite populations. Table 1 gives, for each year, the measured non-null densities in decreasing order of magnitude and the number of sample points with a density of 0.

Several factors, such as the abundance of predators, influence the density of the rainbow smelt. Providing a yearly density estimate is a problem since the largest observation can account for up to 70% of the sample mean in a given year. The CV and the skewness of the 10 samples of non-null values are in the range (1.2, 2.9) and (1.4, 3.9), respectively. Considering Fig. 1, the skewness in these data sets is relatively mild. The exponential Q-Q plots have a convex shape as illustrated in Fig. 2. The sample means are highly unstable, and the nonparametric alternatives suggested in this section are presented in Table 3.

To estimate the optimal cutoffs for Searls' Winsorized mean of Section 2.2, we assumed that the distributions of the non-null smelt densities for each year were the same, up to a scale change. The following algorithm was used:

(a) Discard all the null values;

(b) Normalize the data by dividing each observed density by the first quartile of its yearly sample;

Table 3
The sample mean $\overline{y}$, the optimal Winsorized mean $\overline{y}_R$, the Winsorized mean $\overline{y}_{\hat{R}}$ with R estimated from the current sample, the once-Winsorized mean $\overline{y}_1$ (and their estimated root mean squared errors in adjacent columns), and the preliminary test estimator $\overline{y}_P$ for 10 rainbow smelts inventories. For the two Winsorized estimators, $n_R$ and $n_{\hat{R}}$ give the number of data points above the Winsorization cutoffs

| Year | $\overline{y}$ | $s/\sqrt{n}$ | $n_R$ | $\overline{y}_R$ | *rmse* | $n_{\hat{R}}$ | $\overline{y}_{\hat{R}}$ | *rmse* | $\overline{y}_1$ | *rmse* | $\overline{y}_p$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 96 | 36.12 | 11.87 | 0 | 36.12 | 11.87 | 1 | 31.24 | 9.46 | 28.97 | 7.61 | 36.12 |
| 97 | 6.37 | 4.73 | 1 | 4.16 | 3.37 | 1 | 4.1 | 3.37 | 1.99 | 1.04 | 6.29 |
| 98 | 9.74 | 2.75 | 0 | 9.74 | 2.75 | 3 | 8.79 | 2.46 | 9.21 | 2.55 | 9.74 |
| 99 | 1.28 | 0.77 | 2 | 0.6 | 0.75 | 2 | 0.95 | 0.62 | 0.99 | 0.88 | 0.33 |
| 0 | 9.98 | 3.84 | 0 | 9.98 | 3.84 | 3 | 8.45 | 3.39 | 9.75 | 3.77 | 9.98 |
| 1 | 0.53 | 0.22 | 0 | 0.53 | 0.22 | 2 | 0.44 | 0.19 | 0.5 | 0.25 | 0.44 |
| 2 | 6.96 | 3.6 | 2 | 5.01 | 3.02 | 2 | 5.41 | 2.99 | 6.17 | 4.26 | 3.26 |
| 3 | 0.93 | 0.52 | 2 | 0.57 | 0.42 | 1 | 0.69 | 0.39 | 0.57 | 0.3 | 0.33 |
| 4 | 6.95 | 2.88 | 1 | 6.94 | 2.87 | 2 | 5.85 | 2.44 | 6 | 2.65 | 6.95 |
| 5 | 101.4 | 50.28 | NA | NA | NA | 1 | 80.47 | 40.27 | 77.33 | 40.95 | 101.4 |

(c) Use formula (2), with $F$ equal to the empirical distribution function of the pooled non-null standardized values and $n$ equal to the number of non-null data values in a given year, to calculate the standardized optimal cutoff for that year;

(d) Multiply the standardized optimal cutoff of (c) by the first quartile of (b) to get the final cutoff.

With this algorithm, the value of the optimal cutoff for each year does not depend on the number of null observations for that year. When pooling all the data, this algorithm yielded a severe Winsorization for 2005, with $n_R = n\{1-F_n(R_n)\} = 4$ data values larger than the cut-off. This suggests that 2005 violates the homogeneity assumption underlying the calculation of the optimal cutoffs. This year was discarded and the Winsorization cutoffs for 1995–2004 were calculated with the 187 non-null values for 1995–2004.

In Table 3, the root mses for the optimal Winsorized mean are evaluated as if the optimal cutoffs calculated above were known. A total of eight data points are Winsorized for the nine samples. This is in agreement with the discussion in Section 2.2 that showed that few data points are Winsorized per sample under the optimal scheme.

Table 3 also features the estimator $\overline{y}_{\hat{R}}$ obtained with a Winsorization cutoff that minimizes the estimated MSE 1 year at a time. This estimator does not use the data from the other years to determine$\hat{R}$. It Winsorizes more points in years where no outliers had been detected with the previous scheme. It has an estimated median bias of $-19\%$. The *rmse* column gives the minimal value of $\sqrt{\text{mse}(\overline{y}_R)}$ as a function of $R$; this underestimates the true MSE of $\overline{y}_{\hat{R}}$ as it does not account for the variability in the estimation of $R$. It gives an upper bound on the gains of efficiency achievable with a simple Winsorization; the median of the maximal efficiency, $s/\sqrt{n \times \min_R\{\text{mse}(\overline{y}_R)\}}$, is 123%.

In Table 3, the estimated median bias of the once-Winsorized mean is $-17\%$, whereas the median of $s/\sqrt{n \times \text{mse}(\overline{y}_1)}$ is 108%. Thus, Winsorizing the largest observation improves the estimation of the population mean. The preliminary test estimator $\overline{y}_p$ is equal to the sample mean for 5 analyses out of 10; it is less than 50% of the sample mean for 3 years. The values $j = 3$, $T_j = 4n^{1/2} - 10$, and $K_j = 3.5$ used in the calculations might not be well suited for the data of Table 1. Constructing reliable confidence intervals for the unknown population mean using the Winsorized means presented in Table 3 is still an open research area. Application of empirical likelihood techniques presented in Chen et al. (2003) to Winsorized means could be envisioned for that purpose.

The mean density for the first 9 years of the study is $\overline{\overline{y}} = 8.76$, the average of the nine values of $\overline{y}$. The standard error is estimated by $\sqrt{\sum s_i^2/n_i/9} = 1.60$, where the subscript $i$ refers to the year. The average optimal Winsorized mean is $\overline{\overline{y}}_R = 8.18$, with a root mse $\sqrt{\text{mse}(\overline{\overline{y}}_R)}$ of 1.61, where

$$\text{mse}\left(\overline{\overline{y}}_R\right) = \sum_{i=1}^{9} \frac{1}{81} \frac{\sum\limits_{j=1}^{n_i}\left[\min(y_{ij}, R_i) - \overline{y}_{iR}\right]^2}{n_i(n_i-1)} + \left[\frac{\sum\limits_{i=1}^{9}\sum\limits_{j=1}^{n_i}\max(y_{ij} - R_i, 0)}{9n_i}\right]^2,$$

and subscript $j$ refers to a sample point within a year. The two estimators have approximately the same root mse; however $\overline{\overline{y}}_R$ is biased. In Table 3, the Winsorization cutoffs were selected to optimize the bias-variance tradeoff for yearly estimates. These cutoffs are no longer optimal for estimating the average density over 9 years. Thus, $\overline{\overline{y}}$ has better sampling properties than $\overline{\overline{y}}_R$ for estimating the 9-year average. For the estimation of a sum using a Winsorized mean for each element, Eq. (2) no longer gives the optimal Winsorization cutoff for the elements of the sum. This problem is addressed in the next section in the context of a stratified sampling design.

## 3. The estimation of totals in finite populations containing outliers

When an outlier appears in a sample collected from a finite population, one can consider that this outlier is unique and unrepresentative of the nonsample part of the population (Rao, 1971). One can thus reduce its sampling weight to 1 and redistribute the outstanding sampling weight among the non-outlier sample units. If an outlier is deemed to be representative, one would like the sampling weight to be larger than or equal to 1. The estimators discussed in Section 2 do not satisfy this property. For example, the once-Winsorized mean of Section 2.3 sometimes gives a sampling weight smaller than 1 to the largest observation.

This section focuses on the estimation of the total $T_y = \sum_{i \in U} y_i$ of variable $y$ for the population $U$ of size $N$. Suppose that a simple random sample of size $n$ is drawn from the population, which is itself generated from an asymmetric distribution $F(y)$ with mean $\mu$. Lemma 5 of Fuller (1991) shows that for any estimator $\overline{y}_*$ that has a smaller MSE than the sample mean $\overline{y}_s$ when estimating $\mu$,

$$\hat{T}_y^* = N\left[f\overline{y}_s + (1-f)\,\overline{y}_*\right]$$

is a better estimator of the population total $T_y$ than the expansion estimator $\hat{T}_y = N\overline{y}_s$, where $f = n/N$ is the sampling fraction. In addition, the design MSE of $\hat{T}_y^*$ can be estimated by

$$\text{mse}(\hat{T}_y^*) = N^2\,(1-f)^2\,\text{mse}\left(\overline{y}_*\right) + N(1-f)s^2,$$

where $\text{mse}(\overline{y}_*)$ is an infinite population MSE estimator and $s^2$ is the sample variance.

The Winsorized mean of Section 2.2 can be adapted to sampling from a finite population using $\overline{y}_* = \sum_{i \in s} \min(y_i, R)/n$ in the above equation for $\hat{T}_y^*$. If $R$ stands for the Winsorization cutoff, the resulting Winsorized estimator of $T_y$ is given by

$$\hat{T}_{yR} = N \left[ f\,\overline{y}_s + (1-f)\,\frac{\sum\limits_{i \in s} \min(y_i, R)}{n} \right] = \hat{T}_y - \frac{(1-f)}{f} \sum_{i \in s} \max(y_i - R, 0),$$

where $s$ is the set of the units in the sample. The estimator $\hat{T}_{yR}$ can also be written as $\hat{T}_{yR} = \sum_{i \in s} d_i^* y_i$, where

$$d_i^* = 1 + \left( \frac{N}{n} - 1 \right) \frac{\min(y_i, R)}{y_i}$$

is a reduced sampling weight that is never smaller than 1. Gross et al. (1986) called $\hat{T}_{yR}$ a Winsorized type II estimator. They called estimator (1), where a unit can receive a weight smaller than 1, a Winsorized type I estimator.

The value $R_n$ that minimizes the MSE of $\hat{T}_{yR}$ with respect to the sampling design can be approximated by the solution of

$$\frac{R_n - \overline{y}_U}{n-1} = \frac{(1-f)}{N} \sum_{i=1}^{N} \max(y_i - R_n, 0). \qquad (3)$$

The solution $R_n$ to Eq. (3) is slightly larger than that of Eq. (2) for the infinite population Winsorization, with $F(y)$ equal to the empirical distribution function of $y$ in the population. The next section discusses the extension of (3) to a stratified sampling design.

### 3.1. Winsorization in a stratified sampling design

Suppose that a random stratified design is used to sample a population with an asymmetric distribution. The following notation is used:

- $N_h$ is the size of stratum, $h = 1, \ldots, H$ and $N = \Sigma N_h$
- $s_h$ is the sample, of size $n_h$, collected in stratum $h$, $f_h = n_h/N_h$ is the sampling fraction in stratum $h$ and $n = \Sigma n_h$;
- $\overline{y}_{sh}$ is the sample mean in stratum $h$ and $\hat{T}_y = \sum N_h \overline{y}_{sh}$ is the simple expansion estimator of the total;
- $\overline{y}_{hU}$ is the population mean in stratum $h$;
- $R_h$ is the Winsorization cutoff in stratum $h$.

A Winsorized type II estimator of the total is

$$\begin{aligned}\hat{T}_{yR} &= \sum_{h=1}^{H} N_h \left\{ f_h \overline{y}_{sh} + (1-f_h)\,\frac{\sum\limits_{i \in s_h} \min(y_{hi}, R_h)}{n_h} \right\} \\ &= \sum_{h=1}^{H} N_h \left\{ \overline{y}_{sh} - \frac{1-f_h}{n_h} \sum_{i \in s_h} \max(y_{hi} - R_h, 0) \right\}. \qquad (4)\end{aligned}$$

Alternative expressions for this estimator are $\hat{T}_{yR} = \sum_{h=1}^{H} \sum_{i \in s_h} d^*_{hi} y_{hi}$ and $\hat{T}_{yR} = \sum_{h=1}^{H} \sum_{i \in s_h} (N_h/n_h) y^*_{hi}$, where

$$d^*_{hi} = 1 + \left(\frac{N_h}{n_h} - 1\right) \frac{\min(y_{hi}, R_h)}{y_{hi}} \text{ and } y^*_{hi} = f_h y_{hi} + (1 - f_h) \min(y_{hi}, R_h).$$

They highlight that the Winsorized estimator can be written in the same form as the simple expansion estimator $\hat{T}_y = \sum_{h=1}^{H} \sum_{i \in s_h} (N_h/n_h) y_{hi}$ with either a reduced weight, $d^*_{hi}$, or a reduced $y$-value, $y^*_{hi}$, for the outliers; that is, the units with a $y$-value above the threshold $R_h$. The above expressions also show that the reduced weight $d^*_{hi}$ given to outliers cannot be smaller than 1. The MSE of (4) is given by

$$\text{MSE}(\hat{T}_{yR}) = \text{Var}(\hat{T}_{yR}) + \left[\sum_{h=1}^{H} \sum_{i=1}^{N_h} (1 - f_h)\,(y_{hi} - \min(y_{hi}, R_h))\right]^2,$$

where $\text{Var}(\hat{T}_{yR})$ is the standard design-based variance for the estimator of the total of $y^*_{hi}$.

The minimization of $\text{MSE}(\hat{T}_{yR})$ with respect to $\{R_h\}$ is discussed in Kokic and Bell (1994) and Rivest and Hurtubise (1995). An approximate solution, slightly larger than the true minimizers of $\text{MSE}(\hat{T}_{yR})$, is given by $R_h = \overline{y}_{hU} + Rf_h/(1 - f_h)$, where $R$ is obtained by solving

$$R = \sum_{h=1}^{H} \sum_{i=1}^{N_h} (1 - f_h) \max\left(y_{hi} - \overline{y}_{hU} - R\frac{n_h}{(N_h - n_h)}, 0\right).$$

This is Eq. (3.4) in Rivest and Hurtubise (1995) when $f_h \approx 0$ and is similar to Eq. (7) in Kokic and Bell (1994). In stratum $h$, a data point is Winsorized if $y_{hi} > \overline{y}_{hU} + Rf_h/(1 - f_h)$. This condition can be interpreted in terms of a model $m$ where $E_m(y_{hi}) = \mu_h$. The population residual with respect to this model needs to be larger than a cutoff that increases with the sampling fraction for a $y$-value to be downweighted. Section 4 presents similar results for design-based inference where auxiliary variables are used to improve the estimation of $T_y$.

This Winsorization scheme has properties similar to those for the infinite population model. The number of Winsorized data points decreases with the skewness of the data and the total number of Winsorized data points in all the strata combined should be relatively small.

### 3.2. *Winsorization in a general sampling design*

This section and Sections 4 and 5 consider a population $U$ that is sampled with an arbitrary sampling design with single and joint selection probabilities given by $\pi_i$ and $\pi_{ij}$, respectively. The Horvitz–Thompson estimator of the population total is $\hat{T}_y^{\text{HT}} = \sum_{i \in s} y_i/\pi_i$. A basic, Winsorized type I estimator is $\hat{T}_{yR}^{\text{HT}} = \sum_{i \in s} \min(y_i/\pi_i, R)$. Alternative Winsorization schemes are proposed in Section 4.2. They are developed in the next section where auxiliary information is used in the construction of an estimator for $T_y$.

## 4. The estimation of totals using auxiliary information in finite populations containing outliers

Suppose that a vector $\mathbf{x}_i$ of auxiliary variables is available for all the units in the sample and that the totals $\mathbf{T}_\mathbf{x} = \sum_{i \in U} \mathbf{x}_i$ are known. Within the design-based framework, the sampling weights $d_i = 1/\pi_i$ can be calibrated to the known totals $\mathbf{T}_\mathbf{x}$; see Deville and Särndal (1992). Alternatively, model-based methods (Chambers , 1996, 1997) could also be considered to determine calibrated weights. Let $w_i$ denote the calibrated weight for unit $i$ and $\hat{T}_y^C = \sum_{i \in s} w_i y_i$ denote the calibrated estimator for the total of $y$. The weights $w_i$ satisfy the calibration equation $\sum_{i \in s} w_i \mathbf{x}_i = \mathbf{T}_\mathbf{x}$. The calibration is motivated by the following linear estimation model

$$m : E_m(y_i|\mathbf{X}) = \mathbf{x}_i'\boldsymbol{\beta} \quad \text{and} \quad \text{Var}_m(y_i|\mathbf{X}) \propto v_i = \mathbf{x}_i'\boldsymbol{\lambda}, \quad i \in U,$$

where the subscript $m$ indicates that the moments are evaluated with respect to the model, $\boldsymbol{\beta}$ is a vector of unknown model parameters, $\boldsymbol{\lambda}$ is known, and $\mathbf{X}$ is a $N$-row matrix containing $\mathbf{x}_i'$ in its $i$th row. The quality of the design-based inference drawn with the calibrated weights depends heavily on the validity of the model. Although standard calibration estimators are typically approximately unbiased under the sampling design, they may be quite inefficient if the model fails drastically (e.g., Hedlin et al., 2001); in particular, when the model misspecification results in the presence of outliers. It is thus useful to seek for estimators that are more efficient than standard calibration estimators in outlier-prone populations.

### 4.1. *Robust M-estimation for finite populations*

Let $\hat{\boldsymbol{\beta}}^R$ be an arbitrary outlier-resistant estimator of the regression parameter in model $m$. A decomposition analogous to that of Chambers (1986) is useful to illustrate that calibration estimators are vulnerable to outliers. Since the $\mathbf{x}$-variables satisfy the calibration equation $\sum_{i \in U} \mathbf{x}_i = \sum_{i \in s} w_i \mathbf{x}_i$, one has

$$\hat{T}_y^C = \sum_{i \in s} y_i + \sum_{i \in U-s} \mathbf{x}_i'\hat{\boldsymbol{\beta}}^R + \sum_{i \in s} u_i \frac{\left(y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}^R\right)}{\sqrt{v_i}}, \tag{5}$$

where $u_i = (w_i - 1)\sqrt{v_i}$. Thus,

$$\hat{T}_y^C - T_y = -\sum_{i \in U-s} \left(y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}^R\right) + \sum_{i \in s} u_i \left(y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}^R\right)/\sqrt{v_i}.$$

On the one hand, this expression for the sampling error of $\hat{T}_y^C$ highlights that extreme (positive or negative) residuals $y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}^R$ for nonsample units may have a substantial impact on the sampling error; nothing can be done about that at the estimation stage. On the other hand, a sample unit with a large weight $u_i$, combined to a large standardized residual $(y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}^R)/\sqrt{v_i}$, may account for an important share of the sampling error and may thus be called an influential unit. Methods to downweight its contribution to the sampling error of the calibration estimator are proposed in this section.

We use Schweppe form of the generalized M-estimator to limit the impact of outliers; see Chapter 6 of Hampel et al. (1986). This leads to

$$\hat{T}_y^{RC} = \sum_{i\in s} y_i + \sum_{i\in U-s} \mathbf{x}_i'\hat{\boldsymbol{\beta}}^R + \sum_{i\in s} \frac{u_i}{h_i}\psi\left(h_i\frac{\left(y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}^R\right)}{\sqrt{v_i}}\right), \tag{6}$$

where $h_i$ is a weight that is allowed to depend on both $\mathbf{x}_i$ and $d_i$, and $\psi(t)$ is a bounded function with $\psi(0) = 0$ and $\psi(t) \approx t$ when $t$ is close to 0. If $\psi(t) = t$ for every value of $t$ then $\hat{T}_y^{RC}$ reduces to the nonrobust calibration estimator $\hat{T}_y^{C}$ in (5). The Huber $\psi$-function,

$$\psi(t) = \psi_H^I(t) \equiv \begin{cases} t & \text{if } |t| < c \\ \text{sign}\,(t) \times c & \text{if } |t| \geq c \end{cases},$$

is widely used in this context. An interesting property of estimator (6) is that it is census-consistent in the sense that, no matter the choice of $h_i$ or $\psi(.)$, $\hat{T}_y^{RC}$ reduces to $T_y$ if a census is conducted (i.e., when $s = U$ and $w_i = 1$, for $i \in U$).

The robust calibration estimator (6) can also be written in the form

$$\hat{T}_y^{RC} = \mathbf{T}_\mathbf{x}'\hat{\boldsymbol{\beta}}^R + \sum_{i\in s} w_{r,i}(\hat{\boldsymbol{\beta}}^R)(y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}^R), \tag{7}$$

with weights

$$w_{r,i}\left(\hat{\boldsymbol{\beta}}^R\right) = w_i \frac{\psi_i^*\left(h_i\left(y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}^R\right)/\sqrt{v_i}\right)}{h_i\left(y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}^R\right)/\sqrt{v_i}}$$

and

$$\psi_i^*\,(t) = \frac{t}{w_i} + \frac{(w_i - 1)}{w_i}\psi\,(t)\,.$$

Expression (7) has the same form as the estimator studied in Duchesne (1999). If $\psi(t) = \psi_H^I(t)$, the resulting function $\psi_i^*(t)$ provides an outlier treatment analogous to type II Winsorization since it reduces only $100(w_i - 1)/w_i\%$ of a weighted residual. From this perspective, the standard Huber function $\psi_i^*(t) = \psi_H^I(t)$ gives a two-sided type I Winsorization. Alternatives to (6) are available to limit the impact of large sample residuals. For instance, Lee (1991) multiplies the residual component of (5) by a number $\theta$ between 0 and 1; see also Gwet (1998) and Lee and Patak (1998).

Many choices for the weights $h_i$ have been proposed in the literature. They include $h_i = 1$ (Chambers , 1986, 1997), $h_i \propto \sqrt{v_i}$ (Gwet and Rivest, 1992), and $h_i = w_i\sqrt{v_i}$ or $h_i = d_i\sqrt{v_i}$ (Beaumont and Alavi, 2004). Equation (5) suggests $h_i = u_i = (w_i - 1)\sqrt{v_i}$ as another interesting set of weights.

Several strategies are available to construct $\hat{\boldsymbol{\beta}}^R$. In a model-based approach, $\hat{\boldsymbol{\beta}}^R$ is derived from the theory of Robust Statistics to represent the relationship between $\mathbf{x}$ and $y$ for the bulk of the data values. It should be outlier-resistant and should achieve a high efficiency when the model errors have a normal distribution (see Hampel et al., 1986). Then, $\sum_{i\in U-s}\mathbf{x}_i'\hat{\boldsymbol{\beta}}^R$ is little affected by outliers; it may however suffer from a substantial bias as an estimator of $\sum_{i\in U-s} y_i$ in outlier-prone populations. The role of the third term in (6) is to reduce this bias. The function $\psi$ in (6) brings outliers in the estimation of $T_y$. Hopefully, the resulting increase in variance will be offset by a bias reduction. The management of this bias-variance trade-off is the key to producing estimators of $T_y$ with good sampling properties. Variants of this model-based approach are presented in Chambers (1986) and Welsh and Ronchetti (1998). Several tuning constants, for $\hat{\boldsymbol{\beta}}^R$ and $\psi$, are involved; they cannot be determined using design-based information only.

From a design-based perspective, $\boldsymbol{\beta}$ is a nuisance parameter since the focus is the estimation of $T_y$. Considering (7), a suitable estimating equation for $\boldsymbol{\beta}$ is $\mathbf{U}(\boldsymbol{\beta}) = \mathbf{0}$, where

$$\mathbf{U}(\boldsymbol{\beta}) = \sum_{i \in s} w_{r,i}(\boldsymbol{\beta})(y_i - \mathbf{x}_i'\boldsymbol{\beta})\frac{\mathbf{x}_i}{v_i} = \sum_{i \in s} \frac{w_i}{h_i}\psi_i^*\left(h_i\frac{(y_i - \mathbf{x}_i'\boldsymbol{\beta})}{\sqrt{v_i}}\right)\frac{\mathbf{x}_i}{\sqrt{v_i}}. \tag{8}$$

Let $\hat{\boldsymbol{\beta}}^{GM}$ denote the solution of $\mathbf{U}(\boldsymbol{\beta}) = \mathbf{0}$. The associated robust calibration estimator (7), obtained with $\hat{\boldsymbol{\beta}}^R = \hat{\boldsymbol{\beta}}^{GM}$, takes the simple projection form

$$\hat{T}_y^{RC} = \mathbf{T}_\mathbf{x}'\hat{\boldsymbol{\beta}}^{GM}. \tag{9}$$

Gwet and Rivest (1992); Hulliger (1995), and Beaumont and Alavi (2004) investigated estimator having the form (9). In (9), the outlier treatment is done through the tuning constant $c$ of Huber $\psi$-function. Before discussing the specification of this tuning constant, it is interesting to draw a parallel between estimator (9) and the Winsorized estimator (4) in a stratified sampling design, which was presented in Section 3.1.

### *4.2. Some examples*

The Winsorized estimators of Section 3 are associated to one-sided Huber $\psi$-functions, $\psi(t) = \min(t, R)$. This section shows that they are special cases of the robust calibration estimator (9). Alternatives to the Horvitz-Thompson are also developed in the framework of Section 4.1.

In a stratified design where stratum membership is the only auxiliary information, a unit $i$ in stratum $h$ has $\mathbf{x}_i'\boldsymbol{\beta} = \mu_h$, $v_i = 1, \mathbf{T}_\mathbf{x} = (N_1, \ldots, N_H)', w_i = d_i = 1/f_h$, and $\sum_{i \in s_h} w_i = N_h$. Taking $\psi(t) = \min(t, R)$ and $h_i = u_i = (1 - f_h)/f_h$, the robust estimator of $\mu_h$ obtained from the estimating equation $\mathbf{U}(\boldsymbol{\beta}) = \mathbf{0}$, see (8), is a solution of

$$\mu_h = f_h\bar{y}_{sh} + \frac{(1 - f_h)}{n_h}\sum_{i \in s_h}\min[y_{hi}, f_hR/(1 - f_h) + \mu_h].$$

Thus $\hat{T}_y^{RC} = \sum_{h=1}^H N_h\hat{\mu}_h = \hat{T}_{yR}$, is the Winsorized estimator of Section 3.1 with a cutoff $R_h = \hat{\mu}_h + R\frac{f_h}{1-f_h}$ that has a form similar to the optimal cutoff of Section 3.1. The Winsorized estimator (4) can be regarded as a special case of (9).

The Horvitz–Thompson estimator can be fitted in the framework developed in Section 4.1. First suppose that model $m$ is empty, with no auxiliary variable, and $v_i = 1$. Set $u_i = h_i = (1/\pi_i - 1)$, (7) leads to

$$\hat{T}_{yR}^{HT} = \sum_{i \in s} y_i + \min((1/\pi_i - 1)y_i, R).$$

This is a type II alternative to the estimator presented in Section 3.2.

Suppose now that $m$ has only an intercept $\mu$ and $v_i = 1$. The calibrated weights $w_i$ sum to $N$. If $u_i = h_i = (w_i - 1)$, then the robust estimator of $\mu$ obtained from the estimating equation $\mathbf{U}(\boldsymbol{\beta}) = \mathbf{0}$, see (8), with $\psi(t) = \min(t, R)$ is a solution of

$$\mu = \frac{\sum_{i \in s} y_i + (w_i - 1)\min[y_i, \mu + R/(w_i - 1)]}{\sum_{i \in s} w_i},$$

and $\hat{T}_y^{RC} = \sum_{i \in s} w_i \hat{\mu}$.

Finally, the Horvitz–Thompson weights $d_i$ are sometimes implicitly calibrated to known auxiliary totals; that is, $\sum_{i \in s} d_i \mathbf{x}_i = \mathbf{T_x}$ for a suitably defined vector $\mathbf{x}$. This has already been shown for stratified simple random sampling in the above example. This is also true with probability-proportional-to-size (pps) sampling as the selection probability for unit $i$ is $\pi_i = nx_i/T_x$ and thus $\sum_{i \in s} d_i x_i = T_x$. When a sampling design is used such that the Horvitz–Thompson estimator is implicitly calibrated, estimator (9) may be directly used for its improvement by letting $w_i = d_i$. For pps sampling, this was done by Hulliger (1995), who used $\psi_i^*(t) = \psi_H^I(t)$.

### 4.3. *Choice of the tuning constant when solving* $\mathbf{U}(\boldsymbol{\beta}) = \mathbf{0}$

The determination of the tuning constant $c$ for Huber $\psi$-function appearing in the estimating equation for $\hat{\boldsymbol{\beta}}^{GM}$ is critical for the construction of a good estimator for $T_y$. Section 3 was facing a similar problem when determining the optimal Winsorization cutoff. Too small a value produces a small variance at the expense of a large bias, whereas a large value gives the nonrobust calibration estimator $\hat{T}_y^C$. Therefore, the MSE is a useful criterion for evaluating the quality of outlier-robust estimators of finite population parameters and for choosing an appropriate tuning constant $c$.

A simple strategy sets $c$ equal to a residual scale parameter estimate $Q$ times a fixed constant $c^*$ (e.g., $c^* = 2$ or $c^* = 10$). With this ad hoc approach, one cannot be sure that $\hat{T}_y^{RC}$ is more accurate than the standard calibration estimator. Following the strategy used in Section 3, one can select the value of $c$ that minimizes $\text{MSE}(\hat{T}_y^{RC}) = E(\hat{T}_y^{RC} - T_y)^2$ for a known population similar to that under study, or perhaps by pooling several samples together, as illustrated in Section 2.5.

When additional information is not available, Hulliger (1995) and Beaumont and Alavi (2004) suggested to find the tuning constant $c$ that minimizes an MSE estimator $\text{mse}(\hat{T}_y^{RC})$. As pointed out by Hulliger, this approach does not depend on the choice of the scaling statistic $Q$ so that it can be set to 1. This is the approach pursued now as a known population similar to that under study is often not easily available. Assuming that $E(\hat{T}_y^C) \approx T_y$, the MSE of $\hat{T}_y^{RC}$ can be approximated as

$$\text{MSE}\big(\hat{T}_y^{RC}\big) = E\big(\hat{T}_y^{RC} - T_y\big)^2 \approx \text{Var}\big(\hat{T}_y^{RC}\big) + \Big\{E\big(\hat{T}_y^{RC} - \hat{T}_y^C\big)^2 - \text{Var}\big(\hat{T}_y^{RC} - \hat{T}_y^C\big)\Big\}, \tag{10}$$

The two terms within brackets in (10) are an approximation of the square of the design bias of $\hat{T}_y^{RC}$. Gwet and Rivest (1992) suggested the MSE estimator

$$\text{mse}\big(\hat{T}_y^{RC}\big) = v\big(\hat{T}_y^{RC}\big) + \max\Big\{0, \big(\hat{T}_y^{RC} - \hat{T}_y^C\big)^2 - v\big(\hat{T}_y^{RC} - \hat{T}_y^C\big)\Big\}, \tag{11}$$

where $v\big(\hat{T}_y^{RC}\big)$ and $v\big(\hat{T}_y^{RC} - \hat{T}_y^{C}\big)$ are design-consistent estimators of $\mathrm{Var}\big(\hat{T}_y^{RC}\big)$ and $\mathrm{Var}\big(\hat{T}_y^{RC} - \hat{T}_y^{C}\big)$, respectively. When the goal of MSE estimation is only to find a suitable tuning constant $c$, the following simplified MSE estimator is often more appealing in practice:

$$\mathrm{mse}^*\big(\hat{T}_y^{RC}\big) = v\big(\hat{T}_y^{RC}\big) + \big(\hat{T}_y^{RC} - \hat{T}_y^{C}\big)^2. \tag{12}$$

Then, $v\big(\hat{T}_y^{RC}\big)$ may be estimated using a simple linearization technique. First one can obtain $\hat{\mathbf{V}}\big(\hat{\boldsymbol{\beta}}^{GM}\big)$, an estimator of $\mathrm{Var}\,(\hat{\boldsymbol{\beta}}^{GM})$, using the estimating function (8) and the linearization method of Binder (1983); that is,

$$\hat{\mathbf{V}}(\hat{\boldsymbol{\beta}}^{GM}) = \left\{\frac{\partial \mathbf{U}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right\}^{-1} \hat{\mathbf{V}}\left\{\mathbf{U}(\boldsymbol{\beta})\right\}\left\{\left(\frac{\partial \mathbf{U}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right)'\right\}^{-1}\Bigg|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}^{GM}}. \tag{13}$$

To simplify the estimation of the middle term in the right-hand side of (13), we treat both $w_i$ and $h_i$ as fixed quantities. From (9), we then have $v\big(\hat{T}_y^{RC}\big) = \mathbf{T}_x'\hat{\mathbf{V}}(\hat{\boldsymbol{\beta}}^{GM})\mathbf{T}_{\mathbf{x}}$ and (see Beaumont, 2004, for additional details),

$$v\big(\hat{T}_y^{RC}\big) = \sum_{i\in s}\sum_{j\in s}\frac{\pi_{ij} - \pi_i\pi_j}{\pi_{ij}} e_i e_j, \tag{14}$$

where

$$e_i = \frac{w_{r,i}\big(\hat{\boldsymbol{\beta}}^{GM}\big)\left(y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}^{GM}\right)}{v_i}\mathbf{x}_i'\left(\sum_{j\in s}\gamma_j\mathbf{x}_j\mathbf{x}_j'\right)^{-1}\mathbf{T}_{\mathbf{x}}, \text{ and}$$

$$\gamma_i = \frac{w_i}{v_i}\;\frac{\partial\psi_i^*(t)}{\partial t}\Bigg|_{t=h_i\frac{\left(y_i-\mathbf{x}_i'\hat{\boldsymbol{\beta}}^{GM}\right)}{\sqrt{v_i}}}.$$

Once the $e_i$'s have been computed, the variance estimator (14) is easy to obtain using standard software packages for survey sampling.

The minimization of (12) with respect to $c$ can be achieved using the Newton–Raphson algorithm with the first and second derivatives computed numerically. This is what has been done in Section 4.6. Since current data are used to determine the tuning constant, there is some instability in the estimated $c$. To increase stability, one can consider using the average estimated $c$ or $c*$ over several periods of a survey, when such data are available.

### *4.4. Estimation of the MSE*

Once the tuning constant $c$ has been estimated, it is usually required to evaluate the MSE of the resulting robust estimator $\hat{T}_y^{RC}$. The variability added by the estimation of $c$ must be taken into account in MSE estimation. To this end, we use the bootstrap technique (e.g., Rao and Wu, 1988; Rao et al., 1992) to obtain variance estimators involved in (11). For each bootstrap replicate, the complete estimation process is repeated, including the minimization used to determine the tuning constant. The resulting bootstrap variance accounts for the fact that $w_i$ and $h_i$ are random. Gwet and Lee (2000) studied empirically the bootstrap in the context of outlier-robust estimation and found promising results.

When the residuals in (5) have a skewed distribution, Section 2 suggests that selecting a tuning constant $c$ that curbs the contribution of 1 or 2 data points should bring a reduction in MSE. As argued in Section 2.3, this reduction might not be detectable by standard asymptotic calculations, such as the Taylor or bootstrap procedure, presented in this section. More work is needed to adapt the techniques for estimating the Winsorization cutoff presented in Section 2 to the calibration estimator of Section 4.

### 4.5. Implementation

Users of survey data are accustomed to work with a complete rectangular data file containing a unique set of estimation weights. Unlike the calibration estimator $\hat{T}_y^C$, the robust calibration estimator $\hat{T}_y^{RC}$ cannot be implemented easily if only the calibration weights $w_i$, for $i \in s$, are provided along with the original values of the $y$-variables. In this section, we describe two approaches to deal with this issue: a weighting approach (Beaumont and Alavi, 2004) and an imputation approach (e.g., Beaumont and Alavi, 2004; Chambers and Kokic, 1993; Ren and Chambers, 2002).

Let $\mathbf{T_y} = \sum_{i \in s} \mathbf{y}_i$ be the vector of population totals for the $q$ variables of interest $\mathbf{y} = (y_1, \ldots, y_q)'$, with corresponding robust calibration estimators $\hat{\mathbf{T}}_{\mathbf{y}}^{RC} = \left(\hat{T}_{y_1}^{RC}, \ldots, \hat{T}_{y_q}^{RC}\right)'$. In the weighting approach, $\mathbf{y}_i$, for $i \in s$, are kept intact but the calibration weights $w_i$, for $i \in s$, are replaced by the robust calibration weights $w_i^R$, which are obtained by using the augmented calibration equation

$$\sum_{i \in s} w_i^R \begin{pmatrix} \mathbf{x}_i \\ \mathbf{y}_i \end{pmatrix} = \begin{pmatrix} \mathbf{T_x} \\ \hat{\mathbf{T}}_{\mathbf{y}}^{RC} \end{pmatrix}.$$

Of course, there may be a limit on the number of $\boldsymbol{y}$-variables that can be used for calibration purposes. This may somewhat restrict the applicability of this method when $q$ is very large.

An alternative is to modify the values of the $\boldsymbol{y}$-variables while keeping the calibration weights $w_i$ intact. It is applied separately for each variable of interest and is called the imputation approach. It is straightforward to show that the robust calibration estimator $\hat{T}_y^{RC}$ in (6) can be written as $\hat{T}_y^{RC} = \sum_{i \in s} w_i y_{\cdot i}$, where

$$y_{\cdot i} = \frac{y_i}{w_i} + \frac{(w_i - 1)}{w_i} \left\{ \mathbf{x}_i' \hat{\boldsymbol{\beta}}^R + \frac{\sqrt{v_i}}{h_i} \psi\left( h_i \frac{\left(y_i - \mathbf{x}_i' \hat{\boldsymbol{\beta}}^R\right)}{\sqrt{v_i}} \right) \right\}. \tag{15}$$

When a Huber $\psi$-function, with tuning constant $c$, is used, Eq. (15) reduces to

$$y_{\cdot i} = \begin{cases} y_i & , \text{ if } i \in s - s_o \\ \dfrac{y_i}{w_i} + \dfrac{(w_i - 1)}{w_i} \left\{ \mathbf{x}_i' \hat{\boldsymbol{\beta}}^R + \dfrac{\sqrt{v_i}}{h_i} \text{sign}\left( h_i \dfrac{\left(y_i - \mathbf{x}_i' \hat{\boldsymbol{\beta}}^R\right)}{\sqrt{v_i}} \right) c \right\} & , \text{ if } i \in s_o \ , \end{cases} \tag{16}$$

where $s_o$ is the set of all sample units $i$ for which $w_{r,i}\left(\hat{\boldsymbol{\beta}}^R\right) < w_i$. Thus, $s_o$ contains units that have been detected as being influential by the robust procedure. The modified values (16) can be viewed as type II symmetric two-sided regression Winsorized values and the corresponding robust estimator could be called a Winsorized regression estimator of $T_y$. The expression for the Winsorized values in Section 3.1 of this Chapter as well as those given in Kokic and Bell (1994), Clarke (1995), and Chambers et al. (2000) can be obtained from a one-sided version of (16) with $h_i = u_i = (w_i - 1)\sqrt{v_i}$. Plugging $h_i = 1$ in (16) leads to the Winsorized ratio estimator of Chambers and Kokic (1993).

Considering (7), an alternative form for $\hat{T}_y^{RC}$ is $\hat{T}_y^{RC} = \sum_{i \in s} w_i y_{\cdot i}$, where

$$y_{\cdot i} = \frac{w_{r,i}(\hat{\boldsymbol{\beta}}^R)}{w_i} y_i + \left(1 - \frac{w_{r,i}(\hat{\boldsymbol{\beta}}^R)}{w_i}\right) \mathbf{x}_i' \hat{\boldsymbol{\beta}}^R. \tag{17}$$

The modified values in Eq. (17) are equivalent to those given in (15) and have a simple interpretation: they are a weighted average of the robust predictions $\mathbf{x}_i'\hat{\boldsymbol{\beta}}^R$ and the observed values $y_i$. Less weight is given to the observed value $y_i$ when it has a smaller value of $w_{r,i}(\hat{\boldsymbol{\beta}}^R)/w_i$ and, therefore, when it is highly influential. The imputation approach can also be implemented through reverse calibration by using a suitable distance function (Beaumont and Alavi, 2004; Ren and Chambers, 2002).

### *4.6. An illustration using the Canadian Workplace and Employee Survey*

To illustrate the usefulness of robust M-estimators, we use the data of the 2003 Canadian Workplace and Employee Survey (CWES). The CWES is a longitudinal survey that started in 1999 and that collects information on employers and their employees. In this application, we focus on the employer portion of the CWES. Every other year, a sample from the population of births is selected from the Business Register. Therefore, the 2003 sample contains units selected in 1999, 2001, and 2003. In each of these years, employers are selected by stratified simple random sampling without replacement and the strata are formed by crossing 6 regions, 14 industry groups, and 3 size groups, where the size variable corresponds to the number of employees available on the Business Register. To simplify the example, we restrict to a single region-industry group, which yields a sample of size $n = 112$, and consider only the most important variables of the survey. They are five financial variables, which will be denoted by $Y_1$, $Y_2$, $Y_3$, $Y_4$, and $Y_5$. In this survey, there is a single auxiliary variable $x$ with $v_i = x_i$ and $w_i = d_i T_x / \sum_{i \in s} d_i x_i$, where $d_i$ is the inverse sampling fraction in the stratum of unit $i$. Thus, the calibration estimator is actually a ratio estimator. The sample values of this auxiliary variable are obtained at the collection stage and correspond to the number of employees of each sampled employer. The population total $T_x$ is obtained from a reliable external survey (the Survey of Employment Payroll and Hours); $T_x$ is assumed to be without error to simplify the example.

Figure 3 shows the relationship between $x$ and $Y_1$. For variable $Y_4$, this relationship is illustrated in Fig. 4. The solid line in both figures corresponds to the least-squares fit. In both cases, there are large regression residuals and the ratio model seems to be reasonable although it may not be fully satisfactory. A similar observation was also made for the other three variables, which are not illustrated here. Figures 5 and 6 show a plot of the standardized residuals $(y_i - x_i\hat{\beta})/\sqrt{x_i}$ versus $u_i = (w_i - 1)\sqrt{v_i}$, for $Y_1$ and $Y_4$, respectively, where $\hat{\beta} = \sum_{i \in s} w_i y_i / \sum_{i \in s} w_i x_i$. This plot is useful to determine whether there are influential units in the sample. A unit that has a large value on both axes is most

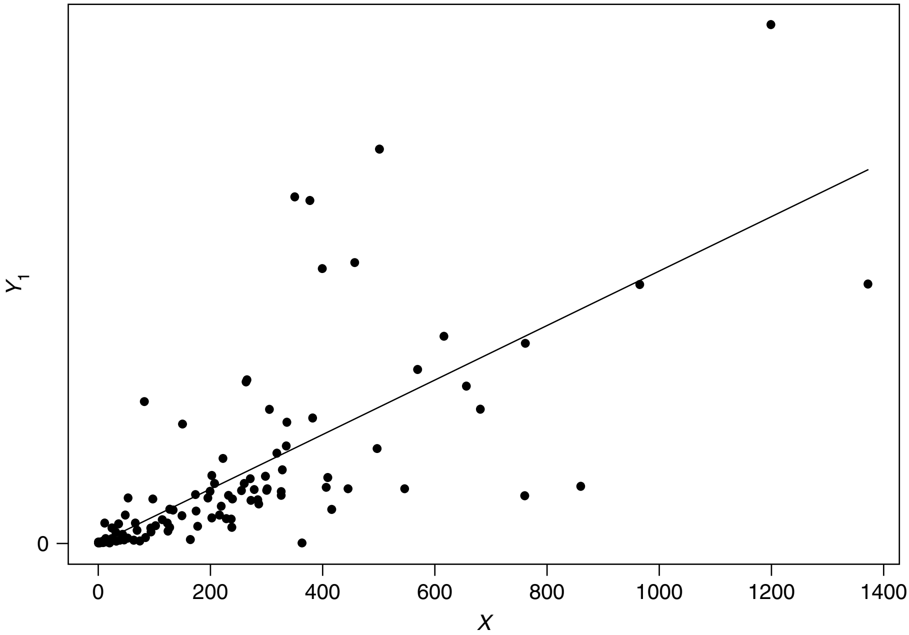


Fig. 3. Plot of $Y_1$ versus the auxiliary variable.

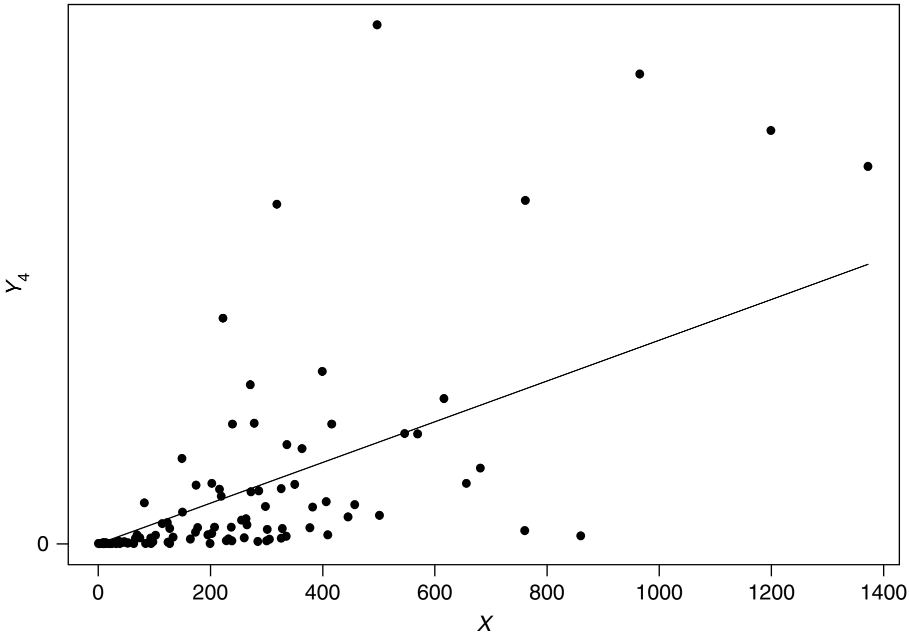


Fig. 4. Plot of $Y_4$ versus the auxiliary variable.

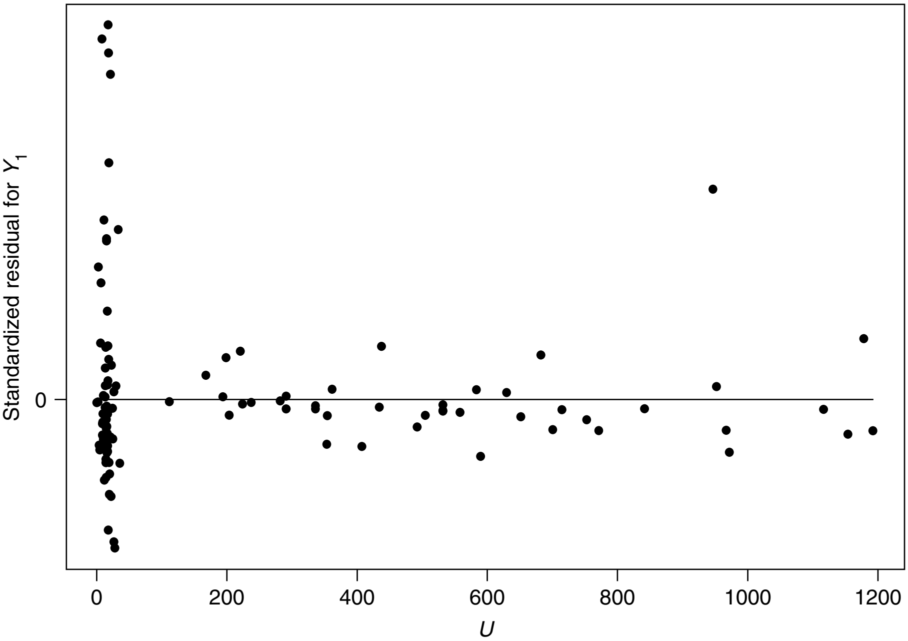


Fig. 5. Plot of the standardized residual for $Y_1$ versus $u$.

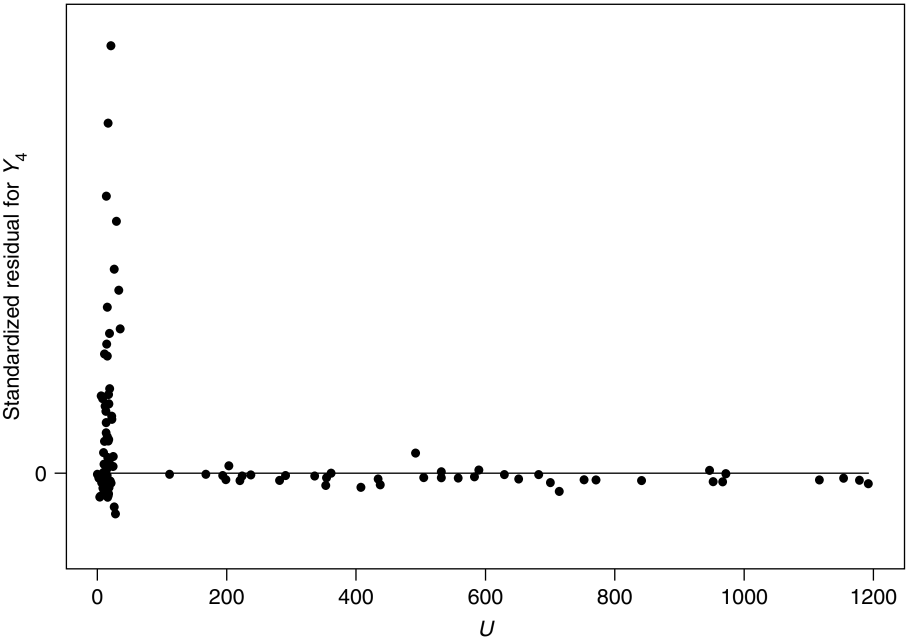


Fig. 6. Plot of the standardized residual for $Y_4$ versus $u$.

likely quite influential. From these two figures, it appears that there is an influential unit for variable $Y_1$ and none for variable $Y_4$. Thus, the large regression residuals in Fig. 4 do not seem to be influential. The other residual plots are not provided here: variable $Y_2$ is similar to $Y_1$, it has one influential unit, whereas variables $Y_3$ and $Y_5$ have two influential units, although less influential than the one for variable $Y_1$.

In our empirical investigation, we considered the robust estimator (9) and solved $\mathbf{U}(\boldsymbol{\beta}) = \mathbf{0}$ (see Eq. 8) using the iteratively reweighted least squares algorithm (Beaton and Tukey, 1974) starting with the least squares solution. We compared two different choices of weight $h_i$: $h_i = u_i = (w_i - 1)\sqrt{v_i}$ and $h_i = 1$. We also compared four different choices of $c$ : $c = 0.00001Q, c = 2Q$, $c = 10Q$, and the optimal value of $c$ obtained by minimizing the estimated MSE given by (12) and (14). The scale statistic $Q$ required in the first three cases is obtained as in Beaumont and Alavi (2004). It is proportional to a weighted median of absolute weighted residuals. Note that we kept this statistic unchanged throughout the iterations of the iteratively reweighted least squares algorithm for simplicity. Results are given in Table 4. Four quantities have been computed: i) the relative difference in percentage RD $= 100\big(\hat{T}_y^{RC} - \hat{T}_y^{C}\big)/\hat{T}_y^{C}$; ii) the Wald statistic Wald $= \big(\hat{T}_y^{RC} - \hat{T}_y^{C}\big)^2/v_B\big(\hat{T}_y^{RC} - \hat{T}_y^{C}\big)$, which can be used to give an indication of the design bias of the robust calibration estimator; iii) the bootstrap relative efficiency in percentage RE_Bootstrap $=$ $100v_B\big(\hat{T}_y^{C}\big)/\mathrm{mse}_B\big(\hat{T}_y^{RC}\big)$, and iv) the Taylor relative

Table 4
Comparison of different robust estimators on the CWES data

| Variable | $C$ | RD | Wald | RE_Bootstrap | RE_Taylor | RD | Wald | RE_Bootstrap | RE_Taylor |
|---|---|---|---|---|---|---|---|---|---|
| | | $h_i = u_i$ | | | | $h_i = 1$ | | | |
| $Y_1$ | $0.00001Q$ | 8.01 | 0.32 | 598.5 | 233.8 | 8.01 | 0.32 | 598.5 | 233.8 |
| | $2Q$ | −8.13 | 0.39 | 671.2 | 263.2 | −20.14 | 4.85 | 59.1 | 48.6 |
| | $10Q$ | −13.76 | 1.29 | 242.1 | 100.6 | −2.67 | 5.74 | 94.5 | 82.7 |
| | Optimal | 1.09 | 0.02 | 246.0 | 667.9 | 1.39 | 0.04 | 198.1 | 442.0 |
| $Y_2$ | $0.00001Q$ | 8.90 | 0.32 | 647.7 | 247.1 | 8.90 | 0.32 | 647.7 | 247.1 |
| | $2Q$ | −9.04 | 0.35 | 1181.6 | 294.4 | −24.43 | 4.52 | 58.9 | 47.6 |
| | $10Q$ | −17.34 | 1.43 | 225.6 | 91.5 | −4.67 | 5.31 | 98.8 | 90.4 |
| | Optimal | 0.72 | 0.01 | 230.6 | 767.7 | 0.97 | 0.02 | 202.2 | 638.1 |
| $Y_3$ | $0.00001Q$ | 19.45 | 8.83 | 10.4 | 9.3 | 19.45 | 8.82 | 10.4 | 9.3 |
| | $2Q$ | 7.77 | 2.42 | 68.8 | 44.0 | −5.85 | 6.64 | 65.1 | 57.4 |
| | $10Q$ | −0.46 | 0.01 | 145.6 | 128.0 | −0.93 | 5.41 | 95.3 | 88.3 |
| | Optimal | 1.86 | 0.63 | 104.9 | 135.3 | 1.28 | 0.75 | 105.9 | 215.9 |
| $Y_4$ | $0.00001Q$ | 157.41 | 76.54 | 0.4 | 0.4 | 157.41 | 76.54 | 0.4 | 0.4 |
| | $2Q$ | 52.21 | 15.00 | 3.6 | 3.4 | −20.34 | 14.37 | 22.2 | 21.3 |
| | $10Q$ | 3.76 | 1.53 | 72.9 | 68.1 | −11.06 | 11.71 | 53.6 | 50.6 |
| | Optimal | 0 | – | 100 | 100.7 | −1.38 | 1.35 | 103.1 | 102.7 |
| $Y_5$ | $0.00001Q$ | 50.03 | 7.02 | 12.6 | 11.0 | 50.02 | 7.02 | 12.6 | 11.0 |
| | $2Q$ | −2.89 | 0.04 | 216.7 | 212.9 | −27.34 | 11.13 | 36.4 | 33.7 |
| | $10Q$ | −20.26 | 3.43 | 71.2 | 54.5 | −3.52 | 5.45 | 93.7 | 83.5 |
| | Optimal | 2.32 | 0.46 | 112.2 | 226.6 | 1.98 | 0.80 | 98.1 | 137.6 |

efficiency in percentage RE_Taylor $= 100v_B(\hat{T}_y^C)/\text{mse}^*(\hat{T}_y^{RC})$. The notation $v_B(.)$ is used to denote the Rao–Wu bootstrap variance estimator. The estimator $\text{mse}_B(\hat{T}_y^{RC})$ is obtained by using the Rao–Wu bootstrap method for the estimation of the two variance terms appearing in (11). One thousand bootstrap replicates have been used in this empirical study. Finally, $\text{mse}^*(\hat{T}_y^{RC})$ is obtained from (12) and (14).

From Table 4, we can make the following remarks:

- On the one hand, the robust calibration estimator $\hat{T}_y^{RC}$ is often more efficient than the nonrobust calibration estimator $\hat{T}_y^C$ when $\hat{T}_y^{RC}$ is not significantly biased; that is, when the Wald statistic is small (say smaller than 3.84). On the other hand, it may be much less efficient than $\hat{T}_y^C$ when the Wald statistic is large.
- The optimal $c$, estimated using the procedure outlined in Section 4.3, always leads to a small value of the Wald statistic and seems to offer a good compromise between bias and variance as it is almost always more efficient than $\hat{T}_y^C$. However, it does not always lead to the most efficient robust calibration estimator due to the increase in variance resulting from estimating $c$. This can be noted by examining the relative efficiencies.
- Note that the optimal choice of $c$ led to a value $c^* = c/Q$ smaller than 2 in most cases but larger than 10 for variable $Y_4$.
- An ad hoc choice of $c^*$ may result in large gains in efficiency, especially when $c^*$ is small. It may also lead to quite an inefficient robust estimator when $c^*$ is small. It thus seems difficult to determine a fixed constant $c^*$ that performs reasonably well in all situations. A larger value of $c^*$ seems preferable to avoid large biases even though it may also reduce efficiency gains.
- The choice $h_i = u_i$ performed better in general than $h_i = 1$ although the difference is small for the optimal choice of $c$.
- The gain in efficiency for variable $Y_4$ was smaller than that for variable $Y_1$. This is not surprising as there was no influential unit in Fig. 6, whereas there was one in Fig. 5. For variable $Y_4$, when $h_i = u_i$, we were indeed not able to find any value of $c$ such that $\text{mse}\,^*(\hat{T}_y^{RC}) < \text{mse}\,^*(\hat{T}_y^C)$ so that the optimal robust estimator was actually the calibration estimator $\hat{T}_y^C$.
- More gains in efficiency could potentially be obtained by finding an optimal value of $c$ using past data. In the absence of past information, the optimal robust calibration estimator shown in this empirical study is attractive and performed well overall.

Finally, it is important to point out that the estimated MSE (12) had more than one local minimum. The local minimum with the largest value of $c$ (not leading to $\hat{T}_y^{RC} = \hat{T}_y^C$) was usually not the global minimum. It was associated to a negative bias with only a few units detected as being influential. It would normally be the global minimum if a one-sided Huber function was used. There was also another minimum with a value of $c$ closer to zero and which was usually the global minimum. This global minimum was associated to a small bias, not necessarily negative. This is in agreement with Kokic (1998) who found out that two-sided Winsorization can substantially reduce the bias compared to one-sided Winsorization. Note that this global minimum may lead to modifying more than 50% of the $y$-values, especially when $h_i = 1$ is used. Indeed, only a few $y$-values were not modified in some cases. Therefore, implementation of $\hat{T}_y^{RC}$

using the imputation approach discussed in Section 4.5 may not be attractive for users. The weighting approach may be more suitable in this context.

Unfortunately, the global minimum was somewhat difficult to find using the Newton–Raphson algorithm as the derivative of the estimated MSE in the neighborhood of this minimum was quite large compared to the derivative in the neighborhood of the local minimum with a larger value of $c$. To find the global minimum in the original sample, we fed the Newton–Raphson algorithm with several different initial values and also scrutinized the estimated MSE curve. This could not be done for the 1000 bootstrap replicates. Instead, for a given bootstrap replicate, we started the iterative process with the optimal value of $c$ found with the original sample and stopped after a maximum of five iterations to save on computer time.

## 5. Dealing with stratum jumpers

Standard design-based estimators, such as the Horvitz–Thompson estimator and the calibration estimators, are known to be inefficient when the estimation weights are highly variable and uncorrelated with the variables of interest (e.g., Basu, 1971; Rao, 1966). A consequence of a high variability in the estimation weights is the presence of extreme weights, sometimes referred to as outlier weights. This problem often occurs in household surveys for a number of reasons, including the use of a series of weight adjustments. In business surveys, stratification by size and common allocation schemes usually results in variable design weights. This may yield inefficient estimates when there are stratum jumpers. This is the problem on which we focus in this section.

### *5.1. The problem of stratum jumpers in business surveys*

In many business surveys, the population of businesses is stratified by region, industry type, and size group. The latter is defined using some measure of business size available on the sampling frame, such as the number of employees or the revenue of the business. Then, the sample is usually selected by stratified simple random sampling without replacement. For efficiency considerations, a large selection probability $\pi_i$ (and thus a small design weight, $d_i = 1/\pi_i$) is usually assigned to a unit $i$ of large size on the sampling frame, whereas a small selection probability (and thus a large design weight) is assigned to a unit of small size. This strategy is justified on the grounds that business survey variables are usually highly skewed and that there is usually a positive correlation between the size measure and the main variables of interest so that large values of these variables are expected to be assigned to a small design weight.

At the time of collection, we often observe discrepancies between the information available at the design stage and the same information collected from the respondent. These discrepancies can be explained by errors on the sampling frame, which are partly due to outdated information, and the time lag between sampling and data collection. They become problematic when a unit that is thought to be of small size at the design stage is actually found to be a large unit. Such units are sometimes called stratum jumpers because they would have been assigned to another stratum had the correct information been available at the time of design. A consequence of this problem is that some units with large values of the variables of interest are unfortunately assigned large design weights, which may result in inefficient design-based estimators.

At the design stage, the potential impact of stratum jumpers can be reduced to some extent by controlling the maximum design weight to be smaller than a certain threshold (e.g., Bocci and Beaumont, 2006). This will usually imply departing from optimal stratification and/or allocation. Rivest (1999) proposed a method for dealing with stratum jumpers, which worked well empirically. It reduces the maximum design weight by a large factor. No matter how carefully the sampling design is chosen, it is likely that the problem will not be completely eliminated as the stratum jumpers occur in a haphazard way.

As an example, suppose that there are two design strata A and B. Stratum A has nine selected units considered to be of large size at the design stage, which are assigned a design weight of 1, whereas stratum B has 41 selected units considered to be of small size at the design stage, which are assigned a design weight of 31. At the collection stage, we observe that one of the 41 units with a large weight is actually a large size unit so that the collection stratum is different from the design stratum for this unit, which is thus called a stratum jumper. Table 5 summarizes the above information. The stratum jumper is found in the middle row of this table.

Assume that the collection strata are homogeneous with respect to the variable of interest $y$ (or at least more homogeneous than the design strata) so that collection stratum A contains large size units associated with large $y$-values, whereas collection stratum B contains small size units associated with small $y$-values. On the one hand, the stratum jumper can be viewed as a unit with a large $y$-value compared to the other 40 units in the same design stratum, which all have the same design weight. Therefore, standard outlier-robust techniques, such as Winsorization (see Section 3) or M-estimation (see Section 4), can be used to handle this stratum jumper. On the other hand, the stratum jumper can also be viewed as a unit with a large design weight compared to the other nine units in the same collection stratum although it may have a similar $y$-value. This is the view we take in Section 5.3 with the weight smoothing approach.

Before describing the weight smoothing approach, let us first briefly discuss an alternative approach to handling the potential problem of extreme design weights; namely, Winsorizing the largest design weights (e.g., Liu et al., 2004; Potter, 1990). This actually seems to be the most popular method whenever something is done to deal with the problem of large design weights. Using this approach in the example given in Table 5 would lead to reducing the design weight of the stratum jumper but also of all other units in the same design stratum, which may be less appealing and may reduce efficiency. The main challenge with this method is to determine an appropriate Winsorization cutoff. Several methods, sometimes more or less ad hoc, have been considered to address this issue, including the use of outlier detection techniques. One data-driven method is to choose the cutoff so as to minimize an estimate of the design MSE. Unfortunately, this

Table 5
Example showing a stratum jumper

| Collection Stratum | Design Stratum | Number of Units | Design Weight |
|---|---|---|---|
| A | A | 9 | 1 |
| A | B | 1 | 31 |
| B | B | 40 | 31 |
| Sum over the sample units | | 50 | 1280 |

leads to a different cutoff for each variable of interest $y$, and thus a different Winsorized weight for each $y$-variable. This is not convenient in multipurpose surveys and, thus, a compromise Winsorization cutoff is needed. Also, it is worth noting that extreme design weights may not cause any problem if there is no stratum jumper and the design strata are homogeneous.

In Section 5.2, we briefly describe a weight smoothing approach, proposed by Beaumont (2008), to deal with the general problem of variable design weights, including the presence of extreme weights. This approach is adapted to handle the problem of stratum jumpers in stratified business surveys in Section 5.3. Unlike Winsorization of design weights, a compromise smoothed weight is obtained naturally when there is more than one $y$-variable. In Section 5.4, both weight smoothing and Winsorization are illustrated using the data of the CWES.

### *5.2. A general weight smoothing approach*

We consider again the problem of estimating the population total $T_y = \sum_{i\in U} y_i$. The probability sampling design that is used to select the sample $s$ is denoted by $p(s|\,\mathbf{Z})$, where $\mathbf{Z} = (\mathbf{z}_1, \ldots, \mathbf{z}_N)'$ and $\mathbf{z}_i$ is a vector of design variables (e.g., the size measure, region, or industry) for the $i$th population unit. Let us also use the notation $\mathbf{I} = (I_1, \ldots, I_N)'$ and $\mathbf{Y} = (y_1, \ldots, y_N)'$, where $I_i$ is the sample inclusion indicator of unit $i$; that is, $I_i = 1$ if the $i$th population unit is selected in the sample $s$ and $I_i = 0$, otherwise. It should be kept in mind that the design weight $d_i = 1/\pi_i$ is a function of $\mathbf{Z}$ only and should thus be written $d_i(\mathbf{Z})$. Nevertheless, we still denote it by $d_i$ for convenience.

The basic idea underlying weight smoothing consists of first viewing the design weights as random (see also Chapter 39 of this volume) and then using the model $\xi$:

$$E_\xi(d_i|\mathbf{I}, \mathbf{X}, \mathbf{Y}) = g_s(\mathbf{x}_i, y_i; \boldsymbol{\alpha}_s), \tag{18}$$

for $i \in s$, where $g_s(.;.)$ is some function that may be sample-dependent and $\boldsymbol{\alpha}_s$ is a vector of unknown model parameters to be estimated from sample data. Specific models for the design weights are given in Pfeffermann and Sverchkov (1999) and Beaumont (2008). Note that nothing precludes $y$ from being a vector so that the problem of finding a smoothed weight in multipurpose surveys boils down to considering more explanatory variables in model $\xi$. If $\tilde{d}_i = g_s(\mathbf{x}_i, y_i; \boldsymbol{\alpha}_s)$ were known, we would obtain the smoothed estimator $\tilde{T}_y^{\mathrm{SDB}}$ by replacing the design weights $d_i$ by the smoothed weight $\tilde{d}_i$ in a design-based estimator $\hat{T}_y^{\mathrm{DB}}$, such as the Horvitz–Thompson estimator or a calibration estimator. For instance, if $\hat{T}_y^{\mathrm{DB}}$ is the Horvitz–Thompson estimator, that is, $\hat{T}_y^{\mathrm{DB}} = \sum_{i\in s} d_i y_i$, then $\tilde{T}_y^{\mathrm{SDB}} = \sum_{i\in s} \tilde{d}_i y_i$. The role of model $\xi$ is to remove the unnecessary variability in the design weights. Beaumont (2008) showed that $\tilde{T}_y^{\mathrm{SDB}}$ is unbiased and never less efficient than the corresponding design-based estimator $\hat{T}_y^{\mathrm{DB}}$ under the model $\xi$ and the sampling design.

Since $\tilde{d}_i = g_s(\mathbf{x}_i, y_i; \boldsymbol{\alpha}_s)$ is unknown, we consider a model-consistent estimator $\hat{\boldsymbol{\alpha}}_s$ of $\boldsymbol{\alpha}_s$ and use this estimator to obtain $\hat{d}_i = g_s(\mathbf{x}_i, y_i; \hat{\boldsymbol{\alpha}}_s)$, for $i \in s$. This leads to the smoothed estimator $\hat{T}_y^{\mathrm{SDB}}$, which is obtained by using $\hat{d}_i$ instead of $d_i$ in the design-based estimator $\hat{T}_y^{\mathrm{DB}}$. Note that classical model selection and validation techniques can be used to determine an appropriate model and to estimate $\boldsymbol{\alpha}_s$ since we are interested in estimating the relationship between the design weight variable $d$ and both $\mathbf{x}$ and $y$ conditional on the realized sample and only for sample units. We expect that $\hat{T}_y^{\text{SDB}}$ keeps the good properties of $\tilde{T}_y^{\text{SDB}}$ in many practical applications provided that the underlying model (18) holds reasonably well. Indeed, Beaumont (2008) showed that if a linear model holds, the smoothed estimator $\hat{T}_y^{\text{SDB}}$ is unbiased and never less efficient than the Horvitz–Thompson estimator under the model $\xi$ and the sampling design.

Similarly to (11), a design-based MSE estimator of $\hat{T}_y^{\text{SDB}}$ is

$$\text{mse}\left(\hat{T}_y^{\text{SDB}}\right) = v\left(\hat{T}_y^{\text{SDB}}\right) + \max\left\{0, \left(\hat{T}_y^{\text{SDB}} - \hat{T}_y^{\text{DB}}\right)^2 - v\left(\hat{T}_y^{\text{SDB}} - \hat{T}_y^{\text{DB}}\right)\right\}. \tag{19}$$

Since $\hat{T}_y^{\text{SDB}}$ may have a complicated form, the bootstrap technique (e.g., Rao and Wu, 1988; Rao et al., 1992) is a natural candidate for obtaining estimators $v\left(\hat{T}_y^{\text{SDB}}\right)$ and $v\left(\hat{T}_y^{\text{SDB}} - \hat{T}_y^{\text{DB}}\right)$ of the design variances $\text{Var}\left(\hat{T}_y^{\text{SDB}}\right)$ and $\text{Var}\left(\hat{T}_y^{\text{SDB}} - \hat{T}_y^{\text{DB}}\right)$ respectively. One can also restrict the estimated MSE not to be greater than $v\left(\hat{T}_y^{\text{DB}}\right)$, the estimator of $\text{Var}\left(\hat{T}_y^{\text{DB}}\right)$, since we expect gains in efficiency if model $\xi$ holds. Such an MSE estimator performed well in the empirical study of Beaumont (2008).

### 5.3. *Weight smoothing to handle stratum jumpers*

Let us now consider the weight smoothing approach in the context of stratum jumpers in business surveys. We have already denoted by $\mathbf{Z}$, the matrix of design information available at the time of the design. Let us denote by $\mathbf{Z}_{\text{col}}$, the matrix of design information at the time of data collection, which is assumed to be measured essentially without errors for sample units. We may hypothesize that, once we know $\mathbf{Z}_{\text{col}}$, $\mathbf{X}$, and $\mathbf{I}$, the initial design matrix $\mathbf{Z}$ brings no extra information about $\mathbf{Y}$. In other words, $\mathbf{Y}$ is independent of $\mathbf{Z}$ after conditioning on $\mathbf{Z}_{\text{col}}$, $\mathbf{X}$, and $\mathbf{I}$; that is, $F\left(\mathbf{Y} \mid \mathbf{Z}, \mathbf{Z}_{\text{col}}, \mathbf{X}, \mathbf{I}\right) = F\left(\mathbf{Y} \mid \mathbf{Z}_{\text{col}}, \mathbf{X}, \mathbf{I}\right)$. This can also be rewritten as $F(\mathbf{Z} \mid \mathbf{Y}, \mathbf{Z}_{\text{col}}, \mathbf{X}, \mathbf{I}\big) = F\left(\mathbf{Z} \mid \mathbf{Z}_{\text{col}}, \mathbf{X}, \mathbf{I}\right)$. The latter implies that the design weights are independent of $\mathbf{Y}$ after conditioning on $\mathbf{Z}_{\text{col}}$, $\mathbf{X}$, and $\mathbf{I}$ and that a suitable model for the design weights would be

$$E_\xi\left(d_i|\mathbf{Z}_{\text{col}}, \mathbf{I}, \mathbf{X}, \mathbf{Y}\right) = g_s(\mathbf{z}_{\text{col},\,i}, \mathbf{x}_i; \boldsymbol{\alpha}_s\big), \tag{20}$$

where $\mathbf{z}_{\text{col},\,i}$ is the vector of design variables at the collection stage for unit $i$. These design variables are treated here like additional $y$-variables. The idea is to keep from the design weights the useful information contained in $\mathbf{z}_{\text{col}}$ and $\mathbf{x}$ (since it may have a strong relationship with $\mathbf{y}$) and remove their extra variability. Model (20) can then be used to construct a smoothed estimator, which should be more efficient than its corresponding design-based estimator.

In many business survey applications, model (20) often reduces to $E_\xi(d_i|\mathbf{Z}_{\text{col}}, \mathbf{I}, \mathbf{X}, \mathbf{Y}) = g_s(\mathbf{z}_{\text{col},\,i}; \boldsymbol{\alpha}_s)$. For instance, one has $\mathbf{X} = \mathbf{Z}_{\text{col}}$ in the example of Section 4.6. A simple approach to approximate the unknown function $g_s(\mathbf{z}_{\text{col},\,i}; \boldsymbol{\alpha}_s)$, is then to discretize $\mathbf{z}_{\text{col},\,i}$, for $i \in s$, into homogeneous categories called collection strata; this is considered in Section 5.4. Assuming that $g_s(\mathbf{z}_{\text{col},\,i}; \boldsymbol{\alpha}_s)$ is constant within each category,

Table 6
Smoothed weights in the example given in Table 5

| Collection Stratum | Design Stratum | Number of Units | Design Weight | Smoothed Weight | Smoothed Weight (with Constraint) |
|---|---|---|---|---|---|
| A | A | 9 | 1 | 4 | 1 × 1.0215 = 1.02 |
| A | B | 1 | 31 | 4 | 4 × 1.0215 = 4.09 |
| B | B | 40 | 31 | 31 | 31 × 1.0215 = 31.67 |
| Sum over the sample units | | 50 | 1280 | 1280 | 1253 × 1.0215 = 1280 |

we approximate the unknown model $\xi$ by a simple analysis-of-variance model. The smoothed weight $\hat{d}_i$ is simply obtained as the average of the design weights within the collection stratum containing unit $i$. The second-to-last column of Table 6 gives the smoothed weight when the above methodology is used in the example provided in Table 5.

For the stratum jumper, the smoothed weight is close to eight times smaller than the design weight. To compensate for this weight reduction, the smoothed weight of other units in collection stratum A became four times larger than the design weight. Since units with a small design weight may be associated to large $y$-values, it is perhaps preferable not to modify too much the weights of these units as they may become quite influential. One option, tested in Section 5.4, is to use the constraint that the smallest design weights are kept unchanged so that the nine units with a design weight of 1 in Table 6 would also be given a smoothed weight of 1. It may then be necessary to adjust all the resulting smoothed weights by a constant factor so that the overall sum of the final smoothed weights is still equal to the overall sum of the design weights. This leads to the last column of Table 6, where the constant factor is $1280/1253 = 1.0215$. This strategy is equivalent to a hybrid approach between Winsorization and weight smoothing, where the largest design weights are Winsorized within each analysis-of-variance cell (collection stratum). Under this scheme, the Winsorization cutoff is simple to compute as it is the average of the design weights within each collection stratum. Perhaps more sophisticated methods of finding the Winsorization cutoff could yield better results. This has yet to be investigated.

### *5.4. An illustration using the CWES*

To illustrate the benefits of the weight smoothing method when handling stratum jumpers, we use the 2003 CWES data, as described in Section 4.6. We have only one auxiliary variable with $x_i = z_{\mathrm{col},i}$ being the number of employees obtained at the collection stage for business $i$. Figure 7 shows the relationship between the design weights and $z_{\mathrm{col},i}$. The solid curve has been obtained using the procedure TPSPLINE of SAS. It is a nonparametric smoothing spline method based on penalized least squares estimation. We can see that there is a unit with a relatively large design weight of about 35 and a large value of $z_{\mathrm{col}}$. Standard Winsorization of the design weights may not reduce at all the weight of this unit, depending on the cutoff point. Weight smoothing will be more efficient by reducing the weight of this unit so that it has less influence on the estimates.

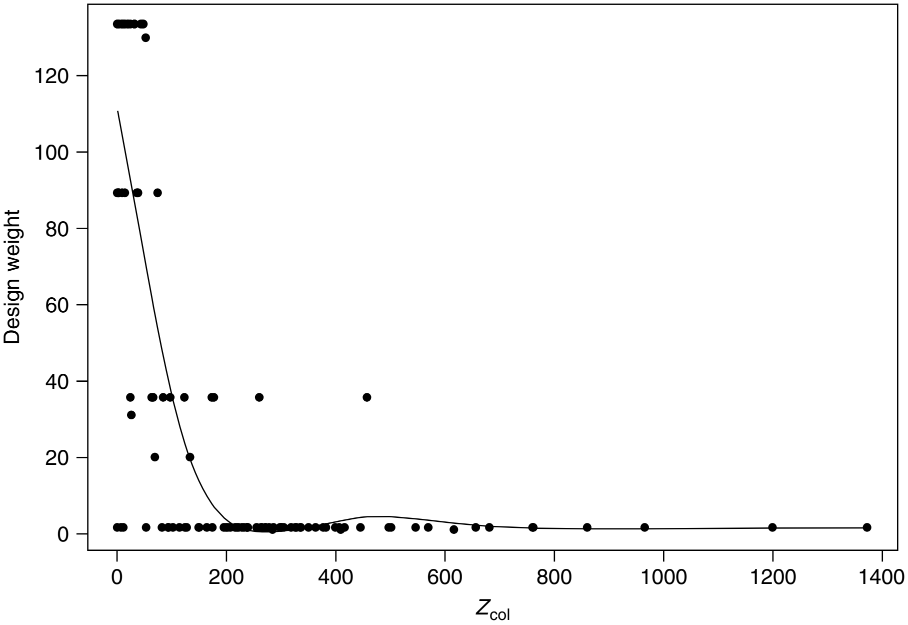


Fig. 7. Plot of the design weights versus $z_{\text{col},i}$.

To smooth the weights, we used a one-way analysis-of-variance model with five categories obtained by discretizing $z_{\text{col}}$ as proposed in Section 5.3. There are three categories for $z_{\text{col}} \leq 200$, as the slope of the smoothed spline curve is quite steep for small $z_{\text{col}}$ and two categories for $z_{\text{col}} > 200$. The resulting smoothed ratio estimator is denoted by SR-5. The analysis-of-variance residuals are plotted against $z_{\text{col}}$ and $Y_4$ in Figures 8 and 9. The smoothing splines in these figures do not show obvious trends in the residuals; thus this model for the design weights is satisfactory. Although they are not provided here, plots for the other $y$-variables were similar. Thus the assumption $F(\mathbf{Z}|\mathbf{Y}, \mathbf{Z}_{\text{col}}, \mathbf{X}, \mathbf{I}) = F(\mathbf{Z}|\mathbf{Z}_{\text{col}}, \mathbf{X}, \mathbf{I})$ made in Section 5.3 is reasonable.

In our empirical study, we have also considered a common mean model for the weights, which led to the smoothed ratio estimator SR-1. Under this model, the smoothed weight $\hat{d}_i$, for $i \in s$, is simply the overall average of the design weights. Versions of SR-1 and SR-5 that left the design weights less than 2 ($d_i < 2$) unchanged were also calculated, as suggested in the example shown in the last column of Table 6. All the smoothed weights were Winsorized to ensure that the final smoothed weight, say $\hat{d}_i^F$, lied in the range $0.1d_i \leq \hat{d}_i^F \leq 10d_i$. This step did not change the results in a significant way; it controlled the bias by preventing large weight adjustments $\hat{d}_i^F/d_i$, especially when the constraint on the smallest design weights was not used. This numerical investigation also features two estimators, WIN10 and WIN100, obtained by Winsorizing the largest design weights with cutoff points of 10 and 100, respectively. A summary of the empirical study is reported in Table 7. The quantities RD, Wald, and RE_Bootstrap are defined in Section 4.6 but with robust estimators replaced by estimators with smoothed or Winsorized weights.

Table 7
Comparison of smoothed and Winsorized estimators using CWES data

| Variable | Method | RD | Wald | RE_Bootstrap | RD | Wald | RE_Bootstrap |
|---|---|---|---|---|---|---|---|
| | | Without Constraint on the Smallest Design Weights | | | With Constraint on the Smallest Design Weights | | |
| $Y_1$ | SR-1 | 8.88 | 0.59 | 434.9 | 11.96 | 3.68 | 56.6 |
| | SR-5 | −3.77 | 0.10 | 323.2 | −10.31 | 0.81 | 238.4 |
| | WIN10 | 10.08 | 1.90 | 143.8 | 10.08 | 1.90 | 143.8 |
| | WIN100 | 2.34 | 1.99 | 92.3 | 2.34 | 1.99 | 92.3 |
| $Y_2$ | SR-1 | 10.10 | 0.62 | 448.9 | 14.19 | 4.05 | 53.6 |
| | SR-5 | −5.49 | 0.16 | 418.8 | −12.98 | 0.98 | 318.7 |
| | WIN10 | 11.70 | 2.14 | 135.2 | 11.70 | 2.14 | 135.2 |
| | WIN100 | 2.53 | 1.78 | 92.8 | 2.53 | 1.78 | 92.8 |
| $Y_3$ | SR-1 | 18.45 | 10.82 | 11.3 | 10.12 | 20.02 | 26.5 |
| | SR-5 | 2.22 | 0.32 | 155.4 | −1.66 | 0.13 | 136.4 |
| | WIN10 | 15.20 | 16.07 | 15.6 | 15.20 | 16.07 | 15.6 |
| | WIN100 | 1.72 | 7.03 | 88.6 | 1.72 | 7.03 | 88.6 |
| $Y_4$ | SR-1 | 137.32 | 73.97 | 0.5 | 29.06 | 26.68 | 10.8 |
| | SR-5 | 39.66 | 6.91 | 6.3 | 12.36 | 3.56 | 40.5 |
| | WIN10 | 95.05 | 61.67 | 1.1 | 95.05 | 61.67 | 1.1 |
| | WIN100 | 7.30 | 34.28 | 64.2 | 7.30 | 34.28 | 64.2 |
| $Y_5$ | SR-1 | 47.46 | 8.65 | 13.7 | 27.00 | 15.33 | 28.4 |
| | SR-5 | −13.01 | 1.26 | 211.0 | −12.22 | 0.92 | 215.3 |
| | WIN10 | 39.47 | 13.17 | 18.2 | 39.47 | 13.17 | 18.2 |
| | WIN100 | 5.32 | 8.60 | 88.9 | 5.32 | 8.60 | 88.9 |

From Table 7, we can make the following remarks:

- For all variables but $Y_4$, the SR-5 estimator was not significantly biased, according to the Wald statistic, and it was more efficient than the ratio estimator. Also, the constraint on the smallest design weights led generally to a small loss of efficiency.
- For variable $Y_4$, the SR-5 estimator was significantly biased although less biased than its competitors according to the Wald statistic. This resulted in an inefficient estimator. The use of the constraint on the smallest design weights substantially reduced the bias and improved the efficiency.
- For variables $Y_1$ and $Y_2$, the SR-1 estimator was the most efficient when no constraint on the smallest design weights were used, although only marginally more efficient than estimator SR-5. However, imposing this constraint resulted in a significant loss of efficiency and an increase in bias for these two variables. For the other three variables, the SR-1 estimator had a very large bias which made the estimator inefficient.
- Both Winsorized estimators did not perform well. The WIN10 estimator was sometimes significantly biased, whereas the WIN100 never led to gains in efficiency. We tried several other Winsorization cutoffs but were not able to find any satisfactory

compromise. Note also that the constraint on the smallest design weights had no effect on the Winsorized estimators.

- Overall, the SR-5 estimator is the best. Its relative efficiency is generally larger than that of the optimal robust calibration estimator of Table 4, except for variable $Y_4$. Also, the constraint on the smallest design weights seems to offer protection against bias at the expense of a slight loss of efficiency when the bias of the smoothed estimator is not significant.

Figures 8 and 9 suggest that the analysis-of-variance model used in this example is adequate. Using an argument similar to Beaumont (2008), the resulting smoothed estimator SR-5 should be asymptotically unbiased and more efficient than the ratio estimator, under the model and the sampling design, provided that this linear model holds. This is in agreement with results of Table 7, except for variable $Y_4$. The bias for variable $Y_4$ may thus be explained either by a slight model misspecification that is difficult to detect by a graphical analysis or by an error of the Wald test since the Wald statistic is not that extreme. From a single sample, it is difficult to determine the exact cause of this bias. It is worth mentioning again that the constraint on the smallest design weights seems to bring some robustness against model misspecification and potential bias. Another alternative could have been to include $Y_4$ in the model, in addition to $z_{\text{col}}$, to reduce the impact of the possible model misspecification.

As a final comment on this example, note that the SR-1 estimator is equivalent to a model-based ratio estimator when no constraint on the design weights is used. Such a model-based estimator ignores completely the design weights and should work well if

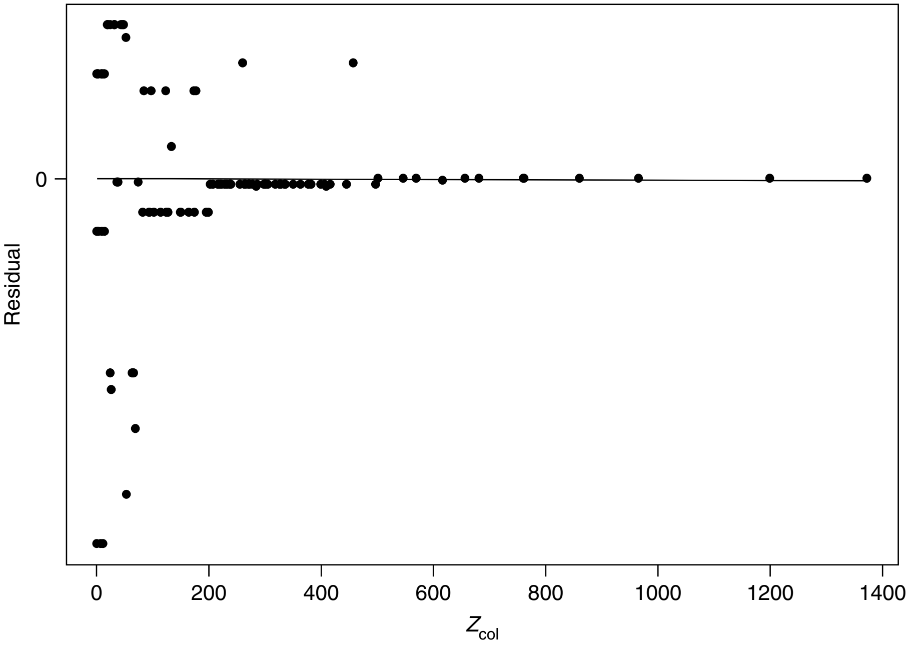


Fig. 8. Plot of the analysis-of-variance model residuals versus $z_{\text{col},i}$.

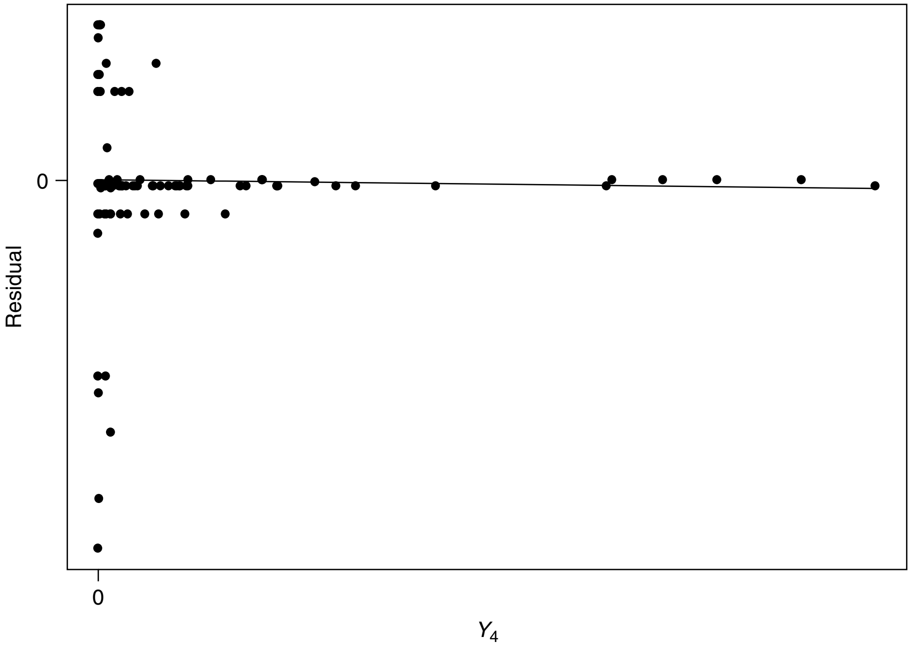


Fig. 9. Plot of the analysis-of-variance model residuals versus $Y_{4,i}$.

its underlying model *m* explains satisfactorily the relationship between the *y*-variables and *x*. Apparently, this might have been the case for variables $Y_1$ and $Y_2$. However, Table 7 also indicates that it may be risky in general to blindly use this estimator unless one is confident that model *m* holds reasonably well.

## 6. Practical issues and future work

Chambers et al. (2000) discussed three practical issues related to Winsorization but that are also applicable to most robust estimation procedures including M-estimation. The first issue is illustrated in the numerical example of Section 2.5, which shows that the bias-variance trade-off for controlling outliers varies with the level of aggregation. On the one hand, we may not reduce the MSE enough in small domains if *y*-values of influential units are modified (e.g., Winsorized) at high levels of aggregation. On the other hand, we may have good performance in small domains if *y*-values of influential units are modified at lower levels of aggregation. However, such a strategy may end up being quite biased at higher levels of aggregations and be far from optimal. The methodology in Rivest and Hidiroglou (2004) could be useful to address this issue so that estimates at higher levels of aggregation remain of good quality.

The second issue is about robust estimation of derived variables. More specifically, if different *y*-variables are Winsorized separately then linear (or nonlinear) relationships that hold among them are likely to be destroyed by Winsorization. For instance, if the sum of variables at unit level must add up to some total then a separate Winsorization of each variable, including the total, is likely to destroy this linear relationship. The same problem is encountered when imputing separately variables with missing values for which some relationships must hold. In practice, some form of pro-rating is usually performed. Chambers et al. (2000) also provided a solution. The last issue that they consider is the treatment of nonrepresentative outliers. It is often assumed that such outliers are not present in the data due to thorough editing. This seems to be a strong assumption in practice. How to deal with such outliers at the estimation stage is not trivial.

There remains research to be done to handle the above practical issues. Also, the choice of the tuning constant involved in most robust methods is important and should not be taken lightly. More research is certainly needed on this topic. Robust estimation of nonlinear population parameters and of changes have not been fully investigated although there has recently been some useful work on these topics. For instance, the method of Zaslavsky et al. (2001) is a good starting point for nonlinear parameters as well as the Winsorized estimators of changes developed by Lewis (2007). Finally, applications of Winsorization or M-estimation to real survey data is important to better understand the properties of robust estimators. A few examples of applications to real survey data are Tambay (1988) and Matthews and Bérard (2002) for Winsorization and Gershunskaya and Huff (2004) and Mulry and Feldpausch (2007) for M-estimation. Nevertheless, more empirical investigations would certainly be useful, especially in the context of M-estimation.

Regarding the weight smoothing approach described in Section 5, it is important to point out that finding an appropriate model is a key aspect of the method. To obtain some robustness against model failures, we partitioned the sample into collection strata in a more or less ad hoc way in Section 5.4. Research into the issue of determining adequate collection strata boundaries could be useful. An alternative to determine collection strata, which remains to be investigated, would be to use nonparametric methods of estimating the smoothed weights. Finally, if it is believed that weight smoothing does not yield a sufficient increase in efficiency then nothing precludes, in principle, to combine weight smoothing with an outlier-robust method, such as Winsorization or M-estimation. This also remains to be investigated.

*Sample Surveys: Design, Methods and Applications, Vol. 29A*
ISSN: 0169-7161

DOI: 10.1016/S0169-7161(08)00012-6

# 12

# Measurement Errors in Sample Surveys

*Paul Biemer*

## 1. Introduction

Measurement errors are errors in the survey observations that may be due to interviewers, respondents, data processors, and other survey personnel. Often, the causes of measurement errors are poor questions or questionnaire design, inadequate personal training or supervision, and insufficient quality control. Measurement errors are often hidden in the data and are only revealed when the measurement process is repeated or responses are compared to a *gold standard* (i.e., error-free measurements). If repeated measurements are collected by the same measurement process, systematic errors may remain hidden. Fortunately, many analytical techniques can be used to account for measurement errors in data analysis and inference, such as structural equation modeling (Bollen, 1989), instrumental variables (Fuller, 1987), and errors-in-variables modeling (Carroll et al., 2006). In general, these techniques consider measurement error components more as nuisance parameters whose effects are to be neutralized in the analysis to achieve greater inferential validity.

In this chapter, measurement error components are considered the primary parameters of interest in an analysis, not as by-products of the analysis. Knowledge of the magnitudes of measurement bias and variance can serve multiple purposes. First, this knowledge can be used to improve data-collection methodology. As an example, a survey methodologist may wish to compare the accuracy of health data collected by telephone with data collected by face-to-face interviewing. Typically, such comparison will entail a mode comparison study based on a split-ballot design, where, say, half the sample is assigned to one mode and half to the other mode. While this comparison may be sufficient for determining whether two modes will produce different results, it is usually not sufficient for determining the better mode of response accuracy. For this purpose, the measurement error components must be estimated and compared. Other error components could be considered as well, for example, nonresponse bias.

Second, the assessment of measurement error can also lead to improved survey questionnaires. Poor reliability of a survey question may indicate a problem with the questionnaire wording. Confusing references, undefined technical terms, vague quantifiers, and so on can lead to respondent confusion, comprehension error and, ultimately, measurement error in responses. Correcting these problems usually begins with a study to identify which questions are subject to measurement error as well as the magnitude of the errors. Only then the sources of errors can be discovered and traced to their root causes so that measurement error can be reduced. Identifying root causes might involve cognitive laboratory methods or special field studies (Biemer and Lyberg, 2003, Chapter 8).

Third, information on the measurement error properties of survey variables can be quite useful to data users and analysts who need to understand limitations of the data. For example, it is easy to show that the *coefficient of determination* (or $\mathrm{R}^2$) for simple regression can never exceed the reliability of the dependent variable. In addition, the estimated regression coefficients are attenuated towards 0 by a factor equal to the reliability of the corresponding dependent variable (Fuller, 1987). Thus, knowledge of the reliability of the analysis variables will help explain the lack of fit of a model or the statistical insignificance of variables thought to be highly explanatory of the dependent variable. Estimates of measurement error can even be used in some cases to correct the analysis for measurement error biases, as previously noted.

This chapter presents five modeling approaches that are appropriate for the study of measurement error: three of which focus primarily on classification errors. Section 2 discusses the model first espoused by Hansen et al. (1964), which can be applied to any type of variable. The essential formulas for this approach follow almost immediately from cluster sampling theory when individuals in a survey are viewed as primary sampling units with secondary "units" corresponding to their potential responses to a survey question. Section 3 extends this model to a classification probability model. In this context, the bias and variance parameters of Section 2 are shown to be complex functions of false positive and false negative probabilities associated with each population unit. Section 3 also introduces the essential concepts underlying latent class analysis (LCA) of measurement error. Section 4 provides greater detail about LCA and extends those concepts to panel data by introducing Markov Latent Class Analysis (MLCA). Section 5 provides a discussion of some common approaches for the assessment of measurement error in continuous data. Finally, the chapter closes with a brief discussion of the main ideas and examines the future of measurement error research.

## 2. Modeling survey measurement error

### 2.1. The general response model

This section introduces a simple model for measurement error originally proposed by Hansen et al. (1964) and revisited in Biemer (2004b), which we refer to as the Hansen-Hurwitz-Pritzker (HHP) model. The focus in this section is inferences about the population mean, $\bar{Y}$, of some characteristic of interest, $\mathcal{Y}$. To fix the ideas, assume a simple random sample (SRS) of individuals is selected, although extension to complex sampling designs is straightforward.

Suppose $\mathcal{Y}$ is measured with error by the survey process; that is, assume that a distribution of responses, $h_i(\mathcal{Y})$, is associated with each individual, $i$, in the population. Each observation on $\mathcal{Y}$ for the $i$th unit is analogous to a single draw from this distribution. Biemer (2004b) showed that, under these assumptions, the measurement process is analogous to a two-stage sampling process, where in the first stage, an individual is selected at random from the finite population of individuals, and in the second stage, $\mathscr{Y}$ is observed for unit $i$ by randomly selecting a value from the distribution $h_i(\mathscr{Y})$. Biemer showed that the usual formulas for two-stage sampling can be directly applied to obtain the population parameter estimation formulas with measurement errors.

To examine these assumptions further, imagine a hypothetical survey process that can be repeated many times under identical survey conditions; that is, under conditions where the same response distribution, $h_i(\mathscr{Y})$, applies for each measurement on $i$. For example, (a) a sampled individual is asked a question; (c) his/her response is recorded; (d) amnesia is then induced; (a)–(c) are repeated some number, say $m$, of times generating $m$ draws from $h_i(\mathscr{Y})$. In a typical survey, only one realization of $\mathscr{Y}$ is obtained for each respondent. Measurement error evaluation studies may obtain two or more realizations of $\mathscr{Y}$. As an example, a test–retest reinterview may be conducted for a subsample of respondents where a subsample of respondents are revisited after the original survey and asked some of the same questions again. Here, the second measurement is obtained solely to estimate the measurement variance. In our discussion, we will consider a sequence of $m \geq 1$ repeated measurements on the $i$th unit denoted by $y_{ij}$, $j = 1, \ldots, m$, all drawn independently from the same distribution $h_i(\mathscr{Y})$.

Let $\overline{Y}_i$ and $\sigma_i^2$ denote the mean and variance of $h_i(\mathscr{Y})$. In the psychometric literature (see, e.g., Nunnally and Berstein, 1994), $\overline{Y}_i$ is called the *true score* of $i$th individual and $\sigma_i^2$ is the *error* (or *response*) *variance* component. The mean of the population is $\overline{Y}$, that is, the population average true score. Later, we will introduce the concept of a *true value*, denoted by $\mu_i$ that is distinct from the true score $\overline{Y}_i$. The present development does not require the acknowledgement of a true value.

An unbiased estimator of $\overline{Y}$ and its variance can be obtained by applying the usual textbook formulas for two-stage sample with SRS at each stage (see, e.g., Cochran, 1977, Chapter 10). We initially consider the general case where all $m$ realizations of $\mathscr{Y}$ are used to estimate $\overline{Y}$, and then, as a special case, we consider the common situation where the first realization (representing the main survey response) is used to estimate $\overline{Y}$, and the repeated measurements are used to evaluate this estimator of $\overline{Y}$. For a SRS of size $n$ from the population of size $N$, let $\overline{y}_i = \sum_j y_{ij}/m$, the mean response for the $i$th individual (sometimes called the individual's *observed score*). An unbiased estimator of $\overline{Y}$ based on all $m$ measurements is $\overline{\overline{y}} = \sum_{i=1}^{n} \overline{y}_i/n$ with variance

$$\mathrm{Var}\left(\overline{\overline{y}}\right) = (1 - f)\frac{S_1^2}{n} + \frac{S_2^2}{nm}, \tag{1}$$

where $f = n/N$, $S_1^2 = \sum (\overline{Y}_i - \overline{Y})^2/(N - 1)$, and $S_2^2 = \sum \sigma_i^2/N$.

Under the measurement error model, the variance of $\overline{\overline{y}}$ is a linear combination of two variance components which HHP referred to as *sampling variance* ($(N-1)N^{-1}S_1^2$, denoted by SV) and *simple response variance* ($S_2^2$, denoted by SRV). Note that the contribution to variance of SV decreases as the sample size increases, and the contribution to SRV decreases as the product of sample size and number of measurements ($nm$) increases. As HHP note, SRV reflects the trial-to-trial variation in responses averaged over all individuals in the population. The larger the SRV, the less *reliable* are the responses obtained by the survey process, since a repetition of the survey under identical conditions would yield very different responses even if the same sample were used. As we shall see, SRV is a key determinant of the *reliability* of a measuring process. When there is no trial-to-trial variation in responses, the SRV is 0. In that case, the variance simplifies to $(1-f)S_1^2/n$, where $S_1^2$ is the variance of $Y_i$. This is the classical formula for the variance of the sample mean in SRS without measurement error. For a typical survey with $m = 1$, the variance of the SRS simple expansion mean of the observations can be obtained by setting $m = 1$ in (1). Note that the contribution to variance of SRV is not 0 in this case; but, as we shall see later, with only one observation per unit, SRV cannot be estimated.

### 2.2. *Estimation of response variance and reliability*

The *reliability*, a survey variable, is key concept in the study of measurement error. It is defined in terms of a single observation rather than an estimator. Consider the variance of a single observation for a randomly selected individual, and let this observation be denoted by $y_{i1}$. From (1), setting $n = m = 1$,

$$\mathrm{Var}(y_{i1}) = \mathrm{SV} + \mathrm{SRV} \tag{2}$$

using the HHP's notation. This gives rise to a useful measure of response quality referred to as the *inconsistency ratio* defined as

$$I = \frac{\mathrm{SRV}}{\mathrm{SV} + \mathrm{SRV}} = \frac{\mathrm{SRV}}{\mathrm{Var}(y_{i1})}. \tag{3}$$

The inconsistency ratio may be interpreted as the proportion of the variance of a single observation that is attributable to measurement error. The complement of $I$, that is, $R = 1 - I$ is called the *reliability ratio*. It is the proportion of total variance that is true score variance (see, e.g., Fuller, 1987).

Note that both $I$ and $R$ are bounded by 0 and 1. The closer the $R$ is to 1, the smaller is SRV, and the observations are said to be more *reliable*. When $R$ is below 0.5 (i.e., $I$ above 0.5), reliability is considered to be poor because more than 50% of the variation in the observations is the result of measurement error (noise). When $(1 - f_1)$ can be ignored and $m = 1$ (i.e., no repeated measurements), $\mathrm{Var}(\overline{\overline{y}})$ is inversely proportional to $R$; that is, $\mathrm{Var}(\overline{\overline{y}}) = \mathrm{SV}/nR$, which follows directly from (1). The product $nR$ is sometimes referred to as the *effective* sample size.

Further application of the classical two-stage sampling formulas yields an unbiased estimator of $\mathrm{Var}(\overline{\overline{y}})$ given by

$$v(\overline{\overline{y}}) = \frac{1-f}{n}s_1^2 + \frac{f}{nm}s_2^2, \tag{4}$$

where $s_1^2 = (n-1)^{-1}\sum(\overline{y}_i - \overline{\overline{y}})^2$, $s_2^2 = [n(m-1)]^{-1}\sum_i\sum_j(y_{ij} - \overline{y}_i)^2$. Note that (4) simplifies to $v(\overline{\overline{y}}) = n^{-1}s_1^2$ when $f << 1$ (Cochran, 1977, Chapter 10).

It can be shown that $\mathrm{E}(s_2^2) = \mathrm{SRV}$ and $\mathrm{E}(s_1^2) = \mathrm{SV} + \mathrm{SRV}/m$. As mentioned previously, quite often the methodologist is interested in the quality of one particular measurement of $\mathscr{Y}$ and repeated measurements are obtained toward that end. Thus, rather than using $\overline{\overline{y}}$, the estimator of $\overline{Y}$ is based upon one set of measurements, say $y_{i1}(i = 1, \ldots, n)$. Denote this estimator by $\overline{y}_{(1)} = \sum_{i=1}^{n} y_{i1}/n$, which has variance given by (1) with $m = 1$. Two estimators of $\mathrm{Var}(\overline{y}_{(1)})$ are possible. One can be obtained from

(1) by replacing SRV and SV by their estimators $s_2^2$ and $s_1^2 - s_2^2/m$, respectively, where $s_1^2$ is based upon all $m$ available measurements. A second estimator is also possible that is based only upon a single set of observations. Let $\widehat{\text{SV}}_{(1)}$ denote the estimator $S_1^2$ based on $y_{i1} (i = 1, \ldots, n)$. Then, $\widehat{\text{SV}}_{(1)}/n$ is an unbiased estimator of $\text{Var}(\bar{y}_{(1)})$ when $f$ can be ignored. This also shows that the usual SRS estimator of $\text{Var}(\bar{y})$ for one set of measurements is also an unbiased estimator of the variance for negligible $f$.

If $m \geq 2$, an estimator of reliability can be obtained by noting that $s_1^2 + \left(\frac{m-1}{m}\right)s_2^2$ is an estimator of SV + SRV. Thus, an estimator of $I$ is obtained by replacing SRV and SV + SRV in (3) by their estimators to obtain

$$\hat{I} = \frac{s_2^2}{s_1^2 + \left(\frac{m-1}{m}\right) s_2^2}. \tag{5}$$

This estimator of $I$ is referred to in the literature as the *index of inconsistency*. It follows that an estimator of $R$ is $1 - \hat{I}$. It can be shown that (5) is a consistent estimator of $I$ for any number ($m \geq 2$) of repeated measures.

### 2.3. *Special case: Dichotomous variables*

Many variables collected in surveys are either inherently categorical or continuous variables that have been discretized. Therefore, much of our study of measurement errors in this chapter will focus on *classification errors*. Classification errors are errors in a categorical variable that cause respondents to be classified in to different categories of the variable over repeated measurements. To simplify the discussion, we initially assume $\mathscr{Y}$ is dichotomous with values 0 and 1. The true score for unit $i$ is $h_i(1) = \text{P}(Y_{ij} = 1|i) = P_i$; that is, $P_i$ is the probability that individual $i$ is classified as a positive by the survey process. The formulas for SV and SRV in the previous section can now be rewritten as

$$\text{SV} = \sum_{i=1}^{N} \frac{(P_i - P)^2}{N - 1} \text{ and } \text{SRV} = \sum_{i=1}^{N} \frac{P_i Q_i}{N} \tag{6}$$

where $Q_i = 1 - P_i$ and $P = N^{-1} \sum P_i$ is the population proportion. Further, the estimators $s_1^2$ and $s_2^2$ can be rewritten as $s_1^2 = (n-1)^{-1} \sum (p_i - p)^2$ and $s_2^2 = [n(m-1)]^{-1} m \sum p_i q_i$, where $p_i$ is the proportion of the $m$ observations on the $i$th unit that are positive, $p = \sum_{i=}^{n} p_i/n$ and $q_i = 1 - p_i$.

For $m = 2$ (e.g., a test–retest reinterview study), the data can be summarized by the *interview–reinterview* (or *crossover*) table shown in Table 1, where $p_{11}$ denotes the proportion of sample members classified as 1 on both occasions, $p_{01}$ is the proportion classified as 0 in the interview and 1 in the reinterview, $p_{01}$ is the proportion classified

Table 1
Interview–reinterview table

| | Reinterview ($B$) | |
|---|---|---|
| Interview ($A$) | 1 | 0 |
| 1 | $p_{11}$ | $p_{10}$ |
| 0 | $p_{01}$ | $p_{00}$ |

as 1 in the interview and 0 in the reinterview, $p_{00}$ is the proportion classified as 0 by both the interview and the reinterview, and $p_{11} + p_{01} + p_{10} + p_{00} = 1$. The notation frequently used in the psychometric literature denotes the original observation ($y_{1i}$) by $A$ and the reinterview classification ($y_{2i}$) by $B$ (note that the subscript $i$ denoting the unit is implicit).

It is easy to show that $s_2^2$ can be further simplified as

$$s_2^2 = \frac{p_{01} + p_{10}}{2} = \frac{g}{2}, \text{ say,} \tag{7}$$

where $g$, referred to as the *gross difference rate*, is the proportion of off-diagonal units in the table or the *disagreement rate*; that is, the proportion of the sample that is classified inconsistently by the two interviews.

As shown for the general case, when $m \geq 2$, there is more than one estimator of SV + SRV, which will lead to different estimators of $I$. Both $p_A q_A$ and $p_B q_B$ estimate SV + SRV under the assumptions of our model. In test–retest reinterview surveys in which only a subsample of respondents are reinterviewed, $p_A q_A$ may be more precise if it is taken from the main survey and is therefore preferred. An estimator of the inconsistency ratio using this denominator is $\hat{I}' = g/(2p_A q_B)$. However, for estimation based only on the cases that have been reinterviewed, an estimator that has somewhat better precision incorporates information from both the interview and the reinterview to estimate SV + SRV. It is given by

$$\hat{I} = \frac{g}{p_A q_B + p_B q_A}, \tag{8}$$

where $p_A$ and $p_B$ are the interview and reinterview proportions, respectively (U.S. Census Bureau, 1985). Hess et al. (1999) provide the interesting result that (8) is identical to $1 - \kappa$ where $\kappa$ is Cohen's kappa measure of reliability (Cohen, 1960) given by $\kappa = (P_0 - P_e)/(1 - P_e)$, where $P_0$ is the agreement rate between the interview and reinterview classifications (i.e., $1 - g$), and $P_e$ is an estimate of the expected agreement by chance alone, that is, $P_e = p_A p_B + q_A q_B$. Kappa may be interpreted as a *chance-corrected* agreement rate since it is the agreement rate adjusted for the probability of chance agreement (i.e., $P_0 - P_e$) divided by the maximum value of this quantity $(1 - P_e)$.

Thus, $\kappa$ has two very different interpretations in the literature: an estimator of $R$ and the chance-corrected agreement rate. Guggenmoos–Holzmann (1996) and Guggenmoos–Holzmann and Vonk (1998) discuss yet a third interpretation of $\hat{R}$ under the so-called *agreement model*. Under their model, reliability is the proportion of units in the population that can be classified consistently by raters (so-called "conclusive" units). For a comparison of this and other methods for computing reliability, see Biemer (2004b).

### *2.3.1. Limitations of parallel measurements*

Recall that, in order for $s_2^2$ to be unbiased for $S_2^2$, the two replicate measures must be parallel, that is, equivalent to a SRS of size $m = 2$ from each individual's response distribution. As Hansen et al. (1964) discuss, these assumptions are seldom satisfied in practice. For example, if the second measurement is provided by a test–retest reinterview, the general survey conditions that existed during the interview likely have changed by the time of the reinterview; thus, the equal error distribution assumptions may not hold. In addition, respondents may have been conditioned by the first interview and their reinterview responses may reflect this conditioning. For example, after the interview, respondents may have obtained additional information on the survey topics that could influence their responses in the reinterviews. The respondent true values may also have changed since the interview. This can be addressed to some extent by modifying the reinterview questions to refer to the same time period referenced in the interview. The reinterview questionnaire is often much shorter than the interview questionnaire and is usually conducted using less expensive methodologies. These design changes threaten the validity of the reliability estimates since they conspire to violate the assumption of identically distributed interview and reinterview errors.

The assumption of conditional (or *local*) independence of response errors is also unlikely to hold in many practical situations. Between-trial correlations can be induced if respondents tend to simply recall their interview responses and repeat them rather than providing a response without referencing the interview. It may be possible to schedule the second interview to allow sufficient time for the respondent to forget their interview responses (see, e.g., Bailar, 1968). However, response errors may still be correlated if respondents tend to misinterpret the survey questions in the same way at both occasions or otherwise use the same process to generate an erroneous response. The risk of correlated errors is particularly great for embedded replicate measurements; i.e., measures that are obtained within the same questionnaire and interview.

Suppose a reinterview study is conducted to estimate the reliability for some characteristic, and let $g$ denote the gross difference rate from the study. It can be shown that, in general,

$$\mathrm{E}(g) = \mathrm{SRV}_A + \mathrm{SRV}_B - \rho_{AB}\sqrt{\mathrm{SRV}_A\mathrm{SRV}_B} + D_{AB}^2 \qquad (9)$$

(U.S. Census Bureau, 1985), where $\mathrm{SRV}_A$ and $\mathrm{SRV}_B$ denote simple response variance for the interview and reinterview, respectively, $\rho_{AB}$ is the between-trial error correlation, and $D_{AB}$ denotes the expected difference between the interview and reinterview responses. Under the parallel assumption, $\rho_{AB} = D_{AB}^2 = 0$, $\mathrm{SRV}_A = \mathrm{SRV}_B$, and thus, $g$ is unbiased for $2\mathrm{SRV}_A$. The failure of any of these three conditions to hold will result in $\mathrm{E}(g) \neq 2\mathrm{SRV}_A$ and biased estimates of reliability. Note that even if $\rho_{AB} = D_{AB}^2 = 0$, $\mathrm{SRV}_A$ and $\mathrm{SRV}_B$ may still differ and thus $\hat{R}$ will be biased.

In that case, $g$ is an estimator of $(\mathrm{SRV}_A + \mathrm{SRV}_B)/2$. If $\mathrm{SRV}_A < \mathrm{SRV}_B$, then $\hat{R}$ will likely overestimate $R_A$. In other words, greater reliability of the reinterview process will tend to make the reliability estimate for the original interview look better than it is. Similarly, if $\mathrm{SRV}_A > \mathrm{SRV}_B$, $\hat{R}$ will likely underestimate $R_A$ (U.S. Census Bureau, 1985); Suppose that SRV is the same for both trials, $D_{AB} = 0$ but $\rho_{AB} > 0$, that is, the errors in the two measurements are positively correlated. In this case, it can be shown from (9) that $\mathrm{E}(g) = 2\mathrm{SRV}(1 - \rho_{AB})$, that is, SRV will be underestimated and reliability will be overestimated. In other words, positive correlations between the errors in the two measurements will make the measurements appear to be more consistent than they are, thus negatively biasing the index of inconsistency. Some violations of the parallel assumptions create negative bias, whereas others may create positive bias in the estimates of $R_A$. In general, the bias in $\hat{R}$ is unpredictable. Examples of this unpredictability are provided in the following illustration.

Table 2
Past-year marijuana use for three measures

| $A$ | $B$ | $C$ | COUNT |
|---|---|---|---|
| 1 | 1 | 1 | 1,158 |
| 1 | 1 | 0 | 2 |
| 1 | 0 | 1 | 2 |
| 1 | 0 | 0 | 82 |
| 0 | 1 | 1 | 7 |
| 0 | 1 | 0 | 313 |
| 0 | 0 | 1 | 308 |
| 0 | 0 | 0 | 15,876 |

### *2.3.2. Example 1: Reliability of questions on marijuana use*

To illustrate the concepts, we use data from a large, national survey on drug use. Data for three dichotomous measures of past-year marijuana use are shown in Table 2. The measures are labeled $A$, $B$, and $C$, and for each, "1" denotes use and "0" denotes no use. Though the sample was drawn by a complex, multistage, unequal probability design, SRS will be assumed for the purposes of the illustration. Cell counts have been weighted and rescaled (see Section 5.2) to the overall sample size, $n = 17{,}748$.

The assumption of parallel measures can be tested by comparing three proportions, viz., $\mathrm{P}(A = 1)$, $\mathrm{P}(B = 1)$, and $\mathrm{P}(C = 1)$. If a test of equality is rejected, the assumption of parallel measures is not supported by the data; however, failure to reject does not suggest that $A$, $B$, and $C$ are parallel. The estimate of $\mathrm{P}(A = 1)$ is $p_A = (1158 + 2 + 2 + 82)/17748 = 0.070$; $\mathrm{P}(B = 1)$ is $p_B = 0.083$, and $\mathrm{P}(C = 1)$ is $p_C = 0.083$. The test is rejected indicating that the parallel assumption is not supported. Thus, we expect that our estimates of reliability will be somewhat biased.

An estimator of marijuana use prevalence that may be considered is the combined estimator of prevalence $\bar{\bar{y}}$. For three measurements, $p_i$ will be 0, 1/3, 2/3, or 1. Table 2 shows that the frequency of each of these values is 15876, 703 (=82 + 313 + 308), 11 (=2 + 2 + 7), and 1158. Thus, $p = (1/3 \times 703 + 2/3 \times 11 + 1 \times 1158)/17748 = 0.079$. The standard error of this estimate may be computed as $s_1^2/n$ using the data in Table 2. Using the estimates of the parameters in (6), the possible values for $(p_i - p)^2$ are $(0 - 0.079)^2$, $(1/3 - 0.079)^2$, $(2/30.079)^2$, and $(1 - 0.079)^2$, which occur with frequencies 15876, 703, 11, and 1158, respectively. Multiplying the values of $(p_i - p)^2$ by their corresponding sample frequencies and dividing by $(17748 - 1)$, which yields $\mathrm{v}(p) = 3.60 \times 10^{-6}$.

Now consider estimating the reliability of the measurements under the parallel assumption. Applying $s_2^2$, an estimate of SRV can be obtained from the data in Table 2. Note that the nonzero values of $p_i q_i$ are (1/3)(2/3) and (2/3)(1/3), both of which yield 0.222 occurring with frequency 714 (=703 + 11). Thus, $s_2^2 = (17748)^{-1}(3/2)(0.222)(714) = 0.0134$. The index of inconsistency from (5) is $0.0134/[(3.60 \times 10^{-6}) \times (17748)]$, which yields $\hat{I} = 0.21$. The reliability of the measures is $\hat{R} = 1 - 0.21 = 0.79$.

The U.S. Census Bureau (1985) suggests the following rule of thumb for interpreting the magnitude of $\hat{I}$: $0 \le \hat{I} \le 0.2$ is good, $0.2 < \hat{I} \le 0.5$ is moderate, and $\hat{I} > 0.5$ is poor. This rule is merely offered as a guideline since whether a particular level of reliability is too low depends upon the purpose to which the data will be put. Nevertheless, by this rule, past-year marijuana use can be said to exhibit moderately good reliability.

Using the first two columns of Table 2, we can demonstrate the calculations for the case where there are only two measures. In this situation, $p_{11} = 0.065$, $p_{01} = 0.0047$, $p_{10} = 0.018$, and $p_{00} = 0.91$. Thus, $g = 0.0227$ and, as calculated previously, $p_A = 0.070$ and $p_B = 0.083$. It follows from (8) that $\hat{I} = 0.0227/(0.065 + 0.078) = 0.16$. To calculate $\kappa$, we first compute $P_0 = p_{11} + p_{00} = 0.977$ and then $P_e = p_A p_B + q_A q_B = 0.86$. Hence, $\kappa = (0.977 - 0.86)/(1 - 0.86) = 0.84$. Alternatively, we could have computed $\kappa = 1 - \hat{I}$ with the same result.

For both the three and two measure cases, the reliability estimates are biased due to the failure of the parallel assumptions to hold. The result suggests that $\hat{I}$ may be biased upward since, as shown previously, $D^2_{AB} > 0$. However, this may not be the case if $\rho_{AB} \neq 0$. Since it appears that $D^2_{BC} = 0$ can be assumed, the assumption of parallel measures seems to be more plausible for $B$ and $C$. Computing $\hat{I}$ using $B$ and $C$ only yields $\hat{I} = 0.25$, which is considerably higher than the value computed for $A$ and $B$. This result is somewhat unexpected since if $B$ and $C$ are parallel and $D^2_{AB} > 0$, $\hat{I}$ computed on $B$ and $C$ should be less than $\hat{I}$ computed for $A$ and $B$. One possible explanation is that $\rho_{AB}\sqrt{\mathrm{SRV}_A \mathrm{SRV}_B} > D^2_{AB}$ which, by (9), results in a negative bias in $\hat{I}$ computed from $A$ and $B$.

## 3. The truth as a latent variable: Latent class models

Further, insights into the structure and effects of classification error can be obtained by considering a classification probability model (Biemer and Stokes, 1991; Tenenbein, 1979; Mote and Anderson, 1965). Such models assume the existence of a true value underlying each observation in a survey. However, the true value is assumed to be unobservable in general. In that sense, the true value may be regarded as a latent variable in a model for the observations. We first consider the simple case of a dichotomous variable and only two measurements: for example, an interview and a reinterview value. Generalizations to polytomous variable and multiple repeated measurements will be considered subsequently.

### *3.1. Two measurements*

As before, let $y_i$ denote a dichotomous observed variable for the $i$th sample unit with corresponding unobserved true value $\mu_i, i = 1, \ldots, n$. The error in $y_i$, viz. $y_i - \mu_i$, is to be assessed. Let $\theta_i = P(y_i = 0|\mu_i = 1)$ and $\phi_i = P(y_i = 1|\mu_i = 0)$ denote the misclassification probabilities for the $i$th unit. The true score can be rewritten for all $i$ as

$$P_i = \mu_i(1 - \theta_i) + (1 - \mu_i)\phi_i. \tag{10}$$

By substituting (10) for $P_i$ in (6) and simplifying, it can be shown that $\mathrm{SV} = \pi(1-\pi)(1-\theta-\phi)^2 + \gamma_{\theta\phi}$ and $\mathrm{SRV} = \pi\theta(1-\theta) + (1-\pi)\phi(1-\phi) - \gamma_{\theta\phi}$ (see Biemer and Stokes, 1991), where $\pi = E(\mu_i)$, the true population proportion, $\gamma_{\theta\phi} = \pi\sigma^2_\theta + (1-\pi)\sigma^2_\phi$, where $\sigma^2_\theta = \mathrm{Var}(\theta_i|\mu_i = 1)$ and $\sigma^2_\phi = \mathrm{Var}(\phi_i|\mu_i = 0)$, $\theta = E(\theta_i)$, and $\phi = E(\phi_i)$. The component $\gamma_{\theta\phi}$ reflects the variation in error probabilities in the population. It is regarded as a nuisance parameter in the literature of classification error. Indeed, the models to be introduced subsequently will assume that it is negligible under the so-called *homogeneity assumption*. Violating this assumption can have serious consequences for classification error analysis. However, as we shall see, the homogeneity assumption can usually be satisfied (at least approximately) in practice by stratifying the population into mutually exclusive groups that are defined so that $\gamma_{\theta\phi} \approx 0$ within each group. The analysis is then carried out in each group assuming that groups are homogeneous with respect to $\theta_i$ and $\phi_i$. This approach will be discussed in more detail subsequently.

From (10), we obtain $E(p) = \pi(1-\theta) + (1-\pi)\phi$ and thus, the bias in the estimate is given by $B(p) = -\pi\theta + (1-\pi)\phi$. Let us consider the implications of these results for data quality evaluations. First, note that the bias can be rewritten as $N \times B(p) = -N_{0|1} + N_{1|0}$, where $N_{0|1} = N\pi\theta$ is the number of false negative classifications and $N_{1|0} = N(1-\pi)\phi$ is the number of false positive classifications. Thus, the bias will be 0 if (a) there are 0 misclassifications or (b) the expected number of false negative classifications equals the expected number of false positive classifications, that is, the two types of misclassifications exactly cancel each other in expectation. If false negatives outnumber false positives, the bias will be negative. If the opposite is true, then the bias will be positive.

Second, consider the form of the index of inconsistency under this model given by

$$I = \frac{\pi\theta(1-\theta) + (1-\pi)\phi(1-\phi) - \gamma_{\theta\phi}}{PQ}. \tag{11}$$

Note that $I$ is a nonlinear function of $\pi$, $\phi$, and $\theta$, which is difficult to interpret, even when $\gamma_{\theta\phi}$ is ignored. In particular, $I$ does not directly reflect the error in a classification process since it can be quite large even if both error probabilities are quite small. To illustrate, suppose the prevalence of an item is quite small, say $\pi = 0.01$ and let the error probabilities also be small, say $\theta = \phi = 0.005$. In this situation, $I$ is still quite large, that is, $I = 0.50$, denoting high inconsistency/ poor reliability. Without changing the error probabilities, suppose now that $\pi$ is higher, say 0.05. The index drops to 0.10, denoting good reliability. These examples illustrate the difficulty in interpreting $I$ for categorical measures.

In addition, $I$ has limited utility for evaluating measurement bias. For example, in the examples above, the parameter values that yield an $I$ of 0.10 (which is small) also yield a relative bias (i.e., $B(p)/\pi$) of $-0.9$ or $-90\%$ (which is quite large). The parameter values that yield a high index ($I = 0.5$) correspond to a relative bias of 0. Similar examples can be constructed to illustrate the point that magnitude of $I$ can and often does belie the magnitude of $B(p)$. These undesirable properties of $I$ (and $R$) suggest that they are poor metrics for gauging the magnitude of measurement error in categorical survey variables.

On the other hand, the parameters $\pi$, $\theta$, and $\phi$ provide all the information one requires to know the bias, the total variance, the SV, and the SRV for the estimator $p$. For categorical variables, we believe that the goal of survey measurement error evaluation should estimate $\pi$, $\theta$, and $\phi$ or comparable parameters whenever possible. The next section discusses methods for obtaining estimates of these parameters.

### 3.2. *Estimation of $\pi$, $\theta$, and $\phi$*

We saw in the last section that estimates of $\pi$, $\theta$, and $\phi$ are the basic building blocks for constructing the estimates of the mean squared error of $p$ and its components including bias, SRV, SV, $I$, and $R$. In this section, we discuss two methods for estimating these parameters. The first relies on knowledge of gold standard measurements, that is, observations $y_i$ that have the property that $y_i \doteq \mu_i$. Such measurements are very difficult to obtain in practice; however, studies have attempted to obtain them from reconciled reinterview surveys (see Forsman and Schreiner, 1991); in-depth probing reinterviews; record check studies (see Biemer, 1988); blood, urine, hair, or other biological specimen collections (Harrison, 1997); or any other method that yields essentially error-free measurements.

To illustrate the gold standard approach to estimation, suppose that in Table 1, the column classification (labeled $B$) is assumed to be the true classification. Then $p_{11} + p_{01}$ is an estimator of $\pi$; $p_{01}(p_{11} + p_{01})^{-1}$ is an estimator of $\theta$; $p_{10}(p_{10} + p_{00})^{-1}$ is an estimator of $\phi$; and $p_{10} - p_{01}$, referred to as the *net difference rate*, is an estimator of the bias in $p$.

Unfortunately, the literature provides few examples in which the gold standard approach has been applied successfully to obtain valid estimates of $\pi, \theta$, and $\phi$. In fact, many articles show that reconciled reinterview data can be as erroneous as the original measurements they were intended to evaluate (see, e.g., Biemer and Forsman, 1992; Biemer et al., 2001; Sinclair and Gastwirth (1996)). In addition, administrative records data are quite often inaccurate and difficult to use (Jay et al., 1994; Marquis, 1978) as a result of differences in time reference periods and operational definitions, as well as errors in the records themselves. Even biological measures such as hair analysis and urinalysis used in studies of drug use contain substantial false positive and false negative errors for detecting some types of drug use (see, e.g., Visher and McFadden, 1991). An alternative approach is to use a model that expresses the likelihood of a sample of observations in terms of $\pi, \theta$, and $\phi$ and then use maximum likelihood estimation to estimate these parameters. This is essentially the idea of LCA.

For example, a survey may ask two questions about smoking behavior that are worded slightly differently but designed to measure the same behavior. This information may be sufficient for estimating the misclassification parameters for both questions using the so-called Hui–Walter method (Hui and Walter, 1980) described in the next section.

### *3.3. The Hui–Walter method*

Assume that the row ($A$) and column ($B$) indicators in Table 1 are parallel measurements. Following notational conventions for LCA, the true value for unit $i$ (previously denoted $\mu_i$) is denoted by $X$ (with implied subscript $i$). The parallel assumptions can be restated in this notation as follows:

$$\text{(i) } \mathrm{P}(A = 1|X = x) = \mathrm{P}(B = 1|X = x) \text{ and}$$

$$\text{(ii) } \mathrm{P}(A = 1,\ B = 1|X = x) = \mathrm{P}(A = 1|X = x)\mathrm{P}(B = 1|X = x).$$

Assumption (i) implies that the error probabilities are equal for both trials and assumption (ii) (referred to as *local independence*) implies that the classification errors are independent between trials. Under these assumptions, the observed cell counts $(n_{11}, n_{10}, n_{01}, n_{00})$, where $n_{ab} = np_{ab}$, follow a multinomial distribution with parameters

$(n, P_{11}, P_{10}, P_{01}, P_{00})$, where $P_{ab} = \mathrm{P}(A = a, B = b)$ for $a, b = 0, 1$. Note that $P_{ab}$ can be written as

$$\begin{aligned} P_{11} &= \mathrm{P}(X=1)\mathrm{P}(A=1, B=1|X = 1) + \mathrm{P}(X = 0)\mathrm{P}(A = 1, B = 1|X = 0) \\ &= \pi(1-\theta)^2 + (1-\pi)\phi^2 \\ P_{10} &= P_{01} = \pi\theta(1-\theta) + (1-\pi)\phi(1-\phi) \qquad (12) \\ P_{00} &= \pi\theta^2 + (1-\pi)(1-\phi)^2. \end{aligned}$$

Thus, the likelihood of $\pi$, $\theta$, and $\phi$, given $\mathbf{n} = (n_{11}, n_{10}, n_{01}, n_{00})$, is given by

$$L(\pi, \theta, \phi|\mathbf{n}) = K \prod_{a,b=0,1} P_{ab}^{n_{ab}}, \qquad (13)$$

where $K$ is the usual combinatorial constant. Maximizing this likelihood with respect to $\pi, \theta$, and $\phi$ will yield the corresponding maximum likelihood estimates (MLEs); however, since the number of parameters (viz., 3) exceeds the degrees of freedom for Table 1 (viz., 2 df), a unique maximum does not exist and the parameters are said to be unidentifiable (Fuller, 1987). This problem can be rectified by constraining the parameters in some way. If the constraint $\theta = \phi = \varepsilon$, say, is imposed, then the number of parameters is reduced by 1 and the model becomes identifiable. In general, ensuring that the number of parameters does not exceed the available degrees of freedom is no guarantee of identifiability of LCA models. Other methods have been proposed for determining whether a model is identifiable (see, e.g., Goodman, 1974). Identifiability is discussed further in Section 3.5.

Suppose we relax assumption (i) for parallel measures but retain assumption (ii). Now the expected cell counts for Table 1 involve five parameters: $\pi$, $\theta_A$, $\theta_B$, $\phi_A$, and $\phi_B$, where $\theta_A$ and $\phi_A$ denote the false positive and false negative probabilities, respectively, for measure $A$ with analogous definitions for $\theta_B$ and $\phi_B$ for measure $B$. To achieve identifiability, we use a device suggested by Hui and Walter (1980).

Let $G$ denote a grouping variable with two categories, $g = 1, 2$, say, for example, gender. Denote the observed counts in each cell of the GAB table by $n_{gab}$ and note that $n_{gab}$, $g = 1, 2$; $a = 0, 1$; and $b = 0, 1$ follows a multinomial distribution with 10 parameters, viz., $\pi_g$, $\theta_{Ag}$, $\theta_{Bg}$, $\phi_{Ag}$, and $\phi_{Bg}$, where $\pi_g = \mathrm{P}(X = 1|G = g)$, $\theta_{Ag} = \mathrm{P}(A = 1|G = g)$, and so on. Since there are only seven degrees of freedom for the GAB table, the number of parameters must be reduced to seven or fewer as a necessary condition for identifiability. Hui and Walter (1980) show that an identifiable model with seven parameters can be obtained by introducing the restrictions: (a) $\theta_{A1} = \theta_{A2} = \theta_A$, (b) $\theta_{B1} = \theta_{B2} = \theta_B$, (c) $\phi_{A1} = \phi_{A2} = \phi_A$, and (d) $\phi_{B1} = \phi_{B2} = \phi_B$, that is, the misclassification probabilities are constrained to be equal across the two groups, $G$. Further, the two groups must be chosen so that $\pi_1 \neq \pi_2$, that is, the prevalence rates for the two groups must differ. These constraints produce a saturated model, that is, the number of parameters exactly equals the number of degrees of freedom, and hence, no residual degrees of freedom are left for testing model fit.

Under the Hui–Walter assumptions, the GAB likelihood is given by

$$L(\pi_g, \theta_{Ag}, \phi_{Ag}, \theta_{Bg}, \phi_{Bg}|\mathbf{n}) = K \prod_{g,a,b} P_{gab}^{n_{gab}}, \qquad (14)$$

where $P_{gab} = \mathrm{P}(G=g,\ A=a,\ B=b)$ for $g=1,2$, $a=0$, 1, and $b=0$, 1. For example, $P_{g11} = \mathrm{P}(G=g)[(1-\theta_A)(1-\theta_B)\pi_1 + \phi_A\phi_B\pi_2)]$, $P_{g10} = \mathrm{P}(G=g)[(1-\theta_A)(1-\theta_B)\pi_1 + \phi_A(1-\phi_B)\pi_2)]$, and so on. The Expectation-Maximization (EM) algorithm (Dempster et al., 1977; Hartley, 1958) has been implemented in a number of software packages for maximizing likelihood functions like (19). One that is freely available and easy to use is $\ell$EM (Vermunt, 1997), which is used extensively for illustrations in this chapter. Others include MLSSA, MPlus, and Latent Gold.

### *3.3.1. Example 2: Estimation of data collection mode bias*

The next illustration presents a useful and interesting variation of the Hui–Walter model when estimating mode effects in split-sample experiments. Taken from Biemer (2001), it is based upon an application of LCA to compare the quality of data obtained by face-to-face interviewing with that obtained by telephone interviewing for the 1994 National Health Interview Survey (NHIS). Classification error was estimated separately for two modes of interview through an LCA of test-retest reinterview data.

The study was conducted in two states: Texas (TX) and California (CA). Face-to-face interviews were provided by the NHIS sample in those states. Simultaneously, random-digit dialing (RDD) samples were selected in each state and interviewed using an abbreviated version of the NHIS questionnaire. Both the NHIS and RDD sample respondents were reinterviewed by telephone within two weeks of the original interview. The NHIS had a response rate of 81% for the main survey and 69% for the reinterview, whereas the RDD survey response rates were 60% for the main survey and 74% for the reinterview. Only cases that responded to both the interview and the reinterviews were retained in the analysis. Data for one characteristic (smoking in the home) are presented in Table 3.

Let $G = 1$ denote the NHIS sample and $G = 2$ denote the RDD sample. As before, $A$ will denote the interview classification and $B$ the reinterview classification. Let $\pi_g$ denote the true prevalence rate for target population members in group $g$, $\theta_{gt}$ denote the false negative probability for group $g$ at time $t$, and $\phi_{gt}$ the false positive probability for group $g$ at time $t$ for $t = 1, 2$. Further, let $P_{ij|g}$ denote the probability of a unit in group $g$ being classified in the $(i,\ j)$ cell of the interview–reinterview table.

Key in this approach is to assume that the error probabilities are the same for all interviews conducted by the same mode of interview, that is, assume

$$\begin{aligned}&\theta_{12} = \theta_{21} = \theta_{22} = \theta_{\mathrm{TEL}} \text{ and } \theta_{11} = \theta_{\mathrm{FF}},\\ &\phi_{12} = \phi_{21} = \phi_{22} = \phi_{\mathrm{TEL}} \text{ and } \phi_{11} = \phi_{\mathrm{FF}}.\end{aligned} \tag{15}$$

Table 3
Typical data table for the 1994 NHIS mode of interview evaluation

| Does Anyone Smoke Inside the Home? | | | |
|---|---|---|---|
| | | Reinterview by Telephone | |
| | | $B = 1$ | $B = 2$ |
| Interview by face-to-face | $A = 1, G = 1$ | 334 | 70 |
| | $A = 2, G = 1$ | 29 | 1233 |
| Interview by telephone | $A = 1, G = 2$ | 282 | 20 |
| | $A = 2, G = 2$ | 9 | 931 |

It therefore follows that, for the NHIS sample ($g = 1$), the probability of being in cell ($i$, $j$) in the GAB table is

$$P_{ij|g=1} = \pi_1 \theta_{\mathrm{FF}}^{1-i} \theta_{\mathrm{TEL}}^{1-j} (1-\theta_{\mathrm{FF}})^i (1-\theta_{\mathrm{TEL}})^j + (1-\pi_1) \phi_{\mathrm{FF}}^i \phi_{\mathrm{TEL}}^j (1-\phi_{\mathrm{FF}})^{1-i} (1-\phi_{\mathrm{TEL}})^{1-j}, \qquad (16)$$

and for the RDD sample ($g = 2$),

$$P_{ij|g=2} = \pi_2 \theta_{\mathrm{TEL}}^{2-i-j} (1-\theta_{\mathrm{TEL}})^{i+j} + (1-\pi_2) \phi_{\mathrm{TEL}}^{i+j} (1-\phi_{\mathrm{TEL}})^{2-i-j}. \qquad (17)$$

Note that the prevalence rates ($\pi_1$ and $\pi_2$) are assumed to differ since the response rates for the NHIS and the RDD data sets differed substantially; thus, the two achieved samples may represent different responding populations with different prevalence rates.

For the data in Table 3, the model implied by (15)–(17) is saturated since there are six degrees of freedom and six parameters: $\pi_g (g = 1, 2)$, $\theta_{\mathrm{FF}}$, $\phi_{\mathrm{FF}}$, $\theta_{\mathrm{TEL}}$, and $\phi_{\mathrm{TEL}}$. The estimates of these parameters, expressed as percentages, are: $\pi_1 = 22.9$, $\pi_2 = 25.1$, $\theta_{\mathrm{FF}} = 7.9$, $\phi_{\mathrm{FF}} = 4.1$, $\theta_{\mathrm{TEL}} = 4.8$, and $\phi_{\mathrm{TEL}} = 0.0$.

Biemer (2001) considered the case where all parameters of the model differed by state as well as when some parameters were set equal across states. As an example, it may be plausible to consider the case where the error parameters associated with telephone interviewing are the same for TX and CA since all telephone interviews in those states were conducted from the same centralized telephone facility using the same staff, interviewers, supervisors, questionnaires, and survey procedures. Thus, letting the subscript $s$ denote parameters specific to the state ($s = 1$ for TX and $s = 2$ for CA), we may also assume that $\theta_{\mathrm{TEL},s} = \theta_{\mathrm{TEL}}$ and $\phi_{\mathrm{TEL},s} = \phi_{\mathrm{TEL}}$ for $s = 1, 2$. Biemer tested this assumption as well as a number of alternative model specifications for each characteristic in the study.

In Section 2, we considered several estimators of reliability when two parallel measurements are available. Using the Hui–Walter model estimates of $\pi$, $\theta_A$, and $\phi_A$, an alternative estimator of $I_A$ can be derived when the two measurements are not parallel. Using the expression of $I$ in (11) and ignoring the variance term $\gamma_{\theta\phi}$, we replace each parameter by its MLE (denoted by a "hat" over the parameter symbol). This yields the following consistent estimator of $I_A$:

$$\hat{I}_A = \frac{\hat{\pi}\hat{\theta}_A(1-\hat{\theta}_A) + (1-\hat{\pi})\hat{\phi}_A(1-\hat{\phi}_A)}{p_A(1-p_A)} \qquad (18)$$

A similar estimator, denoted by $\hat{I}_B$, can be obtained for $I_B$.

## 4. Latent class models for three or more polytomous indicators

### 4.1. Probability model

When three or more indicators of a latent true value are available, more plausible models that impose weaker assumptions on the error parameters can be specified and inference is improved. In some cases, the three measurements are obtained from a survey followed by two reinterview surveys (e.g., see Biemer et al., 2001; Brown and Biemer, 2004). Such situations are rare, however, due to the difficulty and costs of conducting multiple interviews of the same households in a short span of time and solely for the purpose of quality evaluation. Multiple indicators more commonly arise when replicate measurements are embedded within the same interview. For modeling any number of indicators, a more general system of notation will now be introduced that is conventional within the LCA literature.

To fix the ideas, consider three polytomous indicators of a latent variable $X$ denoted by $A$, $B$, and $C$. There are two options for representing marginal and conditional probabilities. For example, $\mathrm{P}(X = x)$ can be denoted as $\pi_x^X$ or $\pi_x$. Similarly, conditional probabilities such as $\mathrm{P}(A = a|X = x)$ can be denoted as $\pi_{a|x}^{A|X}$ or $\pi_{a|x}$. Joint probabilities such as $\mathrm{P}(A = a, B = b, C = c)$ can be denoted as $\pi_{abc}^{ABC}$ or $\pi_{abc}$. In the following, we make use of both notational conventions when it is convenient and informative to do so.

In general, indicators need not be parallel to obtain models that are identified, although local independence usually must be assumed. The local independence assumption is equivalent to assuming

$$\pi_{abc|x} = \pi_{a|x}\pi_{b|x}\pi_{c|x}. \tag{19}$$

Let $\mathbf{n}$ denote the vector of cell counts, $n_{abc}$, for the ABC table and $\boldsymbol{\pi}$ denote the vector of parameters $\pi_x$, $\pi_{a|x}$, $\pi_{b|x}$, and $\pi_{c|x}$ for all values of $x$, $a$, $b$, and $c$. Under these assumptions, $\mathbf{n}$ follows a multinomial distribution with likelihood given by

$$L(\boldsymbol{\pi}|\mathbf{n}) = K\prod_a\prod_b\prod_c\left(\pi_{abc}^{ABC}\right)^{n_{abc}}, \tag{20}$$

where the cell probabilities may be rewritten in terms of the latent variable as

$$\pi_{abc}^{ABC} = \sum_x \pi_x^X \pi_{abc|x}^{ABC|X} = \sum_x \pi_x^X \pi_{a|x}^{A|X} \pi_{b|x}^{B|X} \pi_{c|x}^{C|X}. \tag{21}$$

Equation (21) is the likelihood *kernel* associated with the classical latent class model (LCM) for three indicators. For the situations considered in this chapter (i.e., the latent variable and its indicators having the same dimension), the model parameters are identifiable.

The interpretation of the parameters of the three-indicator LCM extends the two-indicator model interpretation of $\pi$, $\theta$, and $\phi$. For example, using the new notation for two indicators, $A$ and $B$, $\pi$ in the old notation may be rewritten as $\pi_1^X$ in the new notation. Likewise, $\theta_A$ and $\theta_B$ maybe rewritten as $\pi_{0|1}^{A|X}$ and $\pi_{0|1}^{B|X}$, respectively, and $\phi_A$ and $\phi_B$ are rewritten as $\pi_{1|0}^{A|X}$ and $\pi_{1|0}^{B|X}$, respectively. These parameters can be studied for their own intrinsic interest or can be used to compute estimates of reliability and bias as shown in the previous section.

As an example, to compute the estimates of the inconsistency ratios, $I_A$, $I_B$, and $I_C$, for the three-indicator model using MLEs of $\pi_x$, $\pi_{a|x}$, $\pi_{b|x}$ and $\pi_{c|x}$, we can use formulas that are analogous (18) but generalized for three indicators. Suppose $A$, $B$, $C$, and $X$ all have dimension $K$ and consider the estimator of $I_A$ denoted by $\hat{I}_A$. Let $A^*$ denote a hypothetical variable that is parallel to $A$ and consider the form of the $K \times K$ cross-classification table AA*. Let $\hat{n}_{ij}$ for $i = 1, 2, \ldots, K$ and $j = 1, 2, \ldots, K$ denote the estimates of the expected cell counts for the hypothetical AA* table, where

$$\hat{n}_{ij} = n\sum_{x=1}^{K} \hat{\pi}_x^X \hat{\pi}_{i|x}^{A|X} \hat{\pi}_{j|x}^{A^*|X} = n\sum_{x=1}^{K} \hat{\pi}_x^X \left(\hat{\pi}_{i|x}^{A|X}\right)^2, \tag{22}$$

where $\hat{\pi}^{A|X}_{i|x}$ and $\hat{\pi}^{X}_{x}$ are the latent class model parameter estimates of $\pi^{A|X}_{i|x}$ and $\pi^{X}_{x}$, respectively. Using the $\hat{n}_{ij}$, we can compute the index of inconsistency for each category, $i$, by forming the $2 \times 2$ table consisting of the cell frequencies $p_{11}$, $p_{10}$, $p_{01}$, and $p_{00}$ in Table 1, where $np_{11} = n_{ii}, np_{10} = \sum_{j \neq i} n_{ij}, np_{01} = \sum_{j \neq i} n_{ji}$ and $np_{00} = \sum_{j \neq i} \sum_{j' \neq i} n_{jj'}$. In other words, we form the $2 \times 2$ table by collapsing the $K \times K$ dimensional table around the category $i$. The formula for $\hat{I}_A$ is then given by applying to this $2 \times 2$ table. It can be shown that, when $A$ and $B$ are dichotomous, the estimator of $I$ is equal to (18).

The measurement bias for an arbitrary indicator, say $F$, can be written as $\text{Bias}(\pi^{F}_{f}) = \pi^{F}_{f} - \pi^{X}_{f}$ and estimated by replacing the parameter by its MLE. For example, write $\boldsymbol{\pi}^{\mathbf{F}} = \boldsymbol{\pi}^{\mathbf{F|X}}\boldsymbol{\pi}^{\mathbf{X}}$, where $\boldsymbol{\pi}^{\mathrm{F}} = [\pi^{F}_{1}\, \pi^{F}_{2}\, \ldots \pi^{F}_{K}]'$, $\boldsymbol{\pi}^{\mathbf{F|X}} = [\pi^{F|X}_{ij}, i = 1, 2, \ldots K;\ j = 1, 2, \ldots K]$, and $\boldsymbol{\pi}^{\mathbf{X}} = [\pi^{X}_{1}\, \pi^{X}_{2}\, \ldots \pi^{X}_{K}]'$. Then it follows that $\boldsymbol{\pi}^{\mathbf{X}} = (\boldsymbol{\pi}^{\mathbf{F|X}})^{-1}\boldsymbol{\pi}^{\mathbf{F}}$ and

$$\text{Bias}(\boldsymbol{\pi}^{\mathbf{F}}) = [\mathbf{I} - (\boldsymbol{\pi}^{\mathbf{F|X}})^{-1}]\boldsymbol{\pi}^{\mathbf{F}}. \tag{23}$$

An estimator of $\text{Bias}(\boldsymbol{\pi}^{\mathbf{F}})$ is obtained by replacing the parameters in by their corresponding MLEs.

## 4.2. *Model validity and identifiability*

### 4.2.1. *Model validity*

When interpreting the LCM estimates of measurement bias, reliability, or the error probabilities, it is important to keep in mind the four key assumptions underlying the estimation process, viz.:

(a) *unidimensionality*: the indicator variables all measure the same latent variable $X$;
(b) *local independence*: the classification errors are independent across indicators;
(c) *group homogeneity*: the variance of the misclassification probabilities is negligible within groups defined in the model; and
(d) *simple random sampling (SRS)*: all sampling units have equal probability of selection and no clustering was used in the design.

For the Hui–Walter method, two additional assumptions are made, viz.:

(e) misclassification probabilities are equal across levels of the grouping variable, and
(f) prevalence, $\pi_{x|g}$, varies across levels of the grouping variable.

Failure of any of these assumptions to hold could invalidate the estimates.

Assumption (a) is usually satisfied for test–retest reinterview surveys since the same question asked in the original interview is simply asked again in the reinterview survey. In some cases, minor wording changes are needed to reference the same time period in both interviews, which could alter the meaning of the question in some instances. The risk of violating assumption (a) is greater when the indicators are formed from questions embedded in the same interview.

The local independence (or equivalently, uncorrelated errors among indicators) assumption invoked for all the models considered thus far seems a strong assumption for most applications but particularly for embedded repeated measures. When asked about the same topic repeatedly in the same interview, respondents could recall their earlier responses and force response consistency on the topic, thus inducing correlated error. Likewise, correlated errors could result if a respondent uses the same process to formulate responses to multiple questions on the same topic. These problems can often be ameliorated by altering the wording of the questions sufficiently to mask their similarity. Failing that, however, error dependencies can to some extent be modeled. The introduction of additional terms in the model to reflect correlated errors will result in an unidentifiable model unless further restrictions on the model parameters are made (Hagenaars, 1988). As an example, one could impose the constraint $\pi_a^A = \pi_b^B = \pi_c^C$ for $a = b = c$ to free up degrees of freedom for modeling local dependence among the indicators (Hagenaars, 1988). Such constraints may not be plausible, particularly if different methods are used to generate the indicators.

Biemer and Wiesen (2002) describe a study of past-year marijuana smoking that used a number of embedded repeated measurements. Fig. 1 shows the wording of two questions in the study. Indicator $A$ was coded "1" if the response to question 1 in the figure indicated the respondent smoked marijuana anytime within the last 12 months; otherwise $A$ was coded "0." Indicator $B$ was coded "1" if the response to question 2 indicated that the respondent used marijuana on at least one day within the past 12 months; otherwise B was coded "0." Biemer and Wiesen tested for local dependence in their study by introducing grouping variables and then, following the strategy of Hui and Walter, equating some model parameters across groups to free up enough degrees of freedom for estimating the correlated error parameters. Their tests of significance of the dependence parameters provided no evidence of correlated error for the indicators used in their study. In general, however, the possibility of local dependence and its ramifications should always be considered in LCA.

Sinclair (1994) considered the effects of violations of the Hui–Walter assumptions on estimates of the model parameters. He found that the bias in the estimates of the Hui–Walter model parameters depends upon both the size of the prevalence rates in each group (i.e., $\pi_{1|1}^{X|G}$and $\pi_{1|2}^{X|G}$) and the magnitude of the classification error rates of the indicators. Large differences in prevalence rates (i.e., $\pi_{1|1}^{X|G} - \pi_{1|2}^{X|G}$) seemed to quell the biasing effects caused by the failure of the assumption of equal error rates for the two subpopulations to hold. Sinclair also examined the effects of local dependence in the Hui–Walter model. He found that the condition to be most problematic for estimating small error rates. In addition, the problem was exacerbated when the relative difference between the two prevalence rates was small.

As a check on the validity of the Hui–Walter estimates, Biemer and Bushery (2001) proposed a method that they applied to the estimation of labor force status classifications

1. How long has it been since you last used marijuana or hashish?

**A = "Yes" if either "Within the past 30 days" or "More than 30 days but within past 12 months;"**
**A = "No" if otherwise.**

2. Now think about the past 12 months from your 12-month reference date through today. On how many days in the past 12 months did you use marijuana or hashish?

**B = "Yes" if response is 1 or more days;**
**B = "No" otherwise.**

Fig. 1. Two embedded questions on past-year marijuana smoking.

in the Current Population Survey (CPS). Using data from the CPS reinterview program, they computed the index of inconsistency for three labor force categories—employed (EMP), unemployed (UNE) and not in the labor force (NLF)—using two methods: the traditional method (i.e., $\hat{I}$ in (8) and $\hat{I}_A$ in (18)) using the Hui–Walter estimates of the model parameters. Agreement of the two sets of estimates supports (but does not ensure) the validity of both methods since the methods rely on quite different assumptions and agreement would be unlikely if one or more of the assumptions were violated for either method. As an example, the 1993 traditional estimates for $I$ for the categories EMP, UNEMP, and NLF were $\hat{I} = 8.2, 33.5$, and 10.0, respectively, compared with $\hat{I}_A = 7.4$, 34.9, and 10.1, respectively. The close agreement of these estimates from two very different approaches supports the validity of both approaches for this application.

### *4.2.2. Identifiability*

Model *identifiability* for LCMs is related to parameter *estimability* in linear modeling. *Unidentifiability* is similar to the situation of too many unknowns or too few simultaneous equations. Essentially, a model is identified if there is one and only one set of parameter estimates that maximize the model likelihood function. Otherwise, the model is said to be unidentified. A necessary condition for identifiability is that number of parameters to be estimated should not exceed the number of degrees of freedom available for the data table. For $T$ indicator variables with the same number of categories, $C$, this condition can be specified as $\Delta = C^T - LT(C-1) - L \geq 0$, where $L$ is the number of classes in $X$ (Agresti, 2002). For example, if $T = 2, C = 2$, and $L = 2$, then $\Delta = -2$, indicating that the corresponding model is unidentified (as noted in Section 2). Recall that an identified model was obtained by adding constraints that reduced the number of parameters by two. For $T = 3$ indicators and $C = L = 2$ classes and categories, $\Delta = 0$ and the model is "just" identified.

There are situations where $\Delta \geq 0$, but the model is still not identified. This can happen when structural or observed 0s populate the data table rendering estimation of one or more parameters impossible. The model can usually be made identifiable by restricting some parameters or terms in the model to be 0 or equating two or more parameters. It may not be evident that the model is unidentified without testing for this condition. Failure to do so may lead the naive analyst to erroneously accept the estimates from an unidentified model.

Fortunately, it is fairly easy to detect unidentifiability. The preferred method is to test that the model information matrix is positive definite, which is a necessary and a sufficient condition for identifiability. Some software packages compute the information matrix and automatically test that it is of full rank.

Alternative methods can also be used to detect unidentifiability. One method is to run the estimation algorithm two or more times to convergence, using the same data, but different start values. If the same solution is reached using different start values, one can be more confident that the model is identified. However, this method is not a foolproof check on identification.

## *4.3. Log-linear model representations of LCMs*

The models introduced in the last section can easily be extended so that any number of grouping variables can be added either for their intrinsic interest or to improve model fit. In addition, by taking advantage of the equivalence between the categorical probability models and log-linear models, a much larger range of models may be specified. Haberman (1979) showed that the LCM can also be specified as a log-linear model for a table, say XAB, where $A$ and $B$ are observed and $X$ is latent. The log-linear modeling framework also provides greater opportunity to specify more parsimonious models. The misclassification models described previously can be viewed as special cases of a general model log-linear or logistic model for classification error.

To introduce the log-linear modeling framework, let $G$ and $H$ denote two grouping (or *exogenous*) variables; for example, $G$ may denote an individual's gender and $H$ may denote whether the individual is of Spanish origin or descent (i.e., Hispanicity). Let $A$, $B$, and $C$ denote three indicators of the latent true characteristic $X$. As we shall see, it may be advantageous to consider the order in which the indicator variables are observed. For example, suppose $A$ is measured first, then $B$ is measured followed by $C$. Such an ordering is often imposed by the order of questions in the questionnaire where it is assumed that respondents are not allowed to change responses to earlier questions. Alternatively, $A$ may be obtained in an interview, $B$ in a reinterview, and $C$ in a second reinterview where all three indicators are designed to assess the same latent characteristic $X$. Although it is not critical for many types of models of interest, the temporal ordering assumption can result in more parsimonious models since it implies that $A$ may influence $B$ and $C$ and $B$ may influence $C$; however, $C$ cannot influence $B$ or $A$ nor can $B$ influence $A$. These interrelationships can be represented by the path diagram in Fig. 2.

To simplify the discussion somewhat, we first consider models that omit $X$ and then show how the manifest variables only model can be extended to include $X$. The likelihood kernel for the saturated model corresponding to the GHABC table can be written as

$$\pi_{ghabc}^{GHABC} = \pi_{gh}^{GH}\pi_{a|gh}^{A|GH}\pi_{b|gha}^{B|GHA}\pi_{c|ghab}^{C|GHAB}, \tag{24}$$

which imposes the temporal ordering constraints described above. Additional degrees of freedom can be saved if we can also assume that $A$ does not influence $B$ and drop these terms from the conditional probability $\pi_{b|gha}^{B|GHA}$ to obtain $\pi_{b|gh}^{B|GH}$. In this manner, other restrictions can be easily imposed by eliminating one or more conditioning variables. For example, we could further restrict $\pi_{b|gha}^{B|GH}$ by $\pi_{b|g}^{B|G}$, $\pi_{b|h}^{B|H}$, or even $\pi_{b}^{B}$. As this approach saves degrees of freedom at the cost of reduced model fit, the significance of the omitted terms should be tested as described below.

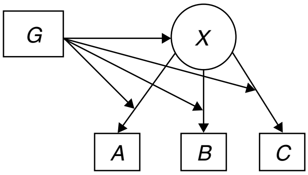


Fig. 2. Saturated path model with grouping variable, $G$.

Now consider how (24) can be expressed as a product of simultaneous logistic regression models. A well-known result in the logistic regression literature is that $\pi_{b|gha}^{B|GHA}$ can be written as

$$\pi_{b|gha}^{B|GHA} = \frac{\exp\left(u_b^B + u_{bg}^{BG} + u_{bh}^{BH} + u_{ba}^{BA} + u_{bgh}^{BGH} + u_{bga}^{BGA} + u_{bha}^{BHA} + u_{bgha}^{BGHA}\right)}{1 + \exp\left(u_b^B + u_{bg}^{BG} + u_{bh}^{BH} + u_{ba}^{BA} + u_{bgh}^{BGH} + u_{bga}^{BGA} + u_{bha}^{BHA} + u_{bgha}^{BGHA}\right)}, \tag{25}$$

where the $u$-variables are the usual log-linear model parameters to be estimated (see, e.g., Agresti, 2002). Taking the logit of both sides of (25), one obtains

$$\begin{aligned}\operatorname{logit}\left(\pi_{b|gha}^{B|GHA}\right) &= \log\left(\frac{\pi_{b|gha}^{B|GHA}}{1 - \pi_{b|gha}^{B|GHA}}\right) \\ &= u_b^B + u_{bg}^{BG} + u_{bh}^{BH} + u_{ba}^{BA} + u_{bgh}^{BGH} + u_{bga}^{BGA} + u_{bha}^{BHA} + u_{bgha}^{BGHA}.\end{aligned} \tag{26}$$

Likewise, the other probabilities on the right hand side of (24) can be rewritten in terms of exponential models with corresponding logit parameterizations. Goodman (1973) named these models *modified path models*. This approach leads to more parsimonious models since certain higher-order interaction terms can be excluded without excluding an entire variable. For example, dropping $u_{bgha}^{BGHA}$ from produces the model

$$\operatorname{logit}\left(\pi_{b|gha}^{B|GHA}\right) = u_b^B + u_{bg}^{BG} + u_{bh}^{BH} + u_{ba}^{BA} + u_{bgh}^{BGH} + u_{bga}^{BGA} + u_{bha}^{BHA}. \tag{27}$$

This model specification cannot be expressed as a simple product of conditional probabilities as in (24). Thus, the modified path model approach provides much greater flexibility for specifying and fitting parsimonious models.

All the models considered thus far are *hierarchical* models in the sense that they include all lower-order interaction terms involving $B$ and the other variables contained in higher-order terms in the model. For example, since $u_{bgh}^{BGH}$ is in the models, so are $u_b^B u_{bh}^{BH}$ and $u_{bg}^{BG}$. In this chapter, only hierarchical logit models will be considered. We shall adopt the convention of representing hierarchical models simply by excluding all implied terms. For example, to represent (26), we can write {BGHA} since all other terms in the model are implied. Likewise, (27) can be represented by {BGH, BGA, BHA}. For the full model in (24), we combine the four submodels in braces separated by semicolons and write as {GH; AGH; BGH BGA BHA; CGH CGB CHB}. In this expression, note that the submodel $\pi_{c|ghab}^{C|GHAB}$ has been reduced to {CGH CGB, CHB}.

Regarding estimation, Goodman (1973) demonstrated that the MLEs for the parameters of an unrestricted modified path model can be estimated by factoring the likelihood into terms corresponding to each submodel and then maximizing each submodel separately. Vermunt (1996, Appendix E.2) shows that the estimates so obtained are equivalent to those obtained by estimating all the submodels simultaneously by maximizing the full likelihood function. For example, the MLE's for $u_b^B$, $u_{bh}^{BH}$, and $u_{bg}^{BG}$ are the same whether they are obtained by maximizing the likelihood of the BGH table separately or by maximizing the likelihood of the full model. This approach has been implemented in the $\ell$ EM software (Vermunt, 1997).

Adding latent variables to modified path models is straightforward. For example, the three-indicator LCM in can be represented in our notation as {X; AX; BX; CX}. If we add grouping variables $G$ and $H$, many more identifiable models can be specified. For example, we can rewrite $\pi_{ghabc}^{GHABC}$ as

$$\pi_{ghabc}^{GHABC} = \pi_{gh}^{GH} \sum_{x} \pi_{x|gh}^{X|GH} \pi_{a|ghx}^{A|GHX} \pi_{b|ghax}^{B|GHAX} \pi_{c|ghabx}^{C|GHABX}. \tag{28}$$

However, the parameters are unidentified. An identifiable but fully saturated LCM can be obtained by eliminating the intra-indicator interactions, viz.,

$$\pi_{ghabc}^{GHABC} = \pi_{gh}^{GH} \sum_{x} \pi_{x|gh}^{X|GH} \pi_{a|ghx}^{A|GHX} \pi_{b|ghx}^{B|GHX} \pi_{c|ghx}^{C|GHX}, \tag{29}$$

which is equivalent to assuming local independence within each level of $GH$. Now using log-linear model notation, more parsimonious models can be specified for each term in this model, for example, $\pi_{a|ghx}^{A|GHX}$ could be simplified to {AGX AHX}, $\pi_{b|ghx}^{B|GHX}$ could be reduced to {BX}, and $\pi_{c|ghx}^{C|GHX}$ replaced by {CGX}. The resulting model is {GHX; AGX AHX; BX; CGX}. Modified path models can be fit using software packages such as ℓEM, Mplus (Muthen and Muthen, 1998–2005) or similar packages that have implemented the EM algorithm (Dempster et al., 1977; Hartley, 1958) as the primary method of estimation.

By adopting a log-linear modeling framework for LCA, many diagnostic measures for estimating and testing logit models are also available for testing and assessing the fit of LCMs. For example, the usual Pearson $\chi^2$ statistic ($X^2$) and the likelihood ratio $\chi^2$ statistic ($L^2$) are both distributed as $\chi^2_{df}$ random variables for LCAs, where $df$ denotes the model degrees of freedom. Their associated p-values can be used to identify models that provide an adequate fit to the data using the usual criterion of p $\geq$ 0.05 for an adequate fit. The $L^2$ statistic is also particularly useful for comparing two nested models by the method of differencing their $L^2$s using standard procedures for log-linear models (see, e.g., Agresti, 2002). Another useful measure of model fit is the dissimilarity index given by $D = \sum_{k} |n_k - \hat{m}_k|/2n$, where $n_k$ is the observed and $\hat{m}_k$ is the model estimated count in cell $k$. It is the smallest proportion of observations that would need to be reallocated to other cells to make the model fit perfectly. As a rule of thumb, $D$ should be less than 0.05 for a well-fitting model. For comparing non-nested models, the Bayesian Information Criterion (BIC) or the Akaike Information Criterion (AIC) can be applied, where

$$\begin{aligned} \text{BIC} &= -2\log L + (\log n) \times npar \\ \text{AIC} &= -2\log L + 2npar. \end{aligned} \tag{30}$$

When several models fit the data, the model with the smallest BIC (or AIC) is deemed the best since, by these measures, the competing criteria of goodness of fit and parsimony are balanced.

### *4.3.1. Example 3. Errors in self-reports of marijuana use*

Biemer and Wiesen (2002) consider the case of three measurements obtained in a single interview in an application to the National Household Survey on Drug Abuse (NHSDA). They defined three indicators of past-year marijuana use in terms of the questions asked at various points during the interview. Indicator $A$ was the response to the recency of use (or *recency*) question (Fig. 1, question 1) and Indicator $B$ was the response to the frequency of use (or *frequency*) question (Fig. 1, question 2). Indicator $C$ was a composite measure combining many questions about past-year marijuana use that appeared on the so-called *drug answer sheet*. It was coded "1" or "yes" if any affirmative response to those questions was obtained; otherwise, it was coded "0" or "no." This study was focused on estimating the false positive and false negative probabilities separately for each indicator. The composite indicator was constructed primarily for the sake of model identifiability. Three years of NHSDA data were analyzed: 1994, 1995, and 1996.

The models Biemer and Wiesen considered were limited to simple extensions of the basic LCM for three measurements incorporating multiple grouping variables defined by age ($G$), race ($R$), and sex ($S$). The simplest model they considered contained 54 parameters. This model allows the prevalence of past-year marijuana use, $\pi_x$, to vary by age, race, and sex; however, the error probabilities, $\pi_{a|x}$, $\pi_{b|x}$, and $\pi_{c|x}$ are constant across these grouping variables. The most complex model they considered contained 92 parameters and allowed error probabilities to vary by the three grouping variables. Since $A$, $B$, and $C$ were collected in the same interview, the possibility of locally dependent errors was also considered in the analysis as noted in Section 4.2.

The best model identified in their analysis was a locally independent model containing 72 parameters that incorporated simple two-way interaction terms between each grouping variable and each indicator variable. (The reader is referred to the original paper for the full specification and interpretation of this model.)

Estimates of the classification error rates for all the three indicators of past-year drug use were derived from the best model. Table 4 shows the estimated classification error rates (expressed as percentages) for the total population and for all the three years. Standard errors, which are provided in parentheses, assume SRS and do not take into account the unequal probability cluster design of the NHSDA. Consequently, they may be understated.

A number of points can be made from these results. First, note that the false positive rates for all the three indicators are very small across all the three years except

Table 4
Comparison of estimated percent classification error by indicator*

| True Classification | Indicator of Past-Year Use | 1994 | 1995 | 1996 |
|---|---|---|---|---|
| Yes ($X = 1$) | Recency = No (A = 2) | 7.29 (0.75) | 8.96 (0.80) | 8.60 (0.79) |
| | Direct = No ($B = 2$) | 1.17 (0.31) | 0.90 (0.28) | 1.39 (0.34) |
| | Composite = No ($C = 2$) | 6.60 (0.70) | 5.99 (0.67) | 7.59 (0.74) |
| No ($X = 2$) | Recency = Yes ($A = 1$) | 0.03 (0.02) | 0.01 (0.01) | 0.08 (0.02) |
| | Direct = Yes ($B = 1$) | 0.73 (0.07) | 0.78 (0.07) | 0.84 (0.07) |
| | Composite = Yes ($C = 1$) | 4.07 (0.15) | 1.17 (0.08) | 1.36 (0.09) |

* Standard errors are shown in parentheses.

for indicator $C$ in 1994 where it is 4.07%: more than four times that of the other two measurements. This was analyzed further by comparing the questions comprising $C$ for all the three years. This analysis revealed that, prior to 1995, one of the questions used for constructing $C$ seemed quite complicated and potentially confusing to many respondents. For 1995 survey, the question was dropped. Therefore, a plausible hypothesis for the high false positive rate for $C$ in 1994 is the presence of this highly complex and confusing question.

To test this hypothesis, a new indicator was created from the 1994 data by deleting the problematic question from indicator $C$ and rerunning the model. This resulted in a drop of the false positive rate for the composite measure from 4.07% to only 1.23%, which was consistent with 1995 and 1996 false positive rates for $C$. Clearly, LCA was successful at identifying the problem item in the composite question.

A second finding of interest that can be observed in Table 4 focused on the false negative error probability estimates. Note that both $A$ and $C$ have much higher false negative rates than $B$, a pattern that is repeated for all the three years. To uncover the possible reason, consider the wording differences for questions 1 and 2 in particular. As shown in Fig. 1, question 1 asks *how long ago* marijuana was used while question 2 asks *how many times* marijuana was used within the last year. Very infrequent users of marijuana may not consider that they are "users" of the substance or may object to being labeled as "marijuana users." These infrequent users might respond "no" to question 1 while responding "yes" to question 2. To test this hypothesis, responses to the two questions were cross-classified. Consistent with the hypothesis, 59% of respondents answering "no" to question 1 were infrequent (i.e., 1–2 days) users compared with only 16% who answered "yes" to both questions 1 and 2.

## 5. Some advanced topics

### 5.1. Markov latent class models

LCMs can also be applied when measurements of the same phenomena are made at different time points as in the case of a panel survey. Suppose now $A$, $B$, and $C$ denote the values of the same categorical variable at three time points. Remarkably, only these data are necessary to estimate the classification error in the measurements under the Markov latent class model (MLCM) assumptions, that is, no other reinterview or data of repeated measurements are required. MLCMs, first proposed by Wiggins (1973), resemble the LCMs described previously, except new parameters must be introduced into the models to allow the true characteristic to vary across the time points.

Let $X$, $Y$, and $Z$ denote latent variables corresponding to the true values of the characteristic at times 1, 2, and 3, respectively. Thus, $A$, $B$, and $C$ are indicators of $X$, $Y$, and $Z$, respectively, observed at times 1, 2, and 3, respectively. A key assumption for these models is the Markov assumption, which states that $P(Z = z|X = x, Y = y) = P(Z = z|Y = y)$, that is, knowing an individual's true state at time 2 is sufficient for predicting his/her true state at time 3. In other words an individual's time 1 status does not provide any additional information once the time 2 status is known. As an example, consider a panel survey question on current smoking behavior. The Markov assumption states an individual's smoking behavior for the last two panel waves is no more predictive of current smoking behavior than the individual's behavior at the last wave. As we shall see, MLCA models are surprisingly robust to violations of this assumption.

We further assume independent classification errors (I.C.E) across time points, referred to in the literature as the ICE assumption (Singh and Rao, 1995). It is analogous to the local independence assumption for LCA. Further restrictions are still needed to attain an identified model. For example, when $X$, $A$, B, and $C$ are all dichotomous random variables, there are 11 parameters to be estimated, viz., $\pi_1^X$, $\pi_{1|x}^{Y|X}(x = 1, 2), \pi_{1|y}^{Z|Y}(y = 1, 2), \pi_{1|x}^{A|X}(x = 1, 2), \pi_{1|y}^{B|Y}(y = 1, 2)$, and $\pi_{1|z}^{C|Z}(z = 1, 2)$, and only seven degrees of freedom. The usual restrictions are to equate the classification error probabilities across time points, that is, assume

$$\pi_{a|x}^{A|X} = \pi_{b|y}^{B|Y} = \pi_{c|z}^{C|Z} \tag{31}$$

for $a = b = c$ and $x = y = z$, referred to as the *time invariant errors* assumption. With these restrictions, the model is saturated but identified, and the likelihood kernel is

$$\pi_{abc}^{ABC} = \sum_x \sum_y \sum_z \pi_x^X \pi_{y|x}^{Y|X} \pi_{z|y}^{Z|Y} \pi_{a|x}^{A|X} \pi_{b|y}^{B|Y} \pi_{c|z}^{C|Z}, \tag{32}$$

subject to the constraints in (31). In shorthand notation, this model can be represented by {X; XY; YZ; AX; BY; CZ}. Additional degrees of freedom can be saved by restricting the transition probabilities, $\pi_{y|x}^{Y|X}$ and $\pi_{z|y}^{Z|Y}$, to be equal whenever $(x,y) = (y,z)$—referred to as *stationary* transitions.

The model can easily be extended to accommodate grouping variables or covariates. For example, consider the model in Fig. 3. In shorthand notation, this model is represented by {GX; GXY; GYZ; GAX; GBY; GCZ} with time invariant errors within groups, that is, $\pi_{a|xg}^{A|XG} = \pi_{b|yg}^{B|YG} = \pi_{c|zg}^{C|ZG}$, when $a = b = c$ and $x = y = z$ for all $g$. A number of additional assumptions can be explored in the search for model parsimony, which are as follows:

(1) nonhomogeneous, stationary transition probabilities: GXY = GYZ;
(2) homogeneous transition probabilities: replaces {GXY} by {XY} and {GYZ} by {YZ}; and

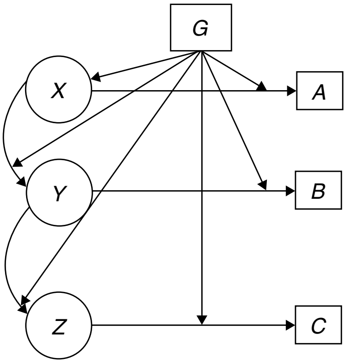


Fig. 3. Fully saturated MLCA with grouping variable, $G$.

(3) homogeneous error probabilities: replaces {GAX} by {AX}, which by the time invariant errors assumption implies {BY} replaces {GBY} and {CZ} replaces {GCZ}.

When four waves of data or more are available, models with less restrictive assumptions are identifiable. For example, Biemer and Tucker (2001) and Tucker et al. (2006) fit second-order MLCAs to four waves of data from the U.S. Consumer Expenditure Survey, thus relaxing the first-order Markov assumption of the present model.

### 5.2. *Methods for complex survey data*

The methods discussed heretofore assume a sample selected by SRS. For survey data analysis, this assumption is rarely satisfied. Survey samples usually involve some form of cluster sampling, stratification, and unequal probability sampling. The survey literature has shown that if data are collected under a complex sampling design and SRS is assumed, parameter estimates from many types of data analysis may be biased and their standard errors may be underestimated (see, e.g., Korn and Graubard, 1999, pp. 159–172). For LCA, local independence could be violated for clustered data if measurement errors also tend to be clustered. Likewise, oversampling certain groups in the population having higher error rates creates heterogeneity that could violate the group homogeneity assumption. As previously discussed, group heterogeneity can usually be rectified by adding grouping variables to the model that, in this case, aligns with the sampling domains. This may be infeasible if the number of sampling domains is quite large.

In some cases, ignoring the sampling design may be acceptable. Although some individual characteristics may be highly geographically clustered in a population, classification errors often exhibit much less clustering. If the focus of an investigation is on classification error, ignoring clustering may not have serious consequences. Patterson et al. (2002) provide some evidence of this in their study of the survey design effects on LCA estimates for dietary data. Other researchers have argued that sample weighting may not even be appropriate when the LCA is focused on misclassification. For example, in his discussion of the Patterson et al. paper, Vermunt (2002) argues that weighting does not appropriately account for the heterogeneity in the misclassification probabilities induced by unequal probability sampling. He suggests that adding grouping variables to account for the heterogeneity is a better way to address the problem.

More research is needed to evaluate the advantages and disadvantages of the various methods for applying LCA to survey data. Until then, we advocate the use of methods that at least take unequal probability sampling into account. One simple method for that advocated by Clogg and Eliason (1985) is to reweight and rescale the cell frequencies using the unit-level sample weights. This method will produce model unbiased parameter estimates although standard errors will tend to be negatively biased. In addition, the usual model diagnostics for assessing fit ($X^2$, $L^2$, BIC, etc) are no longer be valid and may result in model Type I error probabilities smaller than their nominal levels. Thus, models that in truth adequately describe the data may be falsely rejected.

To apply Clogg's scheme, let $(a,b,\ldots, v)$ denote a cell of the AB…V table. Let $n_{ab\ldots v}$ denote the unweighted sample size in the cell and let $W_{ab\ldots v}$ denote the sum of the sample weights (including postsurvey adjustments) for these $n_{ab\ldots v}$ observations. Then, replace $n_{ab\ldots v}$ by

$$n'_{ab\ldots v} = n\frac{W_{ab\ldots v}}{W}, \tag{33}$$

where $W = \sum_{a,b,\ldots,v} W_{ab\ldots v}$ to form the weighted and rescaled table, $\text{AB}\ldots\text{V}_{wtd}$, where the subscript denotes "weighted." The LCA proceeds using $\text{AB}\ldots\text{V}_{wtd}$ under the assumption of SRS. One strategy is to perform the LCA for both unweighted and weighted tables. If the estimates are fairly close, then they opt to use the unweighted estimates since they often exhibit greater stability than the weighted ones.

Fortunately, there have been recent advances in analysis of complex survey data using pseudo maximum likelihood (PML) methods (Pfeffermann, 1993). These methods have been implemented in some latent class software packages such as Latent Gold (Vermunt and Magidson, 2005) and Mplus (Muthen and Muthen, 1998–2005). These packages also take both unequal weighting and clustering into account when estimating standard errors using either replication methods such as jackknife or linearization.

To illustrate the PML method, write the likelihood for the AB…V table under SRS and $K$ latent classes as

$$L = \prod_a \prod_b \cdots \prod_v \left( \sum_{x=1}^{K} \pi_x^X \pi_{a|x}^{A|X} \pi_{b|x}^{B|X} \cdots \pi_{v|x}^{V|X} \right)^{n_{ab\ldots v}}. \tag{34}$$

The PML method provides design-consistent estimates of the model parameters by maximizing $L'$ which is $L$ after replacing $n_{ab\ldots v}$ by $n'_{ab\ldots v}$ defined in (33).

A third method for accounting for unequal probability sampling uses the unweighted cell counts with an offset, consisting of the log of the inverse of the average cell sample weight, in each cell of the contingency table (Clogg and Eliason, 1985). Under a correctly specified log-linear model for the population, this method will produce consistent estimates of model parameters. However, if the log-linear model is misspecified, the estimates will not be consistent. The PML method is often preferred because its estimates will be approximately unbiased for values of the population model parameters regardless of whether the model was correctly specified.

### *5.2.1. Example 4: MLCA estimates of labor force status misclassification*

A common application of MLCM is to model the classification error in labor force survey panel data. For example, Van de Pol and Langeheine (1997) applied these models to the Netherlands Labor Market Survey, Vermunt (1996) to the SIPP labor force series, and Biemer and Bushery (2001) and Biemer (2004a) to the CPS. The latter two papers also evaluated a number of the assumptions of MLCM for the CPS, including the Markov assumption, and provided some evidence of the empirical, theoretical, and external validity of the MLCM estimates of CPS classification error. The latter papers used MLCA to evaluate the accuracy of labor force for the revised CPS questionnaire that was introduced in 1994 and compared it with the accuracy of the original questionnaire that had been in use prior to 1994.

To illustrate the general approach, we use the three labor force categories defined in Example 2 and consider any three consecutive months of the CPS, say January, February, and March. Let $X$ denote an individual's true labor force status in January, that is,

$X = 1$ for EMP, $X = 2$ for UNEMP, and $X = 3$ for NLF. Define $Y$ and $Z$ analogously for February and March. Similarly, $A$, $B$, and $C$ will denote the observed labor force statuses for January February, and March, respectively, with the same categories as their corresponding latent counterparts. Similar models can be defined and fit for February, March, and April; March, April, and May; April, May, and June, etc. The resulting estimates can be compared or more importantly averaged together to form estimators having smaller variance. Models involving four consecutive months can include second-order Markov terms, that is, denoting the latent variables for the four months by $W$, $X$, $Y$, $Z$, the terms $Y|WX$ and $Z|XY$ can be included. The sample size will be reduced slightly if the analysis is restricted only to households that respond to four consecutive months. Vermunt (1996) describes some methods for retaining cases that do not respond at some panel waves.

For grouping variables, Biemer and Bushery (2001) and Biemer (2004a) considered age, race, gender, education, income, and self/proxy response. The last variable is an indicator of whether a subject's labor force status was obtained from the subject or from another person in the household (i.e., a proxy). In the CPS, the interview informant can change from month to month, and thus, self-proxy can be regarded as a time-varying covariate. However, in doing so, Biemer and Bushery found the models to be quite unstable and opted for a time-invariant self/proxy variable denoted by $G$, where $G = 1 \Leftrightarrow$ self-response in all three months, $G = 2 \Leftrightarrow$ self-response in exactly two months, $G = 3 \Leftrightarrow$ proxy response in exactly two months, and $G = 4 \Leftrightarrow$ proxy in all three months. This variable was highly significant in all the models they considered.

Biemer and Bushery analyzed data from three years—1993, 1994, and 1996—of the CPS, fitting models to each year separately. The data were weighted and rescaled data as per (33), and the $\ell$EM software was used. A wide range of models were fit to the data including the following five essential models in Table 5. In addition, all the models they considered assumed stationary classification error probabilities as a condition of identifiability.

Table 5 shows the fit diagnostics for these five models by year. As shown, only Model 4 provided an acceptable fit when the p-value criterion is used. Model 4 is the most general model in the table and allows the January–February and February–March transition probabilities to vary independently across the four self-proxy groups. The model further specifies that the error probabilities are the same for January, February, and March, but may vary by self-proxy group. This latter specification is consistent with the literature of survey methods (see, e.g., O'Muircheartaigh (1991); Moore, 1988). In addition, the dissimilarity index, $d$, for Model 4 is 0.3%, which indicates a very good model fit. Thus, the authors used Model 4 to generate the estimates of labor force classification error. These estimates and their standard errors appear in Table 6. As noted in previous examples, the standard errors are likely to be understated since the effects of unequal weighting and clustering were not taken into account in the analysis.

As shown in Example 2, for the true EMP and true NLF, the probability of a correct response is quite high: 98% and 97%, respectively. However, for the true UNEMP, the probability of a correct response varies across years from 72–84%. However, a surprising result from Table 6 is the magnitude of reporting accuracy for 1994 and 1995 compared to 1993. As the authors note, the CPS questionnaire was substantially redesigned in 1994 to increase the accuracy of the labor force status classifications, as well as other population characteristics. The results in Table 6 suggest that reporting

Table 5
Model diagnostics for alterative MLCA models by year

| Model | Year | *df* | *npar* | $L^2$ | p | BIC | *d* |
|---|---|---|---|---|---|---|---|
| Model 0 | 1993 | 90 | 17 | 645 | 0.00 | –320 | 0.048 |
| {X; XY; YZ; AX; BY; CZ\|XY = YZ} | 1995 | | | 697 | 0.00 | –275 | 0.044 |
| | 1996 | | | 632 | 0.00 | –325 | 0.045 |
| Model 1 | 1993 | 84 | 23 | 632 | 0.00 | –269 | 0.047 |
| {GX; GXY; GYZ; AX; BY; CZ\|GXY = GYZ} | 1995 | | | 668 | 0.00 | –240 | 0.043 |
| | 1996 | | | 585 | 0.00 | –308 | 0.044 |
| Model 2 | 1993 | 66 | 41 | 99 | 0.01 | –609 | 0.007 |
| {GX; XY; YZ; AX; BY; CZ} | 1995 | | | 146 | 0.00 | –567 | 0.008 |
| | 1996 | | | 159 | 0.00 | –543 | 0.010 |
| Model 3 | 1993 | 42 | 65 | 64 | 0.02 | –386 | 0.005 |
| {GX; GXY; GYZ; AX; BY; CZ} | 1995 | | | 82 | 0.00 | –372 | 0.005 |
| | 1996 | | | 83 | 0.00 | –364 | 0.010 |
| Model 4 | 1993 | 24 | 83 | 23 | 0.50 | –234 | 0.002 |
| {GX; GXY; GYZ; GAX; GBY; GCZ} | 1995 | | | 25 | 0.41 | –234 | 0.002 |
| | 1996 | | | 39 | 0.03 | –216 | 0.003 |

Table 6
Estimated labor force classification probabilities by group and year*

| | Observed Classification | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| True classification | EMP | | | UNEMP | | | NLF | | |
| | 1993 | 1995 | 1996 | 1993 | 1995 | 1996 | 1993 | 1995 | 1996 |
| EMP | 98.8 | 98.7 | 98.8 | 0.3 | 0.5 | 0.4 | 0.9 | 0.8 | 0.8 |
| | (0.1) | (0.1) | (0.1) | (0.11) | (0.1) | (0.1) | (0.1) | (0.1) | (0.1) |
| UNEMP | 7.1 | 7.9 | 8.6 | 81.8 | 76.1 | 74.4 | 11.1 | 16.0 | 17.0 |
| | (0.7) | (0.9) | (1.0) | (0.9) | (1.2) | (1.2) | (0.9) | (1.2) | (1.2) |
| NLF | 1.4 | 1.1 | 1.1 | 0.8 | 0.7 | 0.9 | 97.8 | 98.2 | 98.0 |
| | (0.1) | (0.1) | (0.1) | (0.1) | (0.1) | (0.1) | (0.1) | (0.1) | (0.1) |

* Standard errors are shown in parentheses.

accuracy is higher for the year prior to the major redesign (i.e., in 1993) than for the years following the redesign.

Subsequently, Biemer (2004a) applied MLCM to further explore this anomaly. His analysis considered the error associated with the two primary subclassifications of unemployed: persons who are unemployed and on layoff, and persons who are unemployed and looking for work. His analysis found that the primary cause of the anomaly in table 10 was a reduction in classification accuracy of persons who are on layoff. Using MLCA, he found that two questions on the revised questionnaire appear to be responsible. He found considerable error associated with the revised question: "LAST WEEK, did you do ANY work (either) for pay (or profit)?" He found that more than 50% of the error in the revised layoff classification is contributed by this question. In addition, he found considerable classification error in determining whether individuals reporting some type of layoff have a date or indication of a date to return to work. This question contributes between 30 and 40% of the layoff classification error. The combination of these two questions appears to explain the reduction in accuracy in UNEMP in the revised questionnaire.

Biemer and Bushery also compared the MLCA estimates of the CPS classification probabilities with similar estimates from the literature, for example, Fuller and Chua (1985), Poterba and Summers (1995), and the CPS reconciled reinterview program (Biemer and Bushery, 2001). Unlike the MLCA estimates, these other estimates relied on reconciled reinterview data, which was considered a gold standard. Biemer and Bushery found that the relative magnitude of the MLCA estimates across the labor force categories agrees fairly well with the previous estimates. The greatest differences occurred for the true unemployed population where the estimates of response accuracy from the literature were three to seven percentage points higher than corresponding MLCA estimates. One explanation for this difference is that the comparison estimates are biased upward as a result of correlations between the errors in interview and reinterview. Another explanation is that the MLCA estimates are biased downward as a result of the failure of the Markov assumption to hold. Both explanations may be true to some extent. However, Biemer and Bushery (2001) provides some evidence that failure of the Markov assumption in not likely to have an appreciable effect on estimates of classification error.

## 6. Measurement error evaluation with continuous variables

The preceding discussion dealt primarily with the evaluation of classification error, that is, measurement errors when both the latent variable and the manifest variable are discrete. Indeed, categorical variables tend to dominate survey data analysis. For completeness, this section will be devoted to a brief discussion of measurement errors in continuous variables, that is, latent and manifest variables that are measured on an interval or ratio scale. The literature for this area of measurement error analysis is quite extensive and, in many ways, more developed than the methodology for classification error. Disciplines such as econometrics, psychometrics, sociometrics, and statistics have all contributed to this vast area of research. Some recent books that summarize the key developments in this area are Wansbeek and Meijer (2001), Biemer et al. (1991, Section E), and Alwin (2007). Our aim is not to provide a comprehensive exposition of the topic. Rather, we describe a general framework for modeling and estimation emphasizing a few widely used methods. As we did in earlier sections of this chapter, we first consider techniques appropriate for cross-sectional surveys and then extend the ideas to panel surveys. Also as before, the focus will be restricted to quality criteria for single response variables or composite variables that are treated as such.

Section 2.1 considered some of the early developments in the estimation of measurement error variance for categorical variables that arose from the two-stage sampling model of the measurement process. Central to those results is the classic work by Hansen et al. (1961), which considered the estimation of SRV using continuous parallel indicators. The modern approach for estimating reliability with continuous data applies *structural equation models* or SEMs (see, e.g., Saris and Andrews, 1991). Maximum likelihood estimation may be used to obtain estimates of SEM coefficients, which give rise to estimates of the reliability ratio. The origins of SEM can be traced to Wright's (1918, 1921) work on *path analysis*. Since then, further developments have mostly occurred in the behavioral and social sciences. Growth was spurred by the development of software packages that facilitate the estimation of SEMs, particularly LISREL (Jöreskog and Sörbom, 1978).

The next section introduces the basic SEM measurement model that is analogous to the basic LCA model. The concepts of *validity* and *invalidity* will also be introduced. Together with reliability, these concepts are central to the SEM framework.

### 6.1. Reliability, validity, and invalidity in the SEM framework

Extending the notions of general model in Section 2.1, we assume that an observation on the $i$th individual at the $j$th trial follows the linear model

$$y_{ij} = \mu_i + \varepsilon_{ij}, \tag{35}$$

where $\mu_i$ is the true value of the characteristic having mean $\pi$ and variance $\sigma^2_\mu$, and $\varepsilon_{ij}$ is the measurement error with conditional mean and variance, $M_{ij}$ mean and variance $\sigma^2_{\varepsilon ij}$, respectively. In the SEM literature, the subscript $j = 1, \ldots, J$ is referred to as the "method" of measurement since it is assumed that questions, data collection modes, respondent types, or data sources may vary for each replicate measurement or trial. The variable $\mu_i$ is assumed to be latent since none of the methods will always produce the true value.

The model assumes that $\mathrm{Cov}(\varepsilon_{ij}, \varepsilon_{i'j'}) = 0$ for $(i, j) \neq (i', j')$, that is, measurement errors are uncorrelated among units and trials. The model also stipulates that $\mathrm{Cov}(\mu_i, \varepsilon_{ij'}) = 0$, which distinguishes the model from the simple dichotomous model discussed in Section 2, where it can be shown that by using the notation in Section 3, $\mathrm{Cov}(\mu_i, \varepsilon_{ij}) = -\pi(1-\pi)(\phi + \theta)$. While the uncorrelated error assumption seems reasonable for continuous data, it can still be violated. For example, for income data, the size of the measurement error may depend upon the income value. Suppose the error is proportional to the true value, that is, $\varepsilon_{ij} = \gamma_{ij}\mu_i$ and rewrite the right side of (35) as $\mu_i(1 + \gamma_{ij})$. The log transformation can be used to linearize the model. Alternatively, variables believed to be correlated with error such as education, gender, and race (in the case of income) could be included in the model that may result in independent residual error.

One of the strengths of SEM is the ability to simultaneously fit multiequation models that describe a system of measurement involving many related variables. In general, the SEM model incorporates two types of submodels, one for the latent variables (referred to as the *structural model*) and another for the error terms (referred to as the *measurement model* since it describes how the latent variables and the error terms are linked to form the measurement variables). In our discussion, we consider only very simple model structures. Readers interested in exploring more complex measurement error models are referred to Alwin (2007).

The model in (35) is too general to be of any practical use since it is implicitly overspecified. To reduce the number of parameters, various types of restrictions have been considered in the literature. One possibility considered previously is to assume parallel measurements, that is, assume $M_{ij} = M$ and $\sigma^2_{ij} = \sigma^2_\varepsilon$, for all $(i,j)$. However, these assumptions are both impracticable and unnecessarily restrictive. Another is to assume that the distribution of $\varepsilon_{ij}$ depends only on the method $(j)$. We considered a similar assumption in the dichotomous case by allowing the misclassification probabilities to vary across the repeated measurements. For SEM, it is customary to allow the mean and variance of the response distribution to depend only on the method as in the following model:

$$y_{ij} = \mu_i + \xi_{ij} + \varepsilon_{ij}, \tag{36}$$

where $\xi_{ij}$ is an independent, random component representing the interaction between the method and the individual, $\mathrm{E}(\xi_{ij}) = M_j$ and $\mathrm{Var}(\xi_{ij}) = \sigma^2_{\xi j}$. This model postulates that each implementation of the same method under the same identical survey conditions produces the value $\mu_i + \xi_{ij}$ for the $i$th individual apart from a random error term, $\varepsilon_{ij}$. Thus, the true score for the $i$th individual for the $j$th method is $E(y_{ij}|i, j) = \mu_i + \xi_{ij}$, which will be denoted by $T_{ij}$. The SV is $\sigma^2_\mu + \sigma^2_{\xi j}$ and the SRV is $\sigma^2_\varepsilon$. The reliability ratio for the $j$th measurement process can be written as

$$R_j = \frac{\sigma^2_\mu + \sigma^2_{\xi j}}{\sigma^2_\mu + \sigma^2_{\xi j} + \sigma^2_\varepsilon}. \tag{37}$$

Further, $M_j$ may be interpreted as the bias of the $j$th measurement process since $E(\bar{y}_j) = \mu + M_j$, where $\mu = E(\mu_i)$.

In SEM, $y_{ij}$ is often modeled in standardized form. For any variable $z$, the standardized form will be denoted by an asterisk in superscript as $z^* = (z - E(z)/\sqrt{\mathrm{Var}(z)}$. Rewriting model (36) in terms of the standardized variables, we obtain the following equations:

$$\begin{aligned} y^*_{ij} &= h_j T^*_{ij} + \varepsilon^*_{ij} \\ T^*_{ij} &= b_j \mu^*_i + g_j \xi^*_{ij} \end{aligned} \tag{38}$$

$$h_j = \sqrt{\frac{\mathrm{Var}(T_{ij})}{\mathrm{Var}(y_{ij})}},\ b_j = \sqrt{\frac{\sigma^2_\mu}{\mathrm{Var}(T_{ij})}} \text{ and } g_j = \sqrt{\frac{\sigma^2_{\xi j}}{\mathrm{Var}(T_{ij})}} \tag{39}$$

where $h_{ij}$ denotes the *reliability coefficient*, $b_j$ denotes the *validity coefficient* and $g_j$ is the *method coefficient*. Thus, for the $j$th measurement, reliability is $h^2_j$, the *measurement validity* is defined as $b^2_j$, *measurement invalidity* is defined as $(1 - b^2_j)$, which under the present model is equal to the *method effect* or $g^2_j$. This model can be represented by the path diagram in Fig. 4. In the following, we consider some approaches for estimating these parameters under SRS.

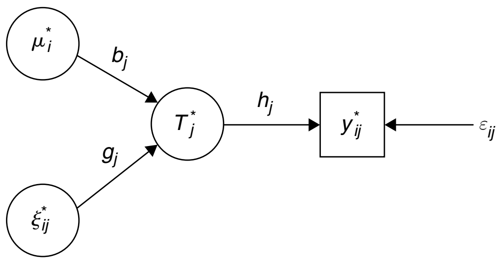


Fig. 4. Basic measurement model for one measurement.

### 6.2. *Estimation of reliability and validity*

#### 6.2.1. *Cross-sectional surveys*

If two measurements are made on the same sample of individuals using different methods, then it can be shown that, under the present model,

$$E(2s_2^2) = \text{SRV}_1 + \text{SRV}_2, \tag{40}$$

where $s_2^2$ is defined in Section 2 and $\text{SRV}_j$ denotes the SRV for $j$th method, $j = 1, 2$. In general, $s_2^2$ is a biased estimator of $\text{SRV}_1$ or $\text{SRV}_2$ unless the two measurements are parallel, in which case $\text{SRV}_1 = \text{SRV}_2$. For parallel measurements, an estimator of reliability can also be obtained from the sample covariance between the two measurements, since from (38) we have (assuming standardized variables but dropping the asterisks for simplicity).

$$\text{Cov}(y_{i1}, y_{i2}) = h_1 h_2 = h^2. \tag{41}$$

When measurements are not parallel, three measurements having the same true score are required for estimating the $h_j^2$, $j = 1, 2, 3$. To see this, note that the three sample covariances between the measurements can be used to form estimates of $h_1h_2$,$h_1h_3$, and $h_2h_3$. These three estimates can be manipulated to obtain estimates of reliability for each measurement, for example, an estimate of $h_1^2$ can be derived from the following relation:

$$\frac{\text{Cov}(y_{i1}, y_{i2})}{\text{Cov}(y_{i2}, y_{i3})}\text{Cov}(y_{i1}, y_{i3}) = \frac{h_1h_2}{h_2h_3}h_1h_3 = h_1^2. \tag{42}$$

The reliability of the other two measurements can be estimated in similar fashion. No estimates of validity and method coefficients can be obtained with this design, however, even if more than three replications (say, $J \geq 3$) of the same true score are available.

The general measurement model for $J$ repeated measurements can be rewritten in matrix form for the $i$th case as

$$\mathbf{y}_i = \mathbf{hT}_i + \varepsilon_i, \tag{43}$$

where $\mathbf{y}_i$ is the $1 \times J$ vector of standardized measurements on the $i$th unit, $\varepsilon_i$ is the $1 \times J$ vector of errors associated with the measurements, $\mathbf{T}_i$ is the $1 \times J$ vector of standardized true scores, and $\mathbf{h} = \text{diag}\,\{h_1, \ldots, h_J\}$. The latent variable model (second model in (38)) will be ignored for the time being since this design does not admit estimates of its coefficients. Let $\Sigma$ denote the $J \times J$ variance–covariance matrix of $\mathbf{y}_i$, which is assumed to be the same for all units $i$. Note that $\Sigma$ can be written as $\mathbf{h}^2 + \mathbf{\Theta}^2$, where $\mathbf{\Theta^2}$ is the $J \times J$ diagonal matrix with diagonal elements, $\sigma_{\varepsilon * j}^2 = \text{Var}(\varepsilon_{ij}^*)$, $j = 1, \ldots, J$. Model (43) is essentially a factor analytic model with one factor, and thus, ordinary factor analysis techniques can be used to estimate $\mathbf{h}$ and $\mathbf{\Theta}$. If the vector $\mathbf{y}_i$ is assumed to follow a multivariate normal distribution, maximum likelihood estimation techniques can be used to obtain efficient estimates and their standard errors in large samples. Let $\hat{\mathbf{h}} = \text{diag}\,\{\hat{h}_1, \ldots, \hat{h}_J\}$ denote the MLE of $\mathbf{h}$. Then the maximum likelihood estimator of the reliability of the $j$th method is $\hat{h}_j^2$.

The estimation of all the components in (38), that is, reliability, validity, and method effects, is possible using the so-called *multitrait-multimethod* (MTMM) approach originally proposed by Campbell and Fiske (1959) and extended to the so-called *true score* *model* in (38) by Saris and Andrews (1991). For estimability, the design requires the same characteristic be obtained using two different methods and two replications for each. However, at least *three* methods and two replications—a minimum of six repeated measurements of the same characteristic—is highly recommended. To apply the MTMM approach, the model in (38) is extended to accommodate more than one trait as follows:

$$\begin{aligned} y^*_{ijk} &= h_{jk} T^*_{ijk} + \varepsilon^*_{ijk} \\ T^*_{ijk} &= b_{jk} \mu^*_{ik} + g_{jk} \xi^*_{ijk}, \end{aligned} \tag{44}$$

where $ijk$ refers to the $i$th individual, $j$th method, and $k$th trait. No restrictions are made on the correlations among the true values for the same individuals, that is, $\mathrm{Cov}(\mu^*_{ik}, \mu^*_{ik'}) \neq 0$; however, the usual assumptions regarding uncorrelated errors is made for $\varepsilon^*_{ijk}$ and $\xi^*_{ijk}$ among individuals, traits, and methods.

Over the past 20 years, researchers have done extensive of analysis of the quality of survey measures using this approach (see, e.g., Scherpenzeel and Saris, 1997, for a review and meta-analysis of the literature). Saris et al. (2004) advocate using past MTMM studies to inform meta-analytic models, where the dependent variables are the MTMM reliability or validity estimates of survey questions and independent variables include item characteristics such as mode of interview, response format or scale type, position in the questionnaire, motivating statements, etc. They use these models to predict the quality of survey questions for future or planned surveys. An excellent discussion of the MTMM approach can be found in (Alwin (2007), Chapter 4).

### 6.2.2. Panel surveys

For panel data, it is possible to estimate validity, method variance, and reliability with fewer than six replicate measurements at each time point by taking advantage the temporal replication inherent in the panel design. Many methods are available including an extension of the MTMM method referred to as the multitrait, multimethod, and multitime (MTMMMT) approach (Saris and Andrews, 1991). This design essentially replicates a MTMM data structure at each time point. One advantage of the design is that fewer replicate measures (e.g., two traits and two repeated measurements) are required at each time point. In addition, this design allows the evaluation of an additional, heretofore unmentioned component of variance, viz., *unique* variance. This component, which is confounded with $\varepsilon$-error term in the previous models, can be regarded as the interaction between the method and the trait (see, e.g., Saris and Andrews, 1991).

For estimating reliability in panel surveys without repeated measurements at each time point, the method of choice is the so-called *quasi-simplex approach* (Heise, 1969, 1970; Jöreskog, 1979; Wiley and Wiley, 1970). This method only requires the same characteristic be measured by the same method (i.e., interview mode, survey question, interview approach, etc.) for at least three panel waves. Like the MLCA model, a first-order Markov latent variable model is assumed for the third and subsequent time points. The covariation of responses within and between the waves provides the basis for an estimate of the reliability of the measurement process. In this sense, the quasi-simplex model is akin to a test–retest reliability assessment, where the correlation between values of the same variable measured at two or more time points estimates the reliability of those values. An important difference is that while test–retest reliability assumes no change in true score across repeated measurements, the quasi-simplex model allows the true scores to change across the repetitions. The quasi-simplex model attempts to model the changing true scores while simultaneously separating the true score variance from error variance. The quasi-simplex path diagram for three time points is very similar to Fig. 3 without the grouping variable, $G$. In fact, it can be regarded as the continuous data version of the MLCA.

Only the model for three waves of data will be illustrated; however, extensions to four or more waves is straightforward. Let $w(=1,2,3)$ denote a wave of a panel survey and, for convenience, drop the subscript $i$ denoting the unit. Assume an observation at wave $w$, say $y_w$, is related to its true score, $T_w$, through the model

$$y_w = T_w + \varepsilon_w, \tag{45}$$

where $\varepsilon_w$ denotes the random measurement error. We further assume that true scores across waves are inter-related as follows:

$$\begin{aligned} T_2 &= \beta_{12} T_1 + \zeta_2 \\ T_3 &= \beta_{23} T_2 + \zeta_3, \end{aligned} \tag{46}$$

Where $\beta_{12}$ is the effect of true score at time 1 on true score at time 2 and $\beta_{23}$ is the effect of true score at time 2 on true score at time 3. The terms $\zeta_2$ and $\zeta_3$ are random error terms that represent the deviations between $t_{w+1}$ and $\beta_{w,w+1}t_w$, sometimes referred to as *random shocks*. Note that Var $(\zeta_w)$ is a component of true score variance at time $w$; for example,

$$\text{Var}(T_2) = \beta_{12}^2 \text{Var}(T_1) + \text{Var}(\zeta_2). \tag{47}$$

Assumptions of the quasi-simplex model include, for all $w$, $w'$ = 1,2,3,

$$E(\varepsilon_w) = 0,\ \text{Cov}(\varepsilon_w, \varepsilon_{w'}) = 0,\ w \neq w',\ \text{Cov}(\varepsilon_w, T_{w'}) = 0, \\ \text{Cov}(\zeta_w, T_{w'}) = 0. \tag{48}$$

For model identification, we impose the further restriction of equal error variances, that is,

$$\text{Var}(\varepsilon_w) = \text{Var}(\varepsilon_{w'}) = \sigma_\varepsilon^2 \quad \text{for } w \neq w'. \tag{49}$$

Under normality assumptions, maximum likelihood estimation can be used to estimate the parameters $\beta_{12}$, $\beta_{23}$, $\sigma_\varepsilon^2$, $\sigma_{t1}^2 = \text{Var}(t_1)$, $\sigma_{\zeta 2}^2 = \text{Var}(\zeta_2)$, and $\sigma_{\zeta 3}^2 = \text{Var}(\zeta_3)$. The reliabilities for the three waves are given by the following:

$$\begin{aligned} R_1 &= \frac{\sigma_{t1}^2}{\sigma_{t1}^2 + \sigma_\varepsilon^2},\ R_2 = \frac{\sigma_{t2}^2}{\sigma_{t2}^2 + \sigma_\varepsilon^2} = \frac{\beta_{12}^2 \sigma_{t1}^2 + \sigma_{\zeta 2}^2}{\beta_{12}^2 \sigma_{t1}^2 + \sigma_{\zeta 2}^2 + \sigma_\varepsilon^2}, \\ R_3 &= \frac{\sigma_{t3}^2}{\sigma_{t3}^2 + \sigma_\varepsilon^2} = \frac{\beta_{23}^2(\beta_{12}^2 \sigma_{t1}^2 + \sigma_{\zeta 2}^2) + \sigma_{\zeta 3}^2}{\beta_{23}^2(\beta_{12}^2 \sigma_{t1}^2 + \sigma_{\zeta 2}^2) + \sigma_{\zeta 3}^2 + \sigma_\varepsilon^2}. \end{aligned} \tag{50}$$

Thus, MLEs of the reliabilities can be obtained by replacing the parameters by their respective MLEs.

A key assumption of the classical quasi-simplex model is the assumption of constant error variance across waves, whereas true score variance is allowed to vary with the wave. However, there are situations when the error variance should also be allowed to vary across the waves. For example, as respondents repeatedly participate in a survey, they may become better respondents and their responses may become less subject to random error. Thus, the error variance at $w = 3$ could be somewhat smaller than the error variance at $w = 2$, which is smaller still than the error variance at $w = 1$.

Unfortunately, specifying both varying true score and error variances will yield an unidentified model due to insufficient number of degrees of freedom to obtain a unique solution to the structural equations. Thus, if the assumption of changing error variances is specified, then true score variances must be held constant across waves. There are no identifiability issues if both true score and error variances are held constant or by restricting the reliabilities to be equal across waves. Biemer et al. (2006) examine this issue for a wide range of composite measures from the National Survey of Child and Adolescent Well-being (NSCAW). They concluded that, for most of the measures, the assumption of unequal error score variances allowing true score variance to vary gave better results using the quasi-simplex model.

## 7. Discussion

Traditional measurement error models made fairly strong assumptions on the error distributions. As an example, the assumption of parallel measurements is difficult, if not impossible, to satisfy for actual survey operations. Estimating measurement bias required error-free measurements obtained from an infallible and unimpeachable source. Through the years, measurement error modeling techniques have evolved under more operationally plausible assumptions that place less stringent demands on evaluation studies. The LCA has an advantage over traditional methods in that it can be used when the assumptions associated with traditional analysis fail or when the remeasurements are collected by methods that do not satisfy traditional assumptions.

But LCA also requires assumptions which can be difficult to satisfy in many situations. Therefore, the use of LCA for evaluating survey error should not preclude the use of classical methods. We recommend using multiple methods and comparing their results since, as some examples in this chapter have demonstrated, insights are often provided about both the error and the validity of the competing methods. Agreement of the results from multiple methods should engender confidence that the findings of the analysis are valid. Disagreement among the results may lead to further investigation of the underlying assumptions of all the methods. In this way, much more knowledge can be discovered about the underlying causes of the errors than if only one method was used.

There is much more work to be done in the area of measurement error modeling. The limitations of LCMs are not well understood despite their use in various types of analysis for more than half a century. The field requires a better understanding of the consequences of model assumption violations for classification error evaluations. The few simulation studies that have been conducted provide valuable insights regarding the sensitivity of LCA estimates to model failures. Future research will continue to add to this knowledge base.

*Sample Surveys: Design, Methods and Applications, Vol. 29A*
ISSN: 0169-7161
© 2009 Published by Elsevier B.V.
DOI: 10.1016/S0169-7161(08)00013-8

13

# Computer Software for Sample Surveys

*Jelke Bethlehem*

## 1. Survey process

Our world faces a growing demand for information on all kinds of topics. Such information can be collected and compiled in a survey. Executing a survey and producing reliable information can be an expensive and time-consuming process. Fortunately, rapid developments in computer technology have made it possible to conduct more surveys and more complex surveys. Over time, computer hardware and software are being used in more and more steps of the survey process.

This chapter describes the various steps in the process, and the software that can be used in it. Also, attention is paid to some of the methodology problems one may encounter. No attempt has been made to give an exhaustive overview of all available software. Only examples of software are mentioned that could or could not be used in specific situations.

The first step is the *design of the survey*. The survey researcher must define the population to be investigated, the data to be collected, and the characteristics to be estimated. Also, a questionnaire must be designed and tested. Usually, only a sample of the population is investigated. This means a sample design and accompanying estimation procedures must be selected. And the sample must be selected accordingly from an appropriate sampling frame.

The second step in the process is *data collection*. Traditionally, in many surveys, paper questionnaires were used. There were three modes of data collection: face-to-face, by telephone, and by mail. Every mode had its advantages and disadvantages. Since the 1970s, paper questionnaire forms were gradually replaced by electronic forms. This is called *computer-assisted interviewing* (CAI). A computer programme asks the questions, checks the answers, and controls the route through the questionnaire. More recent is the use of the Internet for data collection.

Particularly, if the data are collected by means of paper forms, completed questionnaires have to undergo extensive treatment. To produce high-quality statistics, it is vital to remove any errors. This step is called *statistical data editing*.

Detected errors have to be corrected, but this can be very difficult if it has to be done afterwards, when the fieldwork has been completed. In many cases, particularly for household surveys, respondents cannot be contacted again, so other ways have to be found to solve the problem. Sometimes it is possible to determine a reasonable approximation of a correct value by means of an *imputation technique*, but in other cases an incorrect value may have to be replaced by a special code indicating the value is "unknown."

After data editing, the result is a "clean" file, that is, a file in which no more errors can be found. However, this file is not yet ready for tabulation and analysis. In the first place, the sample is sometimes selected with unequal probabilities, for example, establishments are selected with probabilities proportional to their size. The reason is that a clever choice of selection probabilities makes it possible to produce more accurate estimates of population parameters, but only in combination with an estimation procedure that corrects for this inequality. In the second place, representativeness may be affected by *nonresponse*, that is, for some elements in the sample, the required information is not obtained. If nonrespondents behave differently with respect to the population characteristics to be investigated, estimates will be biased.

To correct for unequal selection probabilities and nonresponse, a *weighting adjustment* procedure is often carried out. Every record is assigned some weight. These weights are computed in such a way that the weighted sample distribution of characteristics like sex, age, marital status, and area reflects the known distribution of these characteristics in the population.

In the case of item nonresponse, that is, answers are missing from some, but not all questions, an *imputation procedure* can also be carried out. Using some kind of model, an estimate for a missing value is computed and substituted in the record.

Finally, a clean file is obtained which is ready for *analysis*. The first step in the analysis phase will nearly always be tabulation of the basic characteristics. A more in-depth analysis should reveal underlying structures and patterns, and thus help gain insight in the subject-matter under research.

The results of the analysis will be *published* in some kind of report. Survey agencies experience an increasing demand for releasing survey data files, that is, data sets containing the individual scores on a number of variables for each respondent. An increasing public consciousness concerning the privacy of individuals may lead to a disclosure problem.

## 2. Data collection

### *2.1. Traditional data collection*

Traditionally, paper questionnaires were used to collect survey data. A questionnaire was defined, containing the questions to be asked of respondents. There were three modes of data collection:

- *Face-to-face interviewing*. Interviewers visit respondents, ask questions, and fill in answers on the questionnaire form. The quality of the collected data tends to be good. However, face-to-face interviewing is expensive. It requires a large number of interviewers, all of whom have to do a lot of traveling.
- *Telephone interviewing*. Interviewers call respondents from the survey agency. No more traveling is necessary. Still, telephone interviewing is not always feasible:

only people who have a telephone can be contacted, and the questionnaire cannot be too long or too complicated.

- *Self-interviewing*. This is a mail survey. No interviewers at all are necessary. Questionnaires are mailed to potential respondents with the request to return completed forms. Although reminders can be sent, the persuasive power of the interviewer is lacking, and therefore response tends to be lower in this type of survey.

### *2.2. CAI*

Carrying out a survey is a complex, costly, and time-consuming process. One of the problems is that data collected by means of paper forms usually contain many errors. Extensive data editing is required to obtain data of acceptable quality. The rapid developments of information technology in the last decades made it possible to use microcomputers for CAI. The paper questionnaire was replaced by a computer program containing the questions to be asked. The computer took control of the interviewing process, and it also checked answers to questions on the spot. Application of computer-assisted data collection has three major advantages:

- It simplifies the work of interviewers. They no longer have to pay attention to choosing the correct route through the questionnaire. Therefore, they can concentrate on asking questions and assisting respondents in getting the answers.
- It improves the quality of the collected data because answers can be checked and corrected during the interview. This is more effective than having to do it afterwards in the survey agency.
- Data are entered in the computer during the interview resulting in a clean record, so no more subsequent data entry and data editing is necessary. This considerably reduces the time needed to process the survey data, and thus improves the timeliness of the survey results.

CAI comes in three modes. The first mode implemented was *computer-assisted telephone interviewing*. In the 80s, the laptop computers arrived, and it became possible to implement *computer-assisted personal interviewing*, which is the electronic form of face-to-face interviewing. Another recent mode of CAI is *computer-assisted self-interviewing* (CASI), which is the electronic analog of mail interviewing. Diskettes are sent to respondents, or they can access the interviewing software via telephone and modem, or via the Internet. This latter form of CASI is also denoted by *computer-assisted web interviewing*.

### *2.3. Authoring languages*

The elements of paper questionnaires were rather straightforward: questions for respondents and instructions for interviewers or the respondents to jump to other questionnaires or the end of the questionnaire. Electronic questionnaires can have more elements, for example:

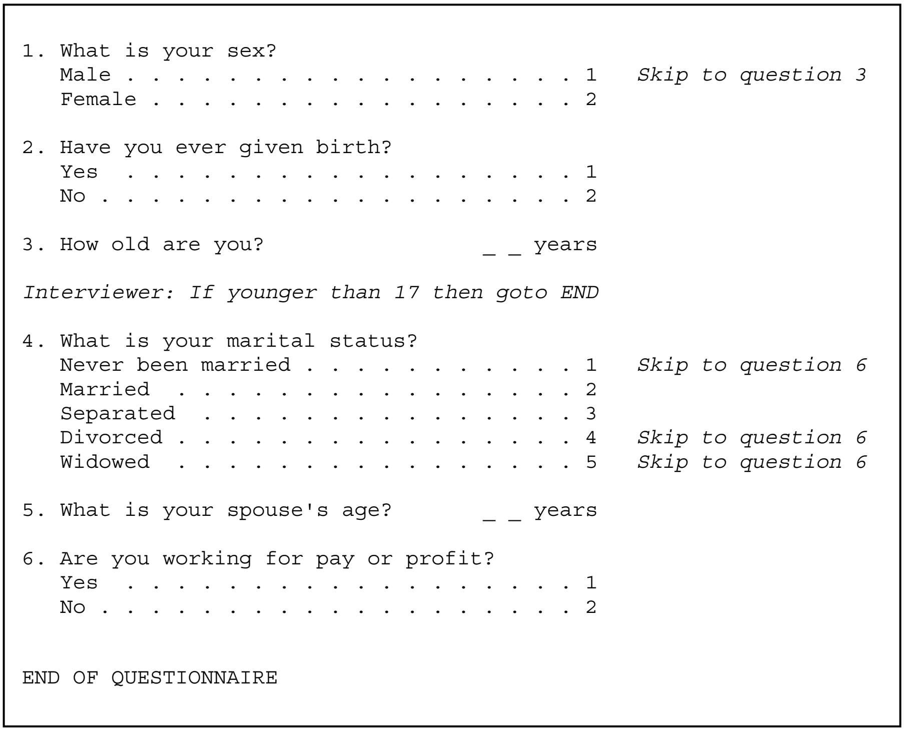
- *Questions*. Each question may have an identification (number or name), a question text, a specification of the type of answer that is expected (text, number, selection from a list, etc.), and a field in which the answer is stored.

- *Checks*. Each check may have identification, a logical expression describing a condition that must be fulfilled, and an error message (which is displayed when the condition is not met).
- *Computations*. Each computation may have identification, an arithmetic expression, and a field in which the result must be stored.
- *Route instructions*. These instructions describe the order in which the objects are processed and also under which conditions they are processed.

These routing instructions can take several forms. This is illustrated using a simple example of a fragment of a questionnaire. Figure 1 shows what this fragment could look like in paper form.

The questionnaire contains two types of routing instructions. In the first place, there are skip instructions attached to answer codes of closed questions. This is the case for questions 1 and 4. The condition deciding the next question asked only depends on the answer to the current question. In the second place, there are instructions for the interviewer that are included in the questionnaire between questions. These instructions are typically used when the condition deciding the next question depends on the answer to several questions, or on the answer to a question that is not the current question. Figure 1 contains an example of such an instruction between questions 3 and 4.

Usually, specification languages of CAI systems (so-called authoring languages) do not contain interviewer instructions. Skip instructions appear in different formats.



```
1. What is your sex?
   Male . . . . . . . . . . . . . . . . . . 1   Skip to question 3
   Female . . . . . . . . . . . . . . . . . 2

2. Have you ever given birth?
   Yes  . . . . . . . . . . . . . . . . . . 1
   No . . . . . . . . . . . . . . . . . . . 2

3. How old are you?                   _ _ years

Interviewer: If younger than 17 then goto END

4. What is your marital status?
   Never been married . . . . . . . . . . . 1   Skip to question 6
   Married  . . . . . . . . . . . . . . . . 2
   Separated  . . . . . . . . . . . . . . . 3
   Divorced . . . . . . . . . . . . . . . . 4   Skip to question 6
   Widowed  . . . . . . . . . . . . . . . . 5   Skip to question 6

5. What is your spouse's age?         _ _ years

6. Are you working for pay or profit?
   Yes  . . . . . . . . . . . . . . . . . . 1
   No . . . . . . . . . . . . . . . . . . . 2


END OF QUESTIONNAIRE
```

Fig. 1. A paper questionnaire.

```
>Sex<
What is your sex?
<1> Male                   [goto Age]
<2> Female
@

>Birth<
Have you ever given birth?
    <1> Yes
    <2> No
    @

>Age<
How old are you ?
<12-20>
@
[@][if Age lt <16> goto End]

>MarStat<
What is your marital status?
<1> Never been married    [goto Work]
    <2> Married
    <3> Separated
    <4> Divorced              [goto Work]
    <5> Widowed               [goto Work]
@

>Spouse<
What is your spouse's age?
<16-20>
@

>Work<
Are you working for pay or profit?
    <1> Yes
    <2> No
    @
```

Fig. 2. The sample questionnaire in *CASES*.

Figure 2 contains a specification of the sample questionnaire of Fig. 1 in the authoring language of the *CASES* system. This system was developed by the University of California in Berkeley. Routing instructions are goto-oriented in *CASES*. There are two types:

(1) Skips attached to answer codes are called *unconditional goto's*
(2) Interviewer instructions are translated into *conditional goto's*.

An example of a CAI system with a different authoring language is the *Blaise System* developed by Statistics Netherlands. The authoring language of this system uses IF-THEN-ELSE structures to specify routing instructions. Figure 3 contains the Blaise code for the sample questionnaire.

There has been an intensive debate about the use of goto-instructions in programming languages. A short paper by Edsger Dijkstra in 1968 ("Go To Statement Considered Harmful") was the start of the structured programming movement. It has become clear

```
DATAMODEL Example

FIELDS

  Sex     "What is your sex?": (Male, Female)
  Birth   "Have you ever given birth?": (Yes, No)
  Age     "How old are you?: 0..120
  MarStat "What is your marital status?":
          (NeverMar "Never been married",
           Married   "Married",
           Separate  "Separated",
           Divorced  "Divorced",
           Widowed  "Widowed")
  Spouse  "What is your spouse's age?": 0..120
  Work    "Are you working for pay or profit?": (Yes, No)

RULES
  Sex
  IF Sex = Female THEN
     Birth
  ENDIF
  Age
  IF Age >= 17 THEN
     MarStat
     IF MarStat = Married) OR (MarStat = Separate) THEN
        Spouse
     ENDIF
     Work
  ENDIF

ENDMODEL
```

Fig. 3. The sample questionnaire in Blaise.

that this also applies to questionnaires. Use of goto-instructions in questionnaires makes these instruments very hard to test and to document.

The way in which the routing structure is specified is not the only difference between Figs. 2 and 3. The developers of Blaise have considered a clear view on the routing structure so important that routing is specified in a separate section of the specification (the rules section).

Note that in the simple example in Fig. 3 only question elements have been used. It contains no checks or computations.

### 2.4. *Modular questionnaires*

Several CAI software systems offer a modular way of specifying electronic questionnaires. This means the questionnaire is split into a number of subquestionnaires, each with its own question definitions and routing structure. Subquestionnaires can be developed and tested separately. It is possible to incorporate such modules as a standard module in several surveys, thereby reducing development time and increasing consistency between surveys.

Also with respect to subquestionnaires, there can be routing instructions. Answers to questions in one subquestionnaire may determine whether or not another subquestionnaire is executed. Furthermore, subquestionnaires can be used to implement hierarchical questionnaires. Such questionnaires allow a subquestionnaire to be executed a number of times. A good example of a hierarchical questionnaire is a household questionnaire. There are questions at the household level, and then there is a set of questions (subquestionnaire) that must be repeated for each eligible member of the household.

On the one hand, a subquestionnaire can be seen as one of the objects in a questionnaire. It is part of the routing structure of the questionnaire, and it can be executed just like a question or a check. On the other hand, a subquestionnaire contains a questionnaire of its own. By zooming into a subquestionnaire, its internal part becomes visible, and that is a questionnaire with its objects and routing conditions.

### *2.5. Testing and documentation*

The growing potential of computer hardware and software has made it possible to develop very large, and very complex electronic questionnaires. It is not uncommon for electronic questionnaires to have thousands of questions. To protect respondents from having to answer all these questions, routing structures and filter questions see to it that only relevant questions are asked and irrelevant questions are skipped. Due to the increasing size and complexity of the electronic questionnaires, it has become more and more difficult for developers, users, and managers to keep control of the content and structure of questionnaires. It takes a substantial amount of knowledge and experience to understand such questionnaires. It has become more and more difficult to comprehend electronic questionnaires in their entirety, and to understand the process that leads to responses for each of the questions as they ultimately appear on data files. See, for example, Kent and Willenborg (1997).

A number of concrete problems have arisen in statistical agencies due to the lack of insight in complex electronic questionnaires:

- It has become very hard to test electronic questionnaires. It is no simple matter to test whether every possible person one might encounter in the field will answer the correct questions in the correct order. Every possible tool providing insight into this matter will help avoid problems in the field.
- Creating textual documentation of an electronic questionnaire has become an enormous task. It is usually a manual task, and is therefore error-prone. There is no guarantee that hand-made documentation exactly describes the real instrument. Making documentation by hand is also very time-consuming.
- There are always managers in organizations who have to approve questionnaire instruments going into the field. In the old days of paper questionnaires, they could base their judgment on the paper questionnaire. However, for modern electronic questionnaire instruments, they have nothing to put their signature on. The printout of the questionnaire specification in the authoring language of the CAI system usually is not very readable for the nonexpert. So, documentation is required that on the one hand is readable and on the other hand describes as exactly as possible what is going in the instrument.
- Interviewers carrying out a survey with paper questionnaires could use the paper questionnaire to get some feeling of where they are in the questionnaire, of what the next questions is about, and of how close they are to the end. If they use an electronic questionnaire, they lack such an overview. Therefore, they often ask for a paper document describing the global content and structure of the questionnaire, which they can use as a tool together with the electronic questionnaire.

All these problems raise the question of the feasibility of a flexible tool capable of representing content and logic of an electronic questionnaire in a human-readable way. Such a tool should not only provide a useful documentation, but also help analyze the questionnaire and report possible sources of problems. Research has shown that there is a need for software capable of displaying the various routes through the questionnaire in the form of a flow chart (see Bethlehem and Hundepool, 2000).

Jabine (1985) described flow charts as a tool to design survey questionnaires. Particularly, flow charts seem to be useful in the early stages of questionnaire development. Sirken (1972) used flow charts to effectively explore alternative structures and sequences for subquestionnaires. He also found that more detailed flow charts, for example, of the structure of subquestionnaires, can be equally effective. Another more recent example

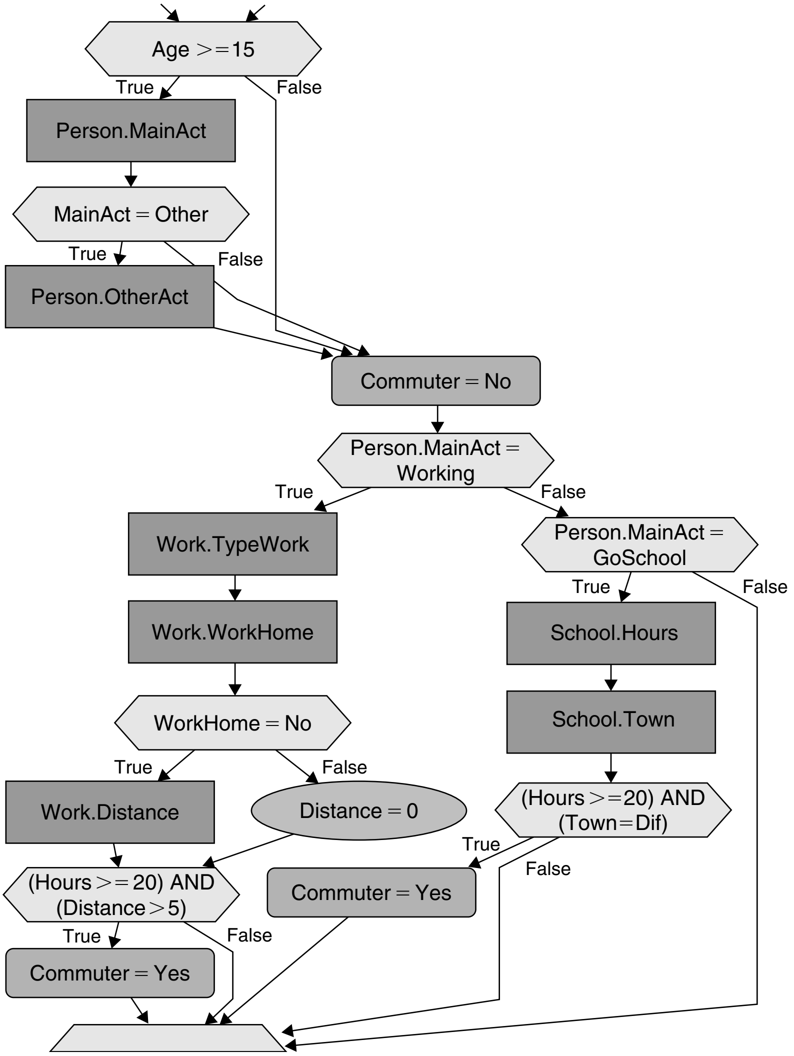


Fig. 4. An example of TADEQ output.

is the *QD* system developed by Katz et al. (1997). A flow chart can also be a useful tool in the documentation of electronic questionnaires. Their strong point is that they can give a clear idea of the routing structure. But they also have the weak point that the amount of textual information that can be displayed about the questionnaire object is limited. Therefore, a flow chart can be a very important component of questionnaire documentation, but it will not be the only component.

A flow chart can also be a useful tool in the documentation of electronic questionnaires. Their advantage is that they can give a clear idea of the routing structure. But they also have the disadvantage that the amount of textual information that can be displayed about the questionnaire object is limited. Therefore, a flow chart can be a very important component of questionnaire documentation, but it will not be the only component. There have been a number of initiatives for automatically producing survey documentation, but they pay little or no attention to documentation of survey data collection instruments. They focus on postsurvey data documentation and not on providing tools to assist in the development and analysis of the operation of the collection instrument. The *TADEQ* project was set up to develop a tool documenting these instruments (see Bethlehem and Hundepool, 2000). Figure 4 shows an example of the output of this tool.

## 3. Statistical data editing

### 3.1. What is statistical data editing?

*Statistical data editing* is the process of detecting errors in survey data and correcting those detected errors, whether those steps take place in the interview or in the survey office after data collection. Traditionally, statistical organizations, especially those in government, have devoted substantial amounts of time and major resources to data editing in the belief that this was a crucial process in the preparation of accurate statistics. Current data editing tools have become so powerful that questions are now raised as to whether too much data editing occurs. A new objective for some is to minimize the amount of data editing performed while guaranteeing a high level of data quality.

Data editing may occur in many phases of the survey process. Edits can be part of a CAI program. In this case, data editing takes place during data collection. Traditionally, data editing has taken place after data collection, either before, after, or during data capture. Editing can also be carried out on tables or graphs of the distribution of one or two variables. Such edits will take place when a substantial part of the data has been completed, for instance 50%. Data editing is not restricted to within record editing. Between record edits and edits on aggregated quantities are also included in the definition.

A more in-depth treatment of statistical data editing is given in Chapter 10 (Statistical Data Editing and Imputation).

### 3.2. Forms of data editing

When data editing takes place at the level of individual records, this is called *micro-editing*. Records are checked and corrected one at a time. Values of the variables in a record are checked without using the values in the other records. Micro-editing typically is an activity that can take place during the interview or during data capture. When data editing takes place at the level of aggregated quantities obtained by using all available records, we call it *macro-editing*. For macro-editing, a file of records is required. This means it is typically an activity that takes place after data collection, data entry, and possibly after micro-editing.

Following Pierzchala (1990), data editing can be seen as addressing four principal types of data errors:

- *Completeness errors.* The first thing to be done when filled-in paper questionnaire forms come back to the survey agency is to determine whether they are complete enough to be processed. Forms that are blank or unreadable, or nearly so, are unusable. They can be treated as cases of unit nonresponse, scheduled for call-back, deleted from the completed sample, or imputed in some way, depending on the importance of the case.
- *Domain errors.* Each question has a domain (or range) of valid answers. An answer outside this domain is considered an error.
- *Consistency errors.* Consistency errors occur when the answers to two or more questions contradict each other. Each question may have an answer in its valid domain, but the combination of answers may be impossible or unacceptable. The occupation of a person may be school teacher, a person may be under 5 years of age but the combination of these answers for the same person is probably an error. A firm known to have 10 employees should not report more than 10,000 person days worked in the past year.
- *Routing errors (skip pattern errors).* Many questionnaires contain routing instructions. A routing error occurs when an interviewer or respondent fails to follow a routing instruction, and a wrong path is taken through the questionnaire. Routing errors are also called *skip pattern errors*. As a result, the wrong questions are answered, leaving applicable questions unanswered and inapplicable items with entries.

### 3.3. Developments

In traditional survey processing, data editing was mainly a manual activity. Domain errors were identified by visually scanning the answers to the questions one at the time. Consistency errors were typically caught only when they involved a small number of questions on the same page or on adjacent pages. Route errors were found by following the route instructions and noting deviations. In general, manual editing could identify only a limited number of the problems in the data.

The data editing process was greatly facilitated by the introduction of computers. Initially, these were mainframe computers, which only permitted batch-wise editing. Tailor-made editing programs, often written in COBOL or FORTRAN, were designed for each survey. Later, general purpose batch editing programs were developed and extensively used in survey agencies. These programs performed extensive checks on each record and generated printed lists of error reports by case identification number. The error lists were then sent to subject-matter experts or clerical staff, who attempted to manually reconcile these errors. This staff then prepared correction forms which were keyed to update the data file, and the process was repeated.

Batch computer editing of data sets improved the data editing process because it permitted a greater number of and more complex error checks. Thus, more data errors could be identified. However, the cycle of batch-wise checking and manual correction was proved to be labor-intensive, time-consuming, and costly. See, for example, Bethlehem (1987) for a more complete analysis of this process and its disadvantages. Garcia and Thompson (2000) also pointed at the disadvantages of manual editing. They showed that a generalized Fellegi–Holt system was able to edit/impute a large economic survey in 24 hours, whereas 10 analysts needed 6 months to make three times as many changes.

With the emergence of microcomputers in the early 1980s, completely new methods of data editing became possible. One of these approaches has been called computer-assisted data input (CADI). The same process also has been called computer-assisted data entry. CADI provides an interactive and intelligent environment for combined data entry and data editing of paper forms by subject-matter specialists or clerical staff. Data can be processed in two ways: either in combination with data entry or as a separate step.

In the first approach, the subject-matter employees process the survey forms with a microcomputer one-by-one. They enter the data "heads up," which means that they tend to watch the computer screen as they make entries. After completion of entry for a form, they activate the check options to test for all kinds of errors (omission, domain, consistency, and check errors). Detected errors are displayed and explained on the screen. Staff can then correct the errors by consulting the form or by contacting the supplier of the information. After elimination of all visible errors, a "clean" record, that is, one that satisfies all check edit criteria, is written to file. If staff members do not succeed in producing a clean record, they can write it to a separate file of problem records. Specialists can later deal with these difficult cases using the same CADI system. This approach of combining capture and editing is efficient for surveys with relatively small samples but complex questionnaires.

In the second approach, clerical staff (data typists or entry specialists) enters data through the CADI system "heads down," that is without much error checking. When this entry step is complete, the CADI system checks all the records in a batch run and flags the cases with errors. Then, subject-matter specialists take over. They examine the flagged records and fields one-by-one on the computer screen and try to reconcile the detected errors. This approach works best for surveys with large samples and simple questionnaires.

A second advance in data editing occurred with the development of CAI. It replaced the paper questionnaire with a computer program that was in control of the interviewing process. Routing and range errors are largely eliminated during data entry. This also reduces the burden on the interviewers, since they need not worry about routing from item to item and can concentrate on getting the answers to the questions. It also becomes possible to carry out consistency checking during the interview. Since both the interviewer and the respondent are available when data inconsistencies are detected, they can immediately reconcile them. In this way, CAI should produce more consistent and accurate data than correcting errors in the survey office after the interview is over.

Performing data editing during a computer-assisted interview is greatly facilitated when the interviewing software allows specification of powerful checks in an easy and user-friendly way. Although edit checks can be hard-coded for each survey in standard programming languages, this is a costly, time-consuming, and error-prone task. Many CAI software packages now offer very powerful tools for micro-editing, permitting easy specification of a large number of checks, including those involving complex relationships among many questions. Editing during CAI is now extensively used both in government and private sector surveys. An example of CAI software used by many statistical institutes is the Blaise System, see Statistics Netherlands (2002).

Whether micro-editing is carried out during or after the interview, the entire process may have major disadvantages, especially when carried to extremes. Little and Smith (1987) and Granquist and Kovar (1997) have mentioned the risk of *over-editing*. Powerful editing software offers ample means for almost any check one can think of, and it is sometimes assumed that the more checks one carries out, the more errors one will correct. But there are risks and costs.

First, the use of too many checks may cause problems in interviewing or postinterview data correction, especially if the checks are not carefully designed and thoroughly tested prior to use. Contradictory checks may cause virtually all records to be rejected, defeating the purpose of editing. Redundant checks may produce duplicate or superfluous error messages slowing the work. And checks for data errors that have little impact on the quality of published estimates may generate work that does not contribute to the quality of the finished product.

Second, since data editing activities make up a large part of the total survey costs, their cost effectiveness has to be carefully evaluated at a time in which many statistical agencies face budget reductions. Large numbers of micro-edits that require individual correction will increase the costs of a survey. Every attempt should be made to minimize data editing activities that do not improve the quality of the survey results.

Third, it must be recognized that not all data problems can be detected and repaired with micro-editing. One such problem is that of outliers. An *outlier* is a value of a variable that is within the domain of valid answers to a question, but it is highly unusual or improbable when compared with the distribution of all valid values. An outlier can be detected only if the distribution of all values is available.

Three alternative approaches to editing are described that address some of the limitations of traditional micro-editing. In some situations, they could replace micro-editing. In other situations, they could be carried out in combination with traditional micro-editing or with each other. They are called *automatic editing*, *selective editing*, and *macro-editing*.

### *3.4. Automatic editing*

In *automatic editing*, checking and correcting the records are carried out automatically by a software package. Since no human activities are involved, this approach is fast and cheap. For automatic editing, the usual two stages of editing, error detection and error correction, are expanded into three stages as follows:

- *Error localization*. As usual, the software detects errors or inconsistencies by reviewing each case using the prespecified edit rules.
- *Determining the cause of the error*. If an edit detects an error that involves several variables, the system must determine which variable caused the error. Several strategies have been developed and implemented to solve this problem.
- *Error correction*. Once the variable causing the error has been identified, its value must be changed so that the new value no longer causes the error message.

There is no straightforward way to determine which of several variables causes a consistency error. One obvious criterion is the number of inconsistencies in which that variable is involved. If variable $A$ is related to three other variables $B$, $C$, and $D$, an erroneous value of $A$ may generate three inconsistencies: with $B$, $C$, and $D$. If $B$, $C$, and $D$ are involved in no other edit failures, $A$ seems the likely culprit. However, it could be that no other edit rules have been specified for $B$, $C$, and $D$. Then $B$, $C$, and $D$ could also be candidates for correction.

The *Fellegi–Holt methodology* takes a more sophisticated approach (see Fellegi and Holt, 1976). To reduce dependence on the number of checks defined, the Fellegi–Holt methodology performs an analysis of the pertinent edit checks for each variable. Logically, superfluous checks are removed and all implied checks that can be logically derived from the checks in question are added. Records are then processed as a whole, and not on a field-by-field basis, with all consistency checks in place to avoid the introduction of new errors as identified ones are resolved. The smallest possible set of imputable fields is located, with which a record can be made consistent with all checks.

In the Fellegi–Holt methodology, erroneous values are often corrected with hot-deck imputation. Hot-deck imputation uses values copied from a similar donor record (another case) not violating any edit checks. When the definition of "similar" is very strict or when the receptor record is unique, it may be impossible to find a similar donor record. In this situation, a simple default imputation procedure is applied instead.

The Fellegi–Holt methodology has been programmed and put into practice for government statistical agencies. Several editing packages exist. For editing categorical variables, they include *DISCRETE* (United States), *AERO* (Hungary), *DIA* and *LINCE* (Spain), and *DAISY* (Italy). The U. S. Census Bureau program *SPEER* was designed for valid value checks on ratios of numerical variables. The Chernikova algorithm, which can only handle numerical variables, was further developed by Statistics Canada for the generalized editing and imputation programs *GEIS* and *Banff*. All these programs identify fields that are likely to contain errors and impute estimates of their values. These programs run on diverse computer platforms and operating systems. *AERO* runs under MS-DOS, *LINCE* under Windows, and *DAISY* runs on an IBM-mainframe. *DISCRETE* and *SPEER* are in portable FORTRAN code. *GEIS* is an Oracle-based system, whereas *Banff* is a SAS-based system.

The current state of affairs of automatic editing only allows for limited applicability of these techniques. This is disappointing because powerful automatic data editing tools can substantially reduce survey costs. Currently, automatic editing should only be used for detecting and correcting errors that have no substantial impact on the published statistics. Furthermore, automatic editing should never be the only data editing activity. To avoid imputation of values of wrong variables, it should be used in combination with other editing techniques.

### *3.5. Selective editing*

Instead of conserving editing resources by fully automating the process, they may be conserved by focusing the process on the most necessary edits. Necessary edits are those which have a noticeable effect on published figures, including outliers. This approach is called *selective editing*.

To establish the effect of edits on population estimates, one can compare estimates based on unedited data with estimates based on edited data. Boucher (1991) and Lindell (1994) compared unedited data with edited data and found that, for each variable studied, 50–80% of the edits had virtually no effect on the estimate of the grand total. Similar results were obtained in an investigation carried out by Van de Pol and Molenaar (1995) on the effects of editing on the Dutch Annual Construction Survey.

If only a few edits have a substantial impact on the final figures, data editing efforts can be reduced by identifying those edits. One way to implement this approach is to use a criterion to split records into a critical and noncritical stream. The *critical stream* contains the records which have a high risk of containing influential errors and therefore requires thorough micro-editing. Records in the *noncritical stream* could remain unedited or could be limited to automatic editing.

At present, there is no standard software available to implement the concepts of selective editing just described. In some situations, selective editing can be applied using existing software. An example is the *Blaise System* developed by Statistics Netherlands (2002). First, the data entry program is used for heads-down data entry. Then, the data manipulation tool (*Manipula*) is used to compute the values of a so-called OK index, and these values are added to the records of each case. One approach could be to split the data file into a file of critical records (i.e., records with an OK index below a specified threshold) and a noncritical file. The critical file will then be subject to micro-editing. Another approach could be to sort the data file by OK index value from low to high, and then allow analysts to continue working with the file. They can decide on a case-by-case basis how far they need to continue editing the records in the file.

Selective editing is a promising approach to data editing. However, methodology is still in its infancy. Selective editing has been shown to work in specific cases, but a general framework is needed to provide more tools for deciding which records must undergo micro-editing.

### *3.6. Macro-editing*

*Macro-editing* provides a solution to some of the data problems left unsolved by micro-editing. It can also address data problems at the aggregate, distribution, and higher levels. The types of edit checks used by macro-editing are similar to those of micro-editing, but the difference is that macro-edit checks involve aggregated quantities. In this chapter, two general approaches of macro-editing will be described.

The first approach is sometimes called the *aggregation method* (see Granquist, 1990; United Nations, 1994). It formalizes and systematizes what statistical agencies routinely do before publishing statistical tables. They compare the current figures with those of previous periods to see if they appear plausible. Only when an unusual value is observed at the aggregate level will the individual records contributing to the unusual quantity be edited at the micro level. The advantage of this form of editing is that it concentrates on editing activities at those points that have an impact on the final results of the survey. No superfluous micro-editing activities are carried out on records that do not produce unusual values at the aggregate level. A disadvantage is that results are bent in the direction of one’s expectations. There is also a risk that undetected errors may introduce undetected biases.

A second approach of macro-editing is the *distribution method*. The available data are formed into distributions of variables, and the individual values are compared with their distributions. Measures of location, spread, and covariation are computed. Records containing values that appear unusual or atypical in their distributions are candidates for further inspection and possible editing. Standard software implementing this macro-editing approach is limited. Three systems mentioned in the literature are described in this chapter.

The first program is *GRED*, a microcomputer program developed by Statistics New Zealand (see Houston and Bruce, 1993). It displays the individual values of a variable for different firms for consecutive survey years. In one plot, it is possible to detect outliers and deviations from trend. For unusual points, the sampling weight of the record can be adjusted or the record can be removed. Outliers can be highlighted by marking them with a different color. Using linked plots, these outliers will also be highlighted in other graphic displays. This permits easy identification of the influence of a specific observation on aggregate statistics.

The second program, *ARIES* was designed for macro-editing the Current Employment Statistics Program of the U. S. Bureau of Labor Statistics (see Esposito and Lin, 1993; Esposito et al., 1994). A session with *ARIES* starts with a so-called anomaly plot. This is a graphical overview of the important estimates in which each node represents a specific industry. Related estimates are connected by lines. Estimates identified as unusual based on month-to-month changes are marked in a different color. Only suspicious estimates are analyzed in more detail.

For industry groups, *ARIES* can generate two types of plots: a scatter plot of the data values of the current month against the data values of the previous month, and a plot of the distribution of the month-to-month changes. By selecting points using a mouse, the data values can be displayed in tabular form on the screen. The adjustment weight of detected outliers can be modified interactively. Future versions of *ARIES* will produce simultaneous scatter plots of multiple variables. Linked plots will help the analyst to study outlier cases from one plot in other plots for the same or additional variables.

A third example of a macro-editing system was developed for use in the Swedish Short Periodic Employment Survey (see Engström and Ängsved, 1994). First, suspect estimates are selected, based on time-series analysis and sample variance. This may be seen as a form of macro-editing by the aggregation method. Next, a scatter plot is made of the data contributing to each suspect estimate, with its data values plotted against the corresponding values of the previous quarter. Outliers are displayed in different colors. Single-clicking on an observation displays information about it, including the weight assigned to the observation and a measure of its contribution to the estimate. Double-clicking on an observation provides the analyst with access to the corresponding data record. The analyst can make changes in the data record. The system then rechecks the record for inconsistencies and updates the scatter plot and all corresponding parameters.

Most current macro-editing systems have been designed for application to specific surveys. Thus, they cannot be directly applied to surveys with different variables and data structure. There is a clear need for general macro-editing software that can be used for a wide variety of surveys.

## 4. Imputation

### *4.1. What is imputation?*

To correct for item nonresponse and for errors in the data, often some kind of imputation technique is applied. *Imputation* means that missing values are replaced by synthetic values. A synthetic value is obtained as the result of some technique that attempts to estimate a missing value. Imputation uses the available information about the specific element and possibly also other available information.

After applying an imputation technique, there are no more "holes" in the survey data set. So, all analysis techniques can be applied without having to worry about missing values. However, there is a downside to this approach. There is no guarantee that an imputation technique will reduce a bias caused by item nonresponse. It depends on the type of missing data pattern and the specific imputation technique that is applied. Three types of missing data mechanisms are distinguished. Let $X$ represent a set of auxiliary variables that are completely observed, $Y$ a target variable that is partly missing, $Z$ represents causes of missingness unrelated to $X$ and $Y$, and $R$ the missingness.

In case of *Missing Completely at Random* (MCAR), missingness is caused by a phenomenon $Z$ that is completely unrelated to $X$ and $Y$. Estimates for parameters involving $Y$ will not be biased. Imputation techniques will not change this.

In case of *Missing at Random* (MAR), missingness is caused partly by an independent phenomenon $Z$ and partly by the auxiliary variable $X$. So, there is an indirect relationship between $Y$ and $R$. This leads to biased estimates for $Y$. Fortunately, it is possible to correct for such a bias by using an imputation technique that takes advantage of the availability of all values of $X$, both for respondents and nonrespondents.

In case of *Not Missing at Random* (NMAR), there may be a relationship between $Z$ and $R$ and between $X$ and $R$, but there is also a direct relationship between $Y$ and $R$ that cannot be accounted for by $X$. This situation also leads to biased estimates for $Y$. Unfortunately, imputation techniques using $X$ will not be able to remove the bias.

There are many imputation techniques available. A number of them are described in the next two sections.

### *4.2. Single imputation*

*Single imputation* means that a missing value is replaced by a synthetic value. By contrast, Section 4.3 is about multiple imputation, where a missing value is replaced by a set of synthetic values. There are many single imputation techniques described in the literature (see, e.g., Kalton and Kasprzyk, 1986). A nonexhaustive overview of the most frequently used techniques is given below.

Sometimes the value of a missing item can be logically deduced with certainty from the nonmissing values of other variables. This is called *deductive imputation*. If strict rules of logic are followed, this technique has no impact on the properties of the distribution of estimators. For example, if a girl is 5-years-old, it is certain that she has had no children. Likewise, if a total is missing but all subtotals are available, the total can easily be computed.

*Imputation by the mean* implies that a missing value of a variable is replaced by the mean of the available values of this variable. Since all imputed values are equal to the sample mean, the distribution of this variable in the completed data set will be affected. It will have a peak at the mean of the distribution. For *imputation by the mean* within groups, the sample is divided into a number of nonoverlapping groups. Within a group, a missing value is replaced by the mean of the available observations in that group. Imputation by the mean within groups will perform better than imputation by the mean if the groups are homogeneous with respect to the variable being imputed. Then, all values are close to each other, and therefore, the imputed group mean will be a good approximation of the true, but unknown, value.

*Random imputation* means that a missing value is replaced by a value that is randomly chosen from the available values for the variable. This imputation is sometimes also called hot-deck imputation. It is a form of donor imputation: a value is taken from an existing record where the value is not missing. The distribution of the values of the variable for the complete data set will look rather natural. However, this distribution does not necessarily resemble the true distribution of the variable. Both distributions may differ if the missing values are not randomly missing. For *random imputation within groups*, the sample is divided into a number of nonoverlapping groups on the basis of one or more qualitative auxiliary variables. Within a group, a missing value is replaced by a randomly chosen value from the set of available values in that group. Random imputation within groups will perform better than random imputation if the groups are homogeneous with respect to the variable being imputed. Since all values are close to each other, the randomly selected value will be a good approximation of the true, but unknown, value.

The idea of *nearest neighbor imputation* is to search for a record which resembles as much as possible the record in which a value is missing. A distance measure is defined to compare records on the basis of values of auxiliary variables that are available for all records.

*Regression imputation* assumes a linear relationship between the target variable (with missing values) and a set of auxiliary variables (without missing values). A regression model is estimated using the available data. After that, this model can be applied to predict missing values of the target variable. Application of regression imputation affects the distribution of the imputed variables in the data set. Therefore, inference for this variable may lead to wrong conclusions. To avoid this, an error term can be added to imputed values. Such an error term can be derived randomly using the nonmissing cases in the data set or from a theoretical distribution.

A number of single imputation techniques are available in an optional module of SPSS (Missing Value Analysis). IVEware is a free package (developed by the University of Michigan) that runs in SAS. It contains a variety of single imputation techniques. See Raghunathan et al. (2002) for a description. Royston (2004, 2005) described a user supplied ado-file for the statistical package *Stata*. It is called Imputation by Chained Equations (*ICE*) and provides a number of model-based single imputation techniques. On the basis of the level of measurement, one can choose a regression, logit, ordered logit, or multinomial regression model. There is a library of functions that can be used in S-plus to perform various kinds of single imputation. For more information, see, for example, Alzola and Harrell (2006). SOLAS is a dedicated package that implements many imputation techniques. Among supported single imputation techniques are imputation by the mean, random imputation, and regression imputations. For more details, see, for example, Scheffer (2002) and Horton and Lipsitz (2001).

### *4.3. Properties of single imputation*

There are many single imputation techniques. So the question arises: which technique to use in a practical situation. There are several aspects that may play a role in this decision. The first aspect is the type of variable for which missing values have to be imputed. In principle, all mentioned imputation techniques can be applied for quantitative variables. However, not every single imputation technique can be used for qualitative variables. A potential problem is that the synthetic value produced by the imputation technique does not necessarily belong to the domain of valid values of the variable. For example, if the variable gender has to be imputed, mean imputation produces an impossible value (what is the mean gender?). Therefore, it is better to stick to some form of "donor imputation" for qualitative variables. These techniques always produce "real" values.

Single imputation techniques can be divided in two groups: deterministic imputation techniques and random imputation techniques. For *deterministic imputation* techniques, imputed values only depend on the realized sample. An example is imputation by the mean. *Random imputation* techniques add an extra source of randomness to the computation of imputed values. An example is random imputation.

For some deterministic imputation techniques (e.g., imputation by the mean), the mean of a variable before imputation is equal to the mean after imputation. This shows that not every imputation technique is capable of reducing a bias due to missing data. For random imputation techniques, the mean before imputation is never equal to the mean after imputation. However, expected values before and after imputation may be equal.

Deterministic imputation may affect the distribution of a variable. It tends to produce synthetic values that are close to the centre of the original distribution. The imputed distribution is more "peaked." This may have undesirable consequences. Estimates of standard errors may turn out to be too small. A researcher using the imputed data (not knowing that the data set contains imputed values) may get the impression that his estimates are very precise, whereas in reality, this is not the case.

Imputation may also have an impact on the correlation between variables. Suppose the variable $Y$ is imputed using imputation by the mean, and suppose the variable $X$ is completely observed for the sample. It can be shown that, in this case, the correlation after imputation is smaller than the correlation in the data set before imputation. The more the observations are missing from $Y$, the smaller the correlation coefficient will be. Researchers not aware of the fact that their data set has been imputed will get the impression that relationships between variables are weaker than they are in reality. Also here, there is a risk of drawing wrong conclusions.

The correlations are also affected if random imputation is applied. These values will generally be too low when computed using the imputed data set. This is caused by the fact that imputed values are randomly selected without taking into account possibly existing relationships with other variables.

### *4.4. Expectation–Maximization imputation*

Expectation–Maximization imputation (or EM imputation) uses an iterative maximum likelihood procedure to provide estimates of the mean and the variance-covariance matrix based on all available data for each respondent. The algorithm assumes that the data are from a multivariate normal distribution (or possibly some other specified distribution), and that, conditional on the reported data, the missing data are MAR.

EM imputation is a widely used technique that derives likelihood-based inferences from incomplete data (see Little and Rubin, 2002). Actually, it is not a pure imputation technique because no missing values themselves are being substituted. Rather, functions of the missing values (sufficient statistics) appearing in the likelihood function are being substituted.

In each cycle of the iterative process, there is an Expectation step (or E-step) followed by an Maximization step (or M-step). The E-step computes expected values based on all available data. This is followed by an M-step in which missing values are replaced by synthetic values computed in the E-step. This process is continued until convergence.

EM imputation is available in several general analysis packages. For example, in *SPSS* it is part of the module Missing Value Analysis (*MVA*). Other-than normal distributions can be assumed here. EM imputation is also available in *SAS* (from version 8.2), and *S-Plus* (from version 6.0).

### *4.5. Multiple imputation*

Single imputation is a technique that solves the missing data problem by filling the holes in the data set by plausible values. This is clearly an advantage in the analysis phase of the survey. However, there are also disadvantages. Application of a single imputation technique may create more problems than that are solved because the distribution of estimates is distorted. Therefore, there is a risk of drawing wrong conclusions about the data set. More details about this aspect of imputation can be found, for example, in Little and Rubin (2002).

To address the problems caused by single imputation techniques, Rubin (1987) proposed *multiple imputation*. This is a technique in which each missing value is replaced by $m > 1$ synthetic values. Typically $m$ is small, say 3–10. This leads to $m$ complete data sets. Each data set is analyzed using standard analysis techniques. For each data set, an estimate of a population parameter of interest is obtained. The $m$ estimates for a parameter are combined to produce estimates and confidence intervals that incorporate missing-data uncertainty.

Rubin (1987) developed his multiple imputation technique primarily to solve the missing data problem in large public-use sample survey data files and censuses files. With the advent of new computational methods and software for multiple imputation, this technique has become increasingly attractive to other sciences where researchers are confronted by missing data. See also Schafer (1997).

Multiple imputation assumes some kind of model. This model is used to generate synthetic values. The imputation model will often be a regression model. The effects of imputation depend on the missing data mechanism that has generated the missing values. The most convenient situation is MCAR. This means that missingness happens completely at random. In this case, multiple synthetic drawings can be generated by applying random imputation a number of times. It is also possible to use imputation by the mean if the variation is modeled properly. For example, this can be done by adding a random component to the mean which has been drawn from a normal distribution with the proper variance.

MCAR is usually not a very realistic assumption. The next-best assumption is that data are MAR. This means that missing information depends on one or more auxiliary variables, but these variables have been observed completely. Sets of synthetic values are generated by using the regression model to predict the missing values. To give the imputed values of the proper variance, usually a random component is added to the predicted value. This component is drawn from a distribution with the proper variance. The worst case is the situation in which data are NMAR. Then, missing information depends on unobserved variables, and therefore no valid imputation model can be built using the available data. The distribution of the estimators cannot be repaired by applying multiple imputation. There is still a risk of drawing wrong conclusions from an analysis.

Rubin (1987) described how to combine estimates for our multiple data sets into one proper estimate. He claims that the number of imputations per missing value should not exceed 10. Multiple imputation can be a useful tool for handling the problems caused by missing data, but if it is not applied carefully, it is potentially dangerous. If an imputation does not model the missing data mechanism properly, analysis of the imputed data sets can be seriously flawed. This means we should check as much as possible the models we use.

A good overview of multiple imputation software can be found in Horton and Lipsitz (2001). *SAS* has the procedures *MI* and *MIANALYZE*. The first procedure is used to create a number of imputed data sets. After analysis of these data sets, the results are combined with the second procedure for making proper inference. The statistical package *SPSS* has the *MVA* module with imputation based on the EM algorithm. It can produce estimates of means, variances, and covariances. It does not compute standard errors. Multiple imputation is not supported. *Stata* contains *ICE*, see Royston (2004, 2005) for multiple imputation. There is program for creating multiple imputed data sets, and another program for analysis of such a data set. *NORM* is a library of functions available for use in *S-Plus*, see Schafer (1997). They just create multiple imputed data sets. Another such library is *MICE*, see Van Buuren and Oudshoorn (1999). It contains tools for generating multiple imputation and pooling of analysis results. The dedicated package *SOLAS* implements various single and multiple imputation approaches, see Horton and Lipsitz (2001). It also contains some analysis tools for imputed data (descriptive statistics, $t$-test, ANOVA, and regression analysis)

## 5. Weighting adjustment

### *5.1. Nonresponse problem*

Surveys suffer from nonresponse. Nonresponse can be defined as the phenomenon where elements (persons, households, companies) in the selected sample do not provide the requested information or where the information provided is useless. The situation in which all requested information on an element is missing is called *unit nonresponse*. If information is missing on some items only, it is called *item nonresponse*. Item nonresponse can be treated by means of imputation (see Section 4).

This section focuses on the treatment of unit nonresponse. Due to this type of nonresponse, the sample size is smaller than expected. This leads to estimates of population characteristics with larger variance, that is, less accurate, but still valid. This is not a serious problem. It can be taken-care-of by making the initial sample size larger. A far more serious problem caused by nonresponse is that estimates of population characteristics may be biased. This situation occurs if some groups in the population are over- or under-represented due to nonresponse, and these groups behave differently with respect to the characteristics to be investigated.

If nonresponse leads to biased estimates, wrong conclusions are drawn from the survey results. Therefore, it is vital to reduce the amount of nonresponse in the fieldwork as much as possible. Nevertheless, in spite of all these efforts, a substantial amount of nonresponse usually remains. To avoid biased estimates, some kind of correction procedure must be carried out. One of the most important correction techniques for nonresponse is *weighting adjustment*. Every observed object in the survey is assigned a weight that is adjusted for nonresponse, and estimates of population characteristics are obtained by processing weighted observations instead of the observations themselves.

## *5.2. Basics of weighting*

Suppose the objective of a survey is assumed to be estimation of the population mean of a variable $Y$. To that end, a simple random sample of size $n$ is selected without replacement. Let $\pi_k$ be the *first-order inclusion probability* of element $k$, for $k = 1, 2, \ldots, N$, where $N$ is the size of the population.

The sample can be represented by a series of $N$ indicators $t_1, t_2, \ldots, t_N$, where the $k$-th indicator $t_k$ assumes the value 1 if element $k$ is selected in the sample, and otherwise it assumes the value 0. Consequently, $E(t_k) = \pi_k$.

In case of complete response, an unbiased estimator for the population mean is defined by Horvitz and Thompson (1952). It can be written as

$$\bar{y}_{\mathrm{HT}} = \frac{1}{N} \sum_{k=1}^{N} \frac{t_k Y_k}{\pi_k}. \tag{1}$$

In the case of nonresponse, estimates may be biased. A frequently used approach to do something about this bias is to apply *weighting adjustment*. Each observed element $k$ is assigned a weight $w_k$. This weight is the product of the inclusion weight $d_k = 1/\pi_k$ and a correction weight $c_k$. So, the Horvitz–Thompson estimator is replaced by a new estimator

$$\bar{y}_{\mathrm{W}} = \frac{1}{N} \sum_{k=1}^{N} w_k t_k Y_k = \frac{1}{N} \sum_{k=1}^{N} c_k d_k t_k Y_k = \frac{1}{N} \sum_{k=1}^{N} \frac{c_k t_k Y_k}{\pi_k}. \tag{2}$$

Correction weights are the result of the application of some weighting technique. The characteristics of the correction weights should be such that the weighted estimator has better properties than the Horvitz–Thompson estimator.

Weighting is based on the use of *auxiliary information*. Auxiliary information is defined as a set of variables that have been measured in the survey and for which information on the population distribution is available. By comparing the population distribution of an auxiliary variable with its sample distribution, it can be assessed whether or not the sample is representative of the population (with respect to this variable). If these distributions differ considerably, one can conclude that nonresponse has resulted in a selective sample.

The auxiliary information can also be used to compute weighting adjustments. Weights are assigned to all records of observed elements. Estimates of population characteristics can now be obtained by using the weighted values instead of the unweighted values. The weights are then adjusted in such a way that population characteristics for the auxiliary variables can be computed without error. So when adjusted weights are applied to estimate the population means of auxiliary variables, the estimates must be equal to the true values, that is,

$$\bar{x}_{\mathrm{W}} = \frac{1}{N} \sum_{k=1}^{N} w_k t_k X_k = \overline{X}. \tag{3}$$

If this condition is satisfied, the weighted sample is said to be *representative* with respect to the auxiliary variable used.

If it is possible to make the sample representative with respect to several auxiliary variables, and if these variables have a strong relationship with the phenomena to be investigated, then the (weighted) sample will also be (approximately) representative with respect to these phenomena, and hence estimates of population characteristics will be more accurate.

A number of weighting techniques are described in the subsequent sections. The simplest technique is *poststratification*. This weighting adjustment technique is a special case of a more general approach called *linear weighting*. It is based on use of the generalized regression estimator. *Multiplicative weighting* is a different kind of weighting. It does not use a regression model. It is also shown that linear and multiplicative weighting are special cases of an even more general framework called *calibration estimation*, where it can be shown that the asymptotic properties of weighted estimates are identical. So, in practical situations, it does not matter whether linear or multiplicative weighting is used.

### *5.3. Poststratification*

To be able to carry out poststratification, one or more qualitative auxiliary variables are needed. Here, only one such variable is considered. The extension to more variables is straightforward. Suppose there is an auxiliary variable $X$ having $L$ categories. So it divides the population into $L$ strata. The strata are denoted by the subsets $U_1, U_2, \ldots, U_L$ of the population $U$. The number of population elements in stratum $U_h$ is denoted by $N_h$, for $h = 1, 2, \ldots, L$. The population size is equal to $N = N_1 + N_2 + \cdots + N_L$.

Poststratification assigns identical correction weights to all elements in the same stratum. The correction weight $c_k$ for an element $k$ in stratum $U_h$ is in its most general form defined by

$$c_k = \frac{N_h}{\sum\limits_{j \in U_h} \dfrac{t_j}{\pi_j}} \tag{4}$$

where the sum is taken over all sample elements $j$ in the stratum $U_h$. In case of simple random sampling, all inclusion probabilities $\pi_k$ are equal to $n/N$, and the correction weight $c_k$ reduces to

$$c_k = \frac{N_h}{N} \frac{n}{n_h} \tag{5}$$

If the values of the inclusion probabilities and correction weights are substituted in expression (2), the result is the well-known poststratification estimator

$$\overline{y}_{\text{ps}} = \frac{1}{N}\sum_{h=1}^{L} N_h \overline{y}_h, \tag{6}$$

where

$$\overline{y}_h = \frac{1}{n_h}\sum_{k\in U_h} t_k Y_k \tag{7}$$

is the sample mean of the observations in stratum $U_h$. So, the poststratification estimator is equal to a weighted sum of sample stratum means.

It can be shown (see, e.g., Bethlehem, 2002) that the bias due to nonresponse vanishes if there is no relationship between response behavior and the target variable of the survey. Two situations can be distinguished in which this is the case:

- The strata are homogeneous with respect to the target variable, that is, this variable shows little variation within strata;
- The strata are homogeneous with respect to the response behavior, that is, response probabilities show little variation within strata.

### *5.4. Linear weighting*

In the case of full response, the precision of simple estimators can be improved if suitable auxiliary information is available. If the auxiliary variables are correlated with the target variable, then for a suitably chosen vector $B$ of regression coefficients for a best fit of $Y$ on $X$, the residuals vary less than the values of the target variable itself. The ordinary least squares solution $B$ can, in the case of full response, be estimated by a vector $b$ which is asymptotically design unbiased. The *generalized regression estimator* is now defined by

$$\overline{y}_{\text{R}} = \overline{y}_{\text{HT}} + (\overline{X} - \overline{x}_{\text{HT}})'b, \tag{8}$$

in which $\overline{x}_{\text{HT}}$ is the vector of Horvitz–Thompson estimators for the vector of population means $\overline{X}$. The generalized regression estimator is asymptotically design unbiased. This estimator reduces the bias due to nonresponse if the underlying regression model fits the data well (see Bethlehem, 1988).

Bethlehem and Keller (1987) have shown that the generalized regression estimator (9) can be rewritten in the form of the weighted estimator (2), where the correction weight $c_k$ for the observed element $k$ is equal to $c_k = v'X_k$, and $v$ is a vector of weight coefficients which is equal to

$$v = N\left(\sum_{k=1}^{N} \frac{t_k X_k X_k'}{\pi_k}\right)^{-1} \overline{X}. \tag{9}$$

Poststratification turns out to be a special case of linear weighting. If the stratification is represented by a set of dummy variables, where each dummy variable denotes a specific stratum, expression (8) reduces to expression (6).

Linear weighting can be applied in more situations than poststratification. For example, poststratification by age class and sex requires the population distribution of the crossing of age class by sex to be known. If just the marginal population distributions of age class and sex are known separately, poststratification cannot be applied. Only one variable can be used. However, linear weighting makes it possible to specify a regression model that contains both marginal distributions. In this way, more information is used, and this will generally lead to better estimates.

The trick is to introduce sets of dummy variables. A complete cross-classification of sex by 10 age classes would require one set of $2 \times 10 = 20$ dummy variables (one for each stratum). For linear weighting only using the marginal distributions of the age classes and sex, two sets of dummy variables are introduced: one set of two dummy variables for the sex categories and another set of 10 dummy variables for the categories of age class. There is always a dummy variable representing the constant term in the regression model. This makes $2 + 10 + 1 = 13$ dummy variables.

Due to the special structure of the auxiliary variables, the computation of the weight coefficients $v$ cannot be carried out without imposing extra conditions. Here for every qualitative variable, the condition is imposed that the sum of the weight coefficients for the corresponding dummy variables must equal zero. The weight for an observed element is now obtained by summing the appropriate elements of this vector $v$.

Linear weighting is, however, not limited to quantitative auxiliary variables. It allows for inclusion in the model a mix of quantitative and qualitative variables.

Linear weighting has the advantage that it is relatively straightforward to compute variances of weighted estimators. It has the disadvantage that some correction weights may turn out to be negative. Such weights are not wrong, but simply a consequence of the underlying theory. Usually, negative weights indicate that the regression model does not fit the data too well. Some analysis packages (e.g., *SPSS*) are able to work with weights, but they do not except negative weights. This may be a reason not to apply linear weighting.

### *5.5. Multiplicative weighting*

Correction weights produced by linear weighting are the sum of a number of weight coefficients. It is also possible to compute correction weights in a different way, namely as the product of a number of weight factors. This weighting technique is usually called *raking* or *iterative proportional fitting*. Here, it is denoted by *multiplicative weighting* because weights are obtained as the product of a number of factors contributed by the various auxiliary variables.

Multiplicative weighting can be applied in the same situations as linear weighting, as long as only qualitative variables are used. Correction weights are the result of an iterative procedure. They are the product of factors contributed by all cross-classifications.

The iterative proportional fitting technique was already described by Deming and Stephan (1940). Skinner (1991) discussed the application of this technique in multiple frame surveys. Little and Wu (1991) described the theoretical framework and showed that this technique comes down to fitting a loglinear model for the probabilities of getting observations in strata of the complete cross-classification, given the probabilities for marginal distributions. To compute weight factors, the following scheme has to be carried out:

(1) Introduce a weight factor for each stratum in each cross-classification term. Set the initial values of all factors to 1.
(2) Adjust the weight factors for the first cross-classification term so that the weighted sample becomes representative with respect to the auxiliary variables included in this cross-classification.
(3) Adjust the weight factors for the next cross-classification term so that the weighted sample is representative for the variables involved. Generally, this will disturb representativeness with respect to the other cross-classification terms in the model.
(4) Repeat this adjustment process until all cross-classification terms have been dealt with.
(5) Repeat steps 2, 3, and 4 until the weight factors do not change any more.

Multiplicative weighting has the advantage that the weights are always positive. There are, however, no simple formulae for the variance of the weighted estimates.

### *5.6. Calibration estimation*

Deville and Särndal (1992) and Deville et al. (1993) have created a general framework for weighting of which linear and multiplicative weighting are special cases. The starting point is that adjusted weights $w_k = c_k/\pi_k = c_k d_k$ have to satisfy two conditions:

- The correction weights $c_k$ have to be as close as possible to 1.
- The weighted sample distribution of the auxiliary variables has to match the population distribution, that is,

$$\bar{x}_{\mathrm{W}} = \frac{1}{N}\sum_{k=1}^{N} t_k w_k X_k = \overline{X}. \tag{10}$$

The first condition sees to it that resulting estimators are unbiased, or almost unbiased, and the second condition guarantees that the weighted sample is representative with respect to the auxiliary variables used.

Deville and Särndal (1992) introduced a distance measure $D(w_k, d_k)$ that measures the difference between $c_k$ and 1 in some way. The problem is now to minimize

$$\sum_{k=1}^{N} t_k D(w_k, d_k) \tag{11}$$

under the condition (10). This problem can be solved by using the method of Lagrange. By choosing the proper distance function, both linear and multiplicative weighting can be obtained as special cases of this general approach. For linear weighting, the distance function $D$ is defined by $D(w_k, d_k) = (w_k - d_k)^2/d_k$, and for multiplicative weighting, the distance $D(w_k, d_k) = w_k \log(w_k/d_k) - w_k + d_k$ must be used.

Deville and Särndal (1992) and Deville et al. (1993) showed that estimators based on weights computed within their framework have asymptotically the same properties.

This means that for large samples, it does not matter whether linear or multiplicative weighting is applied. Estimators based on both weighting techniques will behave approximately the same. Note that, although the estimators behave in the same way, the individual weights computed by means of linear or multiplicative weighting may differ substantially.

### 5.7. *Software*

*Bascula* is a software package for calculating weights for all units in a sample using auxiliary information. The auxiliary information is used to specify a weighting model, which forms the basis for the weighting procedure. Several weighting methods are supported, among which are poststratification, linear weighting, and multiplicative weighting.

*Bascula* can also use the computed weights to estimate population totals, means, and ratios as well as variances based on Taylor linearization and/or balanced repeated replication (BRR). For the purpose of variance estimation, several sampling designs are supported.

*Bascula* is part of the *Blaise System* for computer-assisted survey processing. *Bascula* can be used either as a menu-driven interactive program or as a software component suitable for developing custom weighting/estimation applications. More information on *Bascula* can be found, for example, in Bethlehem (1998).

*CALMAR* is a *SAS* macro developed by the French National Statistics Office (INSEE), see Sautory (1993). It implements the calibration approach developed by Deville and Särndal (1992). *CALMAR* is an acronym for CALibration on MARgins, an adjustment technique which adjusts the margins (estimated from a sample) of a contingency table of two or more qualitative variables to the known population margins. However, the program is more general than mere "calibration on margins," since it also calibrates on the totals of quantitative variables. *CALMAR* offers four calibration methods, corresponding to four different distance functions:

- a linear function: the calibrated estimator is the generalized regression estimator;
- an exponential function with qualitative calibration variables. This comes down to multiplicative weighting;
- a logit function. This approach makes it possible to set upper and lower bounds for the weights;
- a truncated linear function. This approach is similar to that of the logit function.

The last two weighting methods are used to control the range of the distribution of correction weights. The logit function is used more often because it avoids excessively large weights, which can compromise the robustness of the estimates, and excessively small or even negative weights, which can be produced by the linear method.

Work is in progress on a new version, *CALMAR 2*, see Le Guennec and Sautory (2003). It implements the generalized calibration method of handling nonresponse proposed by Deville (1998). Its also does consistent weighting.

Also, for *SPSS*, there are special modules enabling researchers to carry out calibration. Vanderhoeft (2001) developed *g-DESIGN* and *g-CALIB-S* under *SPSS* 9.0. These modules make extensive use of *SPSS's* syntax language, matrix language, and macro facilities and are comparable to *CALMAR*. *g-CALIB-S* allows for virtually any calibration model to be applied. There is no restriction to calibration on margins (or totals of quantitative variables within categories of qualitative variables). The price for this generality is that preparation of input files for *g-CALIB-S* can be very complicated and is therefore not easy to automate. However, the additional module *g-DESIGN* contains several macros that, to a large extent, support construction of these files.

*CLAN* is a system of *SAS* macros developed by Statistics Sweden. It does not explicitly compute weights, but implements the generalized regression estimator, thereby taking into account the sampling design. It has several possibilities for nonresponse correction. For more information, see Andersson and Nordberg (1998).

The generalized estimation system, *GES*, developed in Statistics Canada, is also a system of *SAS* macros. It can produce estimates and variance estimates, taking into account a number of different sampling designs. Also, auxiliary variables can be included in estimation procedures so that implicit weighting is carried out (see Estevao et al., 1995).

## 6. Analysis

### *6.1. Analysis of dirty data*

Carrying out a sample survey is a time-consuming and costly activity. Therefore, attempts will be made to obtain interesting results from the collected data as much as possible. However, one must be careful. A lot can go wrong in the process of collecting and editing the data, and this has an impact on the results of the analysis.

Many data analysis techniques assume some kind of model stating that the data can be seen as an independent identically distributed random sample from some normal distribution. These assumptions are almost never satisfied in practical situations. More often, the *Dirty Data Theorem* applies. It states that data are usually obtained by a *dependent* sample with *unknown* and *unequal* selection probabilities from a *bizarre* and *unspecified* distribution, whereby some values are *missing* and many other values are subject to substantial *measurement errors*.

Researchers have to take into account the fact that data may be affected by measurement errors and nonresponse, that some values may not be observed but imputed, and that they have to use weights to compensate for a possible nonresponse bias.

Many statistical analysis packages assume the ideal model for the data and have no possibilities to account for the effects of dirty data. Therefore, researchers should be very careful in their choice of software for analysis.

### *6.2. Weighting issues*

Most general software packages for statistical analysis assume that the data come from a simple random sample. Then, simple quantities like sample means and sample percentages are unbiased estimators for their population analogs. When the sample is selected using a complex sample design, or if the survey is affected by nonresponse, unbiasedness is at stake.

Analysis packages like *SPSS*, *SAS*, and *Stata* have the possibility to use weights. To that end, a specific variable is assigned the role of weight. The question now is whether the use of these weights can help compute the unbiased estimates.

It should be realized that there are several types of weights. Each statistical package may interpret weights differently. Weights can even be interpreted differently within the same package. The following types are considered here:

- *Inclusion weights*. These weights are the inverse of the inclusion probabilities. Inclusion weights are determined by the sampling design. They must be known and nonzero to compute unbiased estimates (see Horvitz and Thompson, 1952).
- *Correction weights*. These weights are the result of applying some kind of weighting adjustment technique.
- *(Final) Adjusted weights*. These weights combine inclusion weights and correction weights. When applied, they should provide unbiased (or almost unbiased) estimates of population characteristics.
- *Frequency weights*. These weights are whole numbers indicating how many times a record occurs in a sample. They can be seen as a trick to reduce the file size.

To emphasize possible problems that may be encountered, a short overview is given of the treatment of weights in a few major statistical analysis packages.

Weights can be introduced in *SAS* with the WEIGHT statement. However, in one procedure weights may be interpreted as frequency weights, whereas in another they may be used as inclusion weights. If weights only assume integer values, they may be seen as frequency weights, whereas real-valued weights will be interpreted as inclusion weights.

*SPSS* consists of a series of modules. The two modules *SPSS Base* and *SPSS Advanced Models* treat weights as frequency weights. If weights are specified with the "weight by" command, real-valued weights will be rounded to integers. A new module *Complex Samples* was introduced in version 12 of *SPSS*. This module is capable of taking into account the inclusion weights.

*Stata* is more flexible in the use of weights. For many procedures, it is possible to indicate whether weights must be interpreted as either inclusion weights or frequency weights.

Problems may arise if weights are interpreted as frequency weights while in fact they are inclusion weights. Suppose a sample of size $n$ has been selected from a finite population of size $N$. The sample values of the target variable are denoted by $y_1, y_2, \ldots, y_n$. Let $\pi_i$ be the inclusion probability of element $i$, for $i = 1, 2, \ldots, n$. Then, the inclusion weight for element $i$ is equal to $1/\pi_i$. If these inclusion probabilities are interpreted as sample frequency weights, the weighted sample mean is computed as

$$\bar{y}_{\mathrm{W}} = \frac{\sum\limits_{i=1}^{n} w_i y_i}{\sum\limits_{i=1}^{n} w_i} = \frac{\sum\limits_{i=1}^{n} \dfrac{y_i}{\pi_i}}{\sum\limits_{i=1}^{n} \dfrac{1}{\pi_i}} \tag{12}$$

According to the theory of Horvitz and Thompson (1952), the unbiased estimator is equal to

$$\bar{y}_{\mathrm{HT}} = \frac{1}{N}\sum_{i=1}^{n} w_i y_i = \frac{1}{N}\sum_{i=1}^{n} \frac{y_i}{\pi_i} \tag{13}$$

Generally, these two estimators are not the same. However, in the case of simple random sampling, where we have equal probabilities $\pi_i = n/N$, expression (12) reduces to (13).

Similar problems occur when computing estimates for variances. Many statistical packages assume that the sample originated from an independent random sample selected with equal probabilities. If the weights are interpreted as frequency weights, then the sample size is equal to

$$w_+ = \sum_{i=1}^{n} w_i \tag{14}$$

and the proper estimator for the variance of the sample mean is

$$v(\bar{y}_\mathrm{W}) = \frac{\sum\limits_{i=1}^{n} w_i\left(y_i - \bar{y}_W\right)^2}{w_+(w_+ - 1)} \tag{15}$$

Usually, survey samples are selected without replacement, which means that the proper expression for the variance of the estimator is

$$v(\bar{y}_\mathrm{W}) = \left(\frac{1}{w_+} - \frac{1}{N}\right) \frac{\sum\limits_{i=1}^{n} w_i\left(y_i - \bar{y}_W\right)^2}{(w_+ - 1)} \tag{16}$$

If the finite population correction factor $f = w_+/N$ is small, expressions (15) and (16) are approximately the same.

The situation becomes more problematic if the weights $w_i$ contain inclusion weights. In the simple case of an equal probability sample ($w_i = N/n$), expression (15) will be equal to

$$v\left(\bar{y}_\mathrm{W}\right) = \frac{\sum\limits_{i=1}^{n} \left(y_i - \bar{y}_W\right)^2}{n(N - 1)}, \tag{17}$$

which is a factor $(N - 1)/(n - 1)$ too small as a variance estimator.

For general without-replacement sampling designs, a completely different expression should be used to estimate the variance of the estimator as follows:

$$v\left(\bar{y}_\mathrm{W}\right) = \sum_{i=1}^{n} \sum_{j=i+1}^{n} \frac{\left(\pi_i \pi_j - \pi_{ij}\right)}{\pi_{ij}} \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2. \tag{18}$$

Note that expression (18) involves second-order inclusion probabilities $\pi_{ij}$, which do not appear in expression (15).

The problems described above do also occur in a more in-depth analysis of the data. Many multivariate analysis techniques are based on the assumption of an identically distributed independent sample. Due to complex sampling designs, adjusted weighting and imputation (see Section 6.3), estimates for the first- and second-order moments of distributions are likely to be wrong.

### 6.3. Imputation issues

Some imputation techniques affect the distribution of a variable. They tend to produce synthetic values that are close to the centre of the original distribution. The imputed distribution is more "peaked." This may have undesirable consequences. Estimates of standard errors may turn out to be too small. A researcher using the imputed data (not knowing that the data set contains imputed values) may get the impression that his estimates are very precise, whereas in reality this is not the case.

Possible effects of imputation are illustrated by analyzing one single imputation technique: imputation by the mean. This type of imputation does not affect the mean of a variable: the mean $\overline{y}_{\mathrm{IMP}}$ after imputation is equal to the mean $\overline{y}_R$ before imputation. Consequently, the variance of the estimator also does not change.

Problems may arise when an unsuspecting researcher attempts to estimate the variance of estimators, for example, for constructing a confidence interval. To keep things simple, it is assumed that the available observations can be seen as a simple random sample without replacement, that is, missingness does not cause a bias. Then, the variance after imputation is equal to

$$V(\overline{y}_{\mathrm{IMP}}) = V(\overline{y}_R) = \frac{(1 - m/N)}{m} S^2, \tag{19}$$

in which $m \leq n$ is the number of "real" observations, and $S^2$ is the population variance.

It is known from sampling theory that, in case of a simple random sample without replacement, the sample variance $s^2$ is an unbiased estimator of the population variance $S^2$. This also holds for the situation before imputation: the $s^2$ computed using the $m$ available observations is an unbiased estimator of $S^2$.

What would happen if a researcher attempted to estimate $S^2$ using the complete data set, without knowing that some values have been imputed? He would compute the sample variance, and he would assume this is an unbiased estimator of the population variance. However, this is not the case. For the sample variance of the imputed data set, the following expression holds:

$$s^2_{\mathrm{IMP}} == \frac{m-1}{n-1} s^2 \tag{20}$$

Hence,

$$E(s^2_{\mathrm{IMP}}) = \frac{m-1}{n-1} S^2 \tag{21}$$

This is not an unbiased estimator of the population variance. The population variance is under-estimated. This creates the impression that the estimators are very precise, whereas in reality, this is not the case. So there is a substantial risk of drawing wrong conclusions from the data. This risk grows larger with each additional imputed value. For further details, see Lee et al. (2000).

Imputation also has an impact on the correlation between variables. Suppose the variable $Y$ is imputed using imputation by the mean, and suppose the variable $X$ is completely observed for the sample. It can be shown that in this case the correlation after imputation is equal to

$$r_{\mathrm{IMP},X,Y} = \sqrt{\frac{m-1}{n-1}} r_{XY}, \tag{22}$$

where $r_{XY}$ is the correlation in the data set before imputation. So, the more observations are missing for $Y$, the smaller the correlation coefficient will be. Researchers not aware of their data set having been imputed will get the impression that relationships between variables are weaker than they are in reality. Here again, there is a risk of drawing wrong conclusions.

### *6.4. Special software*

A naive researcher, working with standard analysis software, and unaware of the fact that he is working with "dirty data," runs serious risks of making mistakes in his statistical analysis of the data. Fortunately, the functionality offered by analysis software is constantly increasing.

Prior to version 12 of *SPSS*, it was not possible to correctly carry out a weighted analysis because estimates of variances of estimators were wrong. The *Complex Sample* module of version 12 of *SPSS* does conduct a weighted analysis correctly (and also allows for design effects due to clustering and stratification). Unfortunately, this model only includes descriptive statistics. For multivariate analysis, existing *SPSS* modules will have to be used. These will compute variances incorrectly.

Other major multipurpose statistical packages, like *Stata* and *SAS*, can do weighted analysis correctly. *Stata* can take into account the sampling design (stratified, clustered, or multistage) in estimating the variance of measures, such as totals, means, proportions, and ratios (either for the whole population or for different subpopulations), using the Taylor linearization method. There are also commands for jack-knife and bootstrap variance estimation, although these are not specifically oriented to survey data. Multivariate statistical analysis (e.g., linear, logistic, or probit regression) can also be carried out by taking into account the sampling design. However, *Stata* does not allow for variance estimation properly adjusted for poststratification.

*Sudaan* (www.rti.org/sudaan) is a statistical software package for the analysis of data from sample surveys (simple or complex). It uses the *SAS*-language and has similar interface, but it is a stand-alone package. It can estimate the variance of simple quantities (such as totals, means, ratios in the whole population or within domains) as well as more sophisticated techniques (parameter estimates of linear, logistic, and proportional hazard models). The available variance estimation techniques include the Taylor linearization, jack-knife, and BRR. Again, weighting adjustments are not generally supported.

*WesVar* (www.westat.com/wesvar) is a package primarily aimed at the estimation of basic statistics (as well as specific models) and corresponding variances from complex sample surveys using the method of replications (BRR, jack-knife, and bootstrap). Domain estimation and analysis of multiple-imputed data sets are accommodated. It can incorporate sample designs including stratification, clustering, and multistage sampling. Moreover, it can calculate (and take into account in the variance estimation) weights of nonresponse adjustments, complete or incomplete poststratification.

*Caljack* is a *SAS* macro developed by Statistics Canada (see Bernier and Lavallée, 1994). It is an extension of the *SAS* macro *CALMAR*. It implements variance estimation for stratified samples (of elements or clusters). It can compute variance estimates of simple statistics like totals, ratios, means, and percentages. It uses a jack-knife technique. It can take into account all calibration methods provided by *CALMAR*.

*GES* is a *SAS*-based application also developed by Statistics Canada (see Estevao et al., 1995). It can take into account stratified random sampling designs (of both elements and clusters). It does not support multistage designs. It can compute variance estimators for totals, means, proportions, and ratios (not only for the whole population, but also for the domains). Methods of variance estimation available include Taylor linearization and jack-knife techniques. It is possible to apply generalized regression estimation.

*CLAN* is a system of *SAS* macros developed by Statistics Sweden (see Andersson and Nordberg, 1998). Taking into account the sampling design (stratified or clustered), it provides point estimates and variance estimates for totals as well as for means, proportions, ratios, or any rational function of totals (for the whole population or domains). Incorporation of auxiliary information is supported through regression estimation. With respect to the treatment of unit nonresponse, it allows for specific nonresponse models (by defining response homogeneity groups) as well as incorporation of subsampling of nonrespondents.

## 7. Disclosure control

### *7.1. The disclosure problem*

Survey agencies experience an increasing demand for releasing survey data files, that is, data sets containing the scores on a number of variables for each respondent. Because of this trend and an increasing public concern regarding the privacy of individuals, there is a disclosure problem.

The disclosure problem relates to the possibility of identification of individuals in released statistical information and to reveal what these individuals consider to be sensitive information. Identification is a prerequisite for disclosure. Identification of an individual takes place when a one-to-one relationship can be established between a record in released statistical information and a specific individual.

Disclosure is undesirable for several reasons. In the first place, it is undesirable for legal reasons. In several countries, there is a law stating that firms should provide information to the statistical agencies while the agency may not publish statistical information in such a way that information about separate individuals, firms, and institutions becomes available.

In the second place, there is an ethical reason. When collecting data from individuals, survey agencies usually promise respondents that their data will be handled confidentially. The International Statistical Institute (ISI) Declaration on Professional Ethics (see ISI, 1986), states that "Statisticians should take appropriate measures to prevent their data from being published or otherwise released in a form that would allow any subject's identity to be disclosed or inferred."

In the third place, there is a very practical reason: if respondents do not trust statistical agencies, they will cease to respond. In many countries, nonresponse rates in household surveys have increased over the last decade, thereby affecting the quality of the survey results. A further rise is very undesirable.

Having stated that disclosure of data concerning individuals is unacceptable, the question arises, to what extent are statistical publications to be protected to achieve this goal. Too-heavy confidentiality protection of the data may violate another right: the freedom of information. It is the duty of every statistical office to collect and disseminate statistical information. It is this dilemma, right of anonymity versus freedom of information, which is the core of the considerations about disclosure control of microdata.

For more information on the disclosure problem, see, for example, Bethlehem et al. (1990).

### *7.2. Software for disclosure control*

The need for practical tools to establish the disclosure risks, and to reduce these risks, has triggered research in this area. Partly subsidized by the Fifth Framework Research Programme of the European Union, the CASC-project has produced software tools called $\mu$-ARGUS and $\tau$-ARGUS.

The $\mu$-ARGUS software aims at the protection of microdata sets. The starting point for the development of $\mu$-ARGUS was a view on safety/unsafety of microdata that is used at Statistics Netherlands. The incentive for building a package like $\mu$-ARGUS has been to allow data protectors to easily apply general rules for various types of microdata and to relieve them from the chore and tedium that can be involved in producing a safe file in practice.

The aim of statistical disclosure control is to limit the risk that sensitive information of individual respondents can be disclosed from data that are released to third party users. *Identifying variables* are those variables for which the scores are easily known to possible intruders, like municipality, sex, etc. On the basis of the frequency tables of identifying variables, an approximation of the disclosure risk is calculated. Records are considered not to be safe if a combination of its identifying variables does not occur frequently enough in the population. To replace the very basic risk approach, new risk estimators have been included in $\mu$-ARGUS, which could also take into account the structure of the households.

Traditional methods to hamper the possible reidentification of records are global recoding and local suppression. Global recoding will reduce the amount of detail in the identifying variables, whereas local suppression will change individual codes into missing values.

To replace the very basic risk approach, new risk estimators have been include in $\mu$-ARGUS, which could also take into account the structure of the households. Other methods available are

- multivariate micro-aggregation (grouping similar records together and replacing values of numerical variables by its mean),
- rank swapping (exchange the values between neighboring records),
- postrandomization (distort categorical variables with a known random mechanism, such that researches can correct this on an aggregate level),
- top and bottom coding (replacing the tails of the distribution by, for example, the mean of the extreme values),
- rounding, and
- noise addition.

All these methods have in common the fact that they will distort the individual records and make the disclosure much harder. Nevertheless, the resulting data files can be very well used by researchers for their analyses.

It should be noted that application disclosure protection techniques always reduces the amount of information in data file. For example, if a regional classification based on municipalities is replaced by a coarser one based on provinces, a detailed regional analysis is not possible any more. Some protection techniques may also to some extent affect distributional properties of variables.

The $\tau$-ARGUS software aims at the protection of statistical tables. Tables have traditionally been the major form of output of statistical agencies. Even in moderate size tables, there can be large disclosure risks. Protecting tables is usually done in two steps. First, sensitive cells are identified, and next these cells are protected.

To find the sensitive cells in tables containing magnitude data, often the well-known dominance $(n, k)$ rule is used. However, there is a tendency to apply the prior-posterior rule. This rule has several advantages (see Loeve, 2001). Locating the sensitive cells is by far the easiest part of the job. To protect these cells, additional cells have to be found to make the recalculation impossible. This leads to very complex mathematical optimization problems. Some solutions for this have been implemented in $\tau$-ARGUS. These problems have become even more complex due to the hierarchical structures in many of these tables.

For more information about the Argus software, see Hundepool et al. (2004, 2005).

*Sample Surveys: Design, Methods and Applications, Vol. 29A*
ISSN: 0169-7161
© 2009 Published by Elsevier B.V.
DOI: 10.1016/S0169-7161(08)00014-X

# Record Linkage

*William E. Winkler*[1]

## 1. Introduction

Record linkage consists of methods for matching duplicates within or across files using nonunique identifiers such as first name, last name, date of birth, address, and other characteristics. Fields such as first name, last name, date of birth, and address are referred to as *quasi-identifiers*. In combination, quasi-identifiers may uniquely identify an individual. Modern computerized record linkage began with the methods introduced by a geneticist Howard Newcombe (Newcombe et al., 1959; Newcombe and Kennedy, 1962), who used odds ratios (likelihood ratios) and value-specific, frequency-based probabilities. (common value of last name "Smith" has less distinguishing power than rare value "Zabrinsky"). Fellegi and Sunter, (1969, hereafter FS) gave a mathematical formalization of Newcombe's ideas. They proved the optimality of the decision (classification) rule of Newcombe and introduced many ideas about estimating optimal parameters (probabilities used in the likelihood ratios) without training data.

In this chapter, we will give background on the model of FS and several of the practical methods that are necessary for dealing with (often exceptionally) the messy data. Although the methods rely on statistical models, most development has been done by computer scientists using machine learning or database methods (Winkler, 2006a). Computer scientists refer to record linkage as *entity resolution*, *object identification*, or a number of other terms.

Applications of record linkage are numerous. In some situations, we might use a collection of lists to create a large list (survey frame) or update an existing large list. The updating and list maintenance can assure that we have good coverage of a desired population. The largest applications of record linkage are often during a population census or in updating an administrative list such as a national health directory or death index. Large typographical variation or error in fields such as first name, last name, and date of birth in a moderate proportion of records can make the updating quite difficult. Historically, some agencies have a full-time staff devoted to cleaning up the

[1] This report is released to inform interested parties of research and to encourage discussion. The views expressed are those of the author and not necessarily those of the U.S. Census Bureau.

lists (primarily manually). If they did not, then 1–3% error or more might enter the lists every year. The computerized record linkage methods can significantly reduce the need for clerical review and clean-up.

Another application of record linkage might be the matching of one list with another list to estimate the undercoverage/overcoverage of one of the lists that is believed to be reasonably complete. For the U.S. Census (Winkler, 1995), a large number of census blocks (contiguous regions of approximately 70 households) were reenumerated and matched against the main list of individuals. The computerized procedures reduced clerical review from an estimated 3000 individuals for 6 months to 300 individuals for 6 weeks. Because of the high quality of the lists and associated skills of individuals, false match rates of the computerized procedures were approximately 0.2%. More than 85% of matches were found automatically with the remainder of matches easily located among potentially matching individuals in the same household. The potentially matching individuals were often missing both first name and age.

Other applications of record linkage might involve reidentification experiments in which a public-use file only contains fields needed for demographic or economic analyses. Such fields might include a geocode, sex, age or age range, education level, and income level. Agencies release anonymized or masked data so that additional statistical analyses can be performed but do not wish "intruders" to reidentify individuals or data associated with individuals by placing names with individual records. Sweeney (1999) showed that 77+% of individuals can be uniquely identified by ZIP code, sex, and date of birth that are readily available in public lists such as voter registration databases. Until Sweeney's work many public-use health files contained ZIP code, sex, and date of birth. Winkler (1998), Sweeney (1999), and Evfimievski (2004) showed how to reidentify using a combination of analytic properties and record linkage. We do not cover reidentification in this chapter.

Record linkage can both increase the amount of coverage and reduce the amount of duplication in a survey frame. Frame errors can severely bias sampling and estimation. It is nearly impossible to correct errors in estimates that are based on sampling from a frame with moderate error (Deming and Gleser, 1959). After applying sophisticated record linkage, the 1992 Census of Agriculture (Winkler, 1995) contained 2% duplication whereas the 1987 Census of Agriculture contained 10% duplication. The duplication rates are based on field validation. Some estimates from the 1987 Agriculture Census with 10% duplication error may have been substantially biased.

The outline of this chapter is as follows. In the second section following this introduction we give background on the record linkage model of Fellegi and Sunter (1969), methods of parameter estimation without training data, string comparators for dealing with typographical error, an empirical example, and some brief comments on training data. The third section provides details of the difficulties with the preparation of messy data for linkage. Traditionally, file preparation has yielded greater improvements in matching efficacy than any other improvements. In the fourth section, we describe methods for error rate estimation without training data, methods for adjusting statistical analyses of merged files for linkage error, and techniques for speeding up record linkage. The final section consists of concluding remarks.

## 2. Overview of methods

In this section, we provide summaries of certain ideas of record linkage. Although the ideas are based on statistical models, the messiness of the data and the difficulty of developing certain algorithms for estimation and comparison have limited statistical agencies ability to create generalized computer systems that can be used in a variety of their applications.

### *2.1. The Fellegi–Sunter model of record linkage*

Fellegi and Sunter (1969) provided a formal mathematical model for ideas that had been introduced by Newcombe (Newcombe et al., 1959; Newcombe and Kennedy, 1962). They provided many ways of estimating key parameters. The methods have been rediscovered in the computer science literature (Cooper and Maron, 1978) but without proofs of optimality. To begin, notation is needed. Two files A and B are matched. The idea is to classify pairs in a product space $\mathbf{A} \times \mathbf{B}$ from two files A and B into $M$, the set of true matches, and $U$, the set of true nonmatches. Fellegi and Sunter, making rigorous concepts introduced by Newcombe et al. (1959), considered ratios of probabilities of the form:

$$R = P(\gamma \in \Gamma|M)/P(\gamma \in \Gamma|U) \tag{1}$$

where $\gamma$ is an arbitrary agreement pattern in a comparison space $\Gamma$. For instance, $\Gamma$ might consist of eight patterns representing simple agreement or not on the largest name component, street name, and street number. Alternatively, each $\gamma \in \Gamma$ might additionally account for the relative frequency with which specific values of name components such as “Smith,” “Zabrinsky,” “AAA,” and “Capitol” occur. The ratio $R$ or any monotonely increasing function of it such as the natural log is referred to as a *matching weight* (or score).

The decision rule is given by:

$$\begin{aligned} &\text{If } R > T_\mu, \text{ then designate pair as a match.}\\ &\text{If } T_\lambda \le R \le T_\mu, \text{ then designate pair as a possible match and hold for}\\ &\qquad\qquad \text{clerical review.}\\ &\text{If } R < T_\lambda, \text{ then designate pair as a nonmatch.} \end{aligned} \tag{2}$$

The cutoff thresholds $T_\mu$ and $T_\lambda$ are determined by a priori error bounds on false matches and false nonmatches. Rule (2) agrees with intuition. If $\gamma \in \Gamma$ consists primarily of agreements, then it is intuitive that $\gamma \in \Gamma$ would be more likely to occur among matches than nonmatches and ratio (1) would be large. On the other hand, if $\gamma \in \Gamma$ consists primarily of disagreements, then ratio (1) would be small. Rule (2) partitions the set $\gamma \in \Gamma$ into three disjoint subregions. The region $T_\lambda \le R \le T_\mu$ is referred to as the no-decision region or clerical review region. In some situations, resources are available to review pairs clerically.

Figure 1 provides an illustration of the curves of log frequency versus log weight for matches and nonmatches, respectively. The two vertical lines represent the lower and upper cutoffs thresholds $T_\lambda$ and $T_\mu$, respectively. The $x$-axis is the log of the likelihood

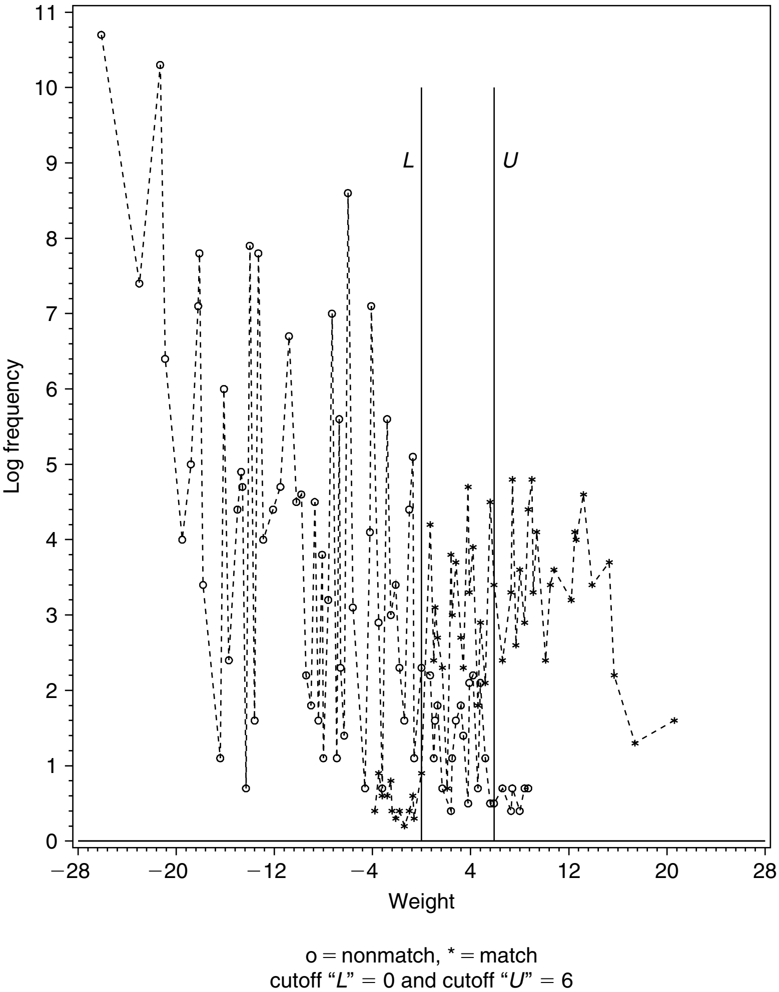


Fig. 1. Log frequency versus weight matches and nonmatches combined.

ratio $R$ given by (1). The $y$-axis is the log of the frequency counts of the pairs associated with the given likelihood ratio. The plot uses pairs of records from a contiguous geographic region that was matched in the 1990 Decennial Census. The clerical review region between the two cutoffs primarily consists of pairs within the same household that are missing both first name and age.

Table 1 provides examples of pairs of records that might be matched using name, address, and age. The pairs give the first indication that matching that might be straightforward for a suitably skilled person might not be easy with naive rules based on (1) and (2). If the agreement pattern $\gamma \in \Gamma$ on the pairs is simple agree or disagree on name, address, and age, then we see none of the pairs would agree on any of the three fields. In most situations, a suitably skilled person would be able to recognize that the first two pairs may be the same but unlikely to put a suitable score (or matching weight)

Table 1
Elementary examples of matching pairs of records (dependent on context)

| Name | Address | Age |
|---|---|---|
| 1a. John A Smith | 16 Main Street | 16 |
| 1b. J H Smith | 16 Main St | 17 |
| 2a. Javier Martinez | 49 E Applecross Road | 33 |
| 2b. Haveir Marteenez | 49 Aplecross Raod | 36 |
| 3a. Gillian Jones | 645 Reading Aev | 24 |
| 3b. Jilliam Brown | 123 Norcross Blvd | 43 |

on the first two pairs. The third pair must be taken in context. If the first record in the pair were individuals in medical school at the University of Michigan 20 years ago and the second record is from a current list of physicians in Detroit, Michigan, then, after suitable follow-up, we might determine that the third pair is a match.

If we had computerized parsing algorithms for separating the free-form name field into first name, middle initial, and last name and address in house number, street name, and other components, then we might have better patterns $\gamma \in \Gamma$ for applying (1) and (2). If we had suitable algorithms for comparing fields (e.g. Javier vs. Haveir) having typographical error, then we might be to give partial agreement to minor typographical error rather than calling a comparison a disagreement. Additionally, we might want standardization routines to replace the commonly occurring words with a common spelling ("Raod" with "Road" in pair two; "Aev" with "Ave" in pair three).

### 2.2. *Learning parameters*

Early record linkage systems were often for large administrative lists such as a national health index. The typical fields were name, address, date of birth, city of birth, and various fields associated with health information. The main administrative list might be cleaned-up in the sense that many components of the name, address, and other fields were reviewed and changed manually. As time progressed, the easiest of the manual procedures were replaced by computerized procedures that mimicked the manual procedures. For instance, it is straightforward to convert nicknames to possible legal names ("Bob" → "Robert") or obvious spelling errors ("Smitn" to "Smith") using lookup tables from prior manual review.

In virtually all real-world situations of which we are aware, training data have been unavailable. Practitioners have developed a number of ways for learning "optimal" record linkage parameters without training data. In all but one of the following subsections, we will describe methods of *unsupervised learning* where training data are unavailable.

#### 2.2.1. *Ideas of Newcombe*

Newcombe's ideas (Newcombe et al., 1959; Newcombe and Kennedy, 1962) are based on odds-ratios that are effectively likelihood ratios. He began with a large administrative list representing an entire population. The list had been cleaned up in the sense that duplicates were removed and inconsistent spelling or formatting was eliminated. Let file $C = (c_{ij}), 1 \leq i \leq N, 1 \leq j \leq N_c$, be a file with $Ns$ records (rows) and $N_c$

fields (columns). Newcombe wished to divide pairs in **C x C** into matches $M$ and nonmatches $U$. Although he knew the answer, he wished to be able to match external files A against C using the odds (conditional probabilities) that he developed from matching C against itself. Let $A_i$ represent agreement on field $i$, $A_i^c$ represent disagreement on field $i$, and $A_i^x$ represent agreement or disagreement on field $i$ but not both. Newcombe's first simplifying assumption is the conditional independence (CI) assumption that conditional on being in the set of matches $M$ or nonmatches $U$ agreement on field $i$ is independent of agreement on field $j$.

$$\textbf{(CI)} \quad P(A_i^x \cap A_j^x | D) = P(A_i^x | D) P(A_j^x | D) \tag{3}$$

where $D$ is either $M$ or $U$. Under condition (CI), Newcombe then computed the odds associated with each value of a specific field. The intuition is to bring the pairs together on common values of individual fields. For instance, with last name we might consider pairs agreeing on Smith or Zabrinsky. Let $(f_{ij})$, $1 \leq j \leq I_j$, represent the specific frequencies (number of values) of the $i$th field. The number of matches in $N$ and the number of nonmatches is $N \times N - N$. Among matches $M$, there are $f_{ij}$ pairs that agree on the $j$th value of the $i$th field. Among nonmatches $U$, there are $f_{ij} \times f_{ij} - f_{ij}$ pairs that agree on the $j$th value of the $i$th field. Then the odds ratio of agreement on the $j$th value of the $i$th field is

$$\begin{aligned} R_{1i} &= P(\text{agree } j\text{th value of } i\text{th field} | M) / P(\text{agree } j\text{th value of } i\text{th field} | U) \\ &= (f_{ij}/N)/(f_{ij} \times f_{ij} - f_{ij})/(N \times N - N). \end{aligned} \tag{4}$$

If pairs are taken from two files (i.e., product space of **A x B**), then we can use $f_{ij}$ as the frequency in A, $g_{ij}$ as the frequency in B, $h_{ij}$ as the frequency in A $\cap$ B (that is usually approximated with $h_{ij} = \min(f_{ij}, g_{ij})$), and make the appropriate changes in (4).

We notice that the sum of the probabilities of the numerator in Eq. (4) sum to 1. In practice, we assume that the sum of the probabilities is $1 - \varepsilon$ where $\varepsilon > 0$ and multiply all of the numerators in Eq. (4) by $1 - \varepsilon$. This allows a small probability of disagreement $\varepsilon > 0$ and $P(A_1|M) = 1 - \varepsilon$. The values of the $\varepsilon > 0$ were chosen via experience. In some situations there was clerical review on a subset of pairs and the $P(A_1|M)$ were reestimated. Although the reestimation (possibly after several iterations) was cumbersome, it did work well in practice.

Newcombe and others had observed the probabilities in the denominator could be approximated by random agreement probabilities

$$P(A_i|U) \approx P(A_i) = \sum_j f_{ij} f_{ij} / N^2, \tag{5}$$

Formula (5) is a reasonable approximation when the set of matches $M$ is not known. There are equivalent random agreement probabilities in the case of **A x B**.

There were only a few methods for dealing with typographical error. On receipt and keying of data, certain obvious misspelling ("William" vs. "Willam" or "Bill" vs. "William") might be changed by an analyst. Previously determined typographical variations might be placed in lookup tables that could be used for replace one spelling with another. The intent in all situations was to increase the proportion of matches that were found.

### 2.2.2. The methods of Fellegi and Sunter

Fellegi and Sunter (1969, Theorem 1) proved the optimality of the classification rule given by (2). Their proof is very general in the sense in it holds for any representations $\gamma \in \Gamma$ over the set of pairs in the product space **A x B** from two files. As they observed, the quality of the results from classification rule (2) were dependent on the accuracy of the estimates of $P(\gamma \in \Gamma|M)$ and $P(\gamma \in \Gamma|U)$.

Fellegi and Sunter (1969) were the first to give very general methods for computing these probabilities in situations that differ from the situations of Newcombe in the previous section. As the methods are useful, we describe what they introduced and then show how the ideas led into more general methods that can be used for *unsupervised learning* (i.e., without training data) in a large number of situations.

Fellegi and Sunter observed several things. First,

$$P(A) = P(A|M)P(M) + P(A|U)P(U) \tag{6}$$

for any set $A$ of pairs in **A x B**. The probability on the left can be computed directly from the set of pairs. If sets $A$ represent simple agreement/disagreement, under condition (CI), we obtain

$$P(A_1^x \cap A_2^x \cap A_3^x|D) = P(A_1^x|D)P(A_2^x|D)P(A_3^x|D), \tag{7}$$

then (6) and (7) provide seven equations and seven unknowns (as $x$ represent agree or disagree) that yield quadratic equations that they solved. Here $D$ is either $M$ or $U$. Equation (or set of equations) (7) is essentially the same as Eq. (3) and can be expanded to $K$ fields. Although there are eight patterns associated with the equations of the form (7), we eliminate one because the probabilities must add to one. In general, with more fields but still simple agreement/disagreement between fields, the equations can be solved via the EM algorithm in the next section. Probabilities of the form $P(A_i|D)$ are referred to as m-probabilities if $D = M$ and u-probabilities if $D = U$.

Fellegi and Sunter provided more general methods for frequency-based matching (value-specific) matching than those of Newcombe. Specifically, they obtained the general probabilities for simple agree/disagree and then scaled the frequency-based probabilities to the agree/disagree weights. If $A_1$ represents agreement on the first field and $v_j$, $1 \leq j \leq I_1$, are the values of the first field, then

$$P(A_1|D) = \sum_j P(A_1 \cap v_j|D) \tag{8}$$

where $D$ is either $M$ or $U$. Typically, $P(A_i|M) < 1$ for the simple agree/disagree weights on field $i$. This reflects the fact that there is less than 100% agreement on the $i$th field. Superficially, we can think of the $1 - P(A_i|M)$ as the average "typographical error" rate in the $i$th field. To make Eq. (8) valid under certain restrictions, FS assumed that the typographical error rate was constant over all values $v_j$, $1 \leq j \leq I_1$, associated with the $i$th field. Winkler (1989b) extended the frequency-based ideas of FS by showing how to do the computation under significantly weaker assumptions. The details of the computations (that we have greatly simplified) are given in their papers (FS, Winkler 1989b).

There are a number of implicit assumptions that are often made when matching two files and computing probabilities using (6)–(8). The first is that there is a significant overlap between two files A and B. This essentially means that $A \cap B$ is either most of $A$ or most of $B$. If this assumption is not true, then the probabilities obtained via Newcombe's methods or the FS methods may not work well. The second assumption is that neither file $A$ nor $B$ can simultaneously be samples from two larger files $A_2$ and $B_2$. Deming and Gleser (1959) provided theory demonstrating the unreliability of determining the sampling overlap (i.e., number of duplicates) from two sample files. As a case in point, if $A_2 = B_2$ each contain 1000 records on which 1% have the last name of Smith, among the matches $M$ between $A_2$ and $B_2$, there is a 1% probability of being a pair agreeing on Smith actually being a match. If $A$ and $B$ are 10% samples of $A_2$ and $B_2$, respectively, then among matches between $A$ and $B$, there is a 0.1% probability of a pair agreeing on Smith actually being a match. The third assumption is that the typographical error rates are quite low so the frequency-based computations based on the different observed values of the fields are valid. If a relatively rare value of last name such as Zabrinsky has six different spellings in the six records in which it appeared, then it is not possible to compute accurate frequency-based probabilities directly from the file.

In practice, it is necessary to perform *blocking* of two files that effect how pairs are brought together. If two files $A$ and $B$ each contain 10,000 records, then there are $10^8$ pairs in the product $\mathbf{A} \times \mathbf{B}$. Until very recently, we could not do the computation of $10^8$ pairs. In *blocking*, we only consider pairs that agree on certain characteristics. For instance, we may only consider pairs that agree on first initial of first name, last name, and date of birth. If we believe (possibly based on prior experience) that we are not getting a sufficiently large proportion of matches with a first blocking criteria, we may try a second. For instance, we may only consider pairs that agree on first initial of first name, first initial of last name, and the ZIP+4 code (that represents approximately 50 households). FS gave the straightforward theoretical extensions for blocking. In performing computation over pairs $P_1$ in $\mathbf{A} \times \mathbf{B}$ obtained via blocking, there is a fourth implicit assumption: that the pairs in $P_1$ contain a moderately high proportion of matches (say 3+% of $P_1$ consists of matches). In the next section, we return to the minimal needed proportion of pairs needing to be matches in more general situations. The methods of obtaining the probabilities given by (6)–(8) break down when the proportion of matches from $M$ in the set of pairs $P_1$ is too low. The computations also break down if we do the computation over all $10^8$ pairs in $\mathbf{A} \times \mathbf{B}$. In $\mathbf{A} \times \mathbf{B}$, at most 0.01% of the pairs are matches. In the next section, we will show how we can effectively find reasonable probabilities in a variety of situations.

### *2.2.3. EM algorithm*

In this section, we do not go into much detail about the basic EM algorithm because the basic algorithm is well understood. We provide a moderate amount of detail for the record linkage application so that we can describe a number of the limitations of the EM and some of the extensions.

For each $\gamma \in \Gamma$, we consider

$$P(\gamma) = P(\gamma|M)P(M) + P(\gamma|U)P(U) \tag{8a}$$

$$P(\gamma) = P(\gamma|C_1)P(C_1) + P(\gamma|C_2)P(C_2) \tag{8b}$$

$$P(\gamma) = P(\gamma|C_1)P(C_1) + P(\gamma|C_2)P(C_2) + P(\gamma|C_3)P(C_3) \tag{8c}$$

and note that the proportion of pairs having representation $\gamma \in \Gamma$ [i.e., left-hand side of Eq. (8)] can be computed directly from available data. In each of the variants, either $M$ and $U$, $C_1$ and $C_2$, or $C_1$, $C_2$, and $C_3$ partition **A x B**.

If the number of fields associated with $P(\gamma)$ is $K > 3$, then we can solve the combination of equations given by (8) and (7) using the EM algorithm. Although there are alternate methods of solving the equation such as methods of moments and least squares, the EM is greatly preferred because of its numeric stability. Under CI, programming is simplified and computation is greatly reduced (from $2^k$ to $2k$).

Caution must be observed when applying the EM algorithm to real data. The EM algorithm that has been applied to record linkage is a *latent class algorithm* that is intended to divide **A x B** into the desired sets of pairs $M$ and $U$. The probability of a class indicator that determines whether a pair is in $M$ or $U$ is the missing data must be estimated along with the m- and u-probabilities. It may be necessary to apply the EM algorithm to a particular subset S of pairs in **A x B** in which most of the matches $M$ are concentrated, for which the fields used for matching clearly can separate $M$ from $U$, and for which suitable initial probabilities can be chosen. Because the EM is a local maximization algorithm, the starting probabilities may need to be chosen with care based on experience with similar types of files. Because the EM latent-class algorithm is a general clustering algorithm, there is no assurance that the algorithm will divide **A x B** into two classes $C_1$ and $C_2$ that almost precisely correspond to $M$ and $U$.

The following example characterizes some of the cautions that must be observed when applying the EM. As we will observe, the EM, when properly applied, can supply final limiting parameters that are quite effective. In extensive Decennial Census work, we observed that the final limiting parameters often reduced the size of the clerical review region by 2/3 from the region that might have been obtained by the initial parameters obtained from knowledgeable guesses. In the following we use 1988 Dress Rehearsal Census data from one of the 457 regions of the U.S. that we used for the 1990 Decennial Census. The matching fields consist of last name, first name, house number, street name, phone, age, and sex. In actuality, we also used middle initial, unit (apartment identifier), and marital status. The first file A is a sample of blocks from the region and the second file is an independent enumeration of the same sample of blocks. The first file size is 15,048 and the second file size is 12,072. In the first part of the example, we only consider 116,305 pairs that agree on Census block id and first character of surname and, in the second part, we only consider the 1,354,457 pairs that agree on Census block id only. A census block consists of approximately 70 households whereas a ZIP+4 area represents approximately 50 households. We observe that there can be at most 12,072 matches if the smaller file is an exact subset of the larger file. As is typical in population censuses, the work begins with address lists of households in which the data from the survey forms are used to fill-in information associated with individuals. In many situations (such as with families), there will be more than one individual associated with each address (housing unit).

We begin by applying the (2-class) EM to the set of 110,305 pairs. We use knowledgeable initial probabilities that we believe correspond to the probabilities we need for matching individuals. We also use a precursor program to get the counts (or probabilities) of the form $P(\gamma)$ that we use in the EM algorithm. In the limit, we get the final probabilities given in Table 2. The final proportion of matches in the first class $P(M) = 0.2731$ is much too large. The m-probability $P(\text{agree first} \mid M) = 0.31$ is much

Table 2
Initial and final probabilities from 2-class EM fitting

| | Initial | | Final | |
|---|---|---|---|---|
| | *m* | *u* | *m* | *u* |
| Last | 0.98 | 0.24 | 0.95 | 0.07 |
| First | 0.98 | 0.04 | 0.31 | 0.01 |
| Hsnm | 0.94 | 0.24 | 0.98 | 0.03 |
| Stnm | 0.66 | 0.33 | 0.99 | 0.47 |
| Phone | 0.70 | 0.14 | 0.68 | 0.01 |
| Age | 0.88 | 0.11 | 0.38 | 0.07 |
| Sex | 0.98 | 0.47 | 0.61 | 0.49 |

Table 3
Initial and final probabilities from 3-class EM fitting

| | Initial | | | Final | | | |
|---|---|---|---|---|---|---|---|
| | *m* | *i* | oh | *m* | *i* | oh | *u* |
| Last | 0.98 | 0.90 | 0.24 | 0.96 | 0.92 | 0.07 | 0.25 |
| First | 0.98 | 0.24 | 0.04 | 0.96 | 0.02 | 0.01 | 0.01 |
| Hsnm | 0.94 | 0.90 | 0.24 | 0.97 | 0.97 | 0.04 | 0.23 |
| Stnm | 0.66 | 0.90 | 0.33 | 0.98 | 0.99 | 0.47 | 0.58 |
| Phone | 0.70 | 0.60 | 0.14 | 0.72 | 0.64 | 0.01 | 0.14 |
| Age | 0.88 | 0.20 | 0.11 | 0.88 | 0.14 | 0.07 | 0.08 |
| Sex | 0.98 | 0.70 | 0.47 | 0.98 | 0.45 | 0.49 | 0.49 |

too small. What has gone wrong? We observe that addresses are of high quality. Because we are in very small contiguous regions (blocks), last name, house number, street name, and phone are likely to be the same in most housing units associated with families. The higher quality household information outweighs the person fields of first name, age, and sex that might be used to separate individuals within household.

We overcome the situation by creating a 3-class EM that we hope divides records agreeing on household variables into 2-classes and leaves a third class that would be nonmatches outside of households. The initial ideas were due to Smith and Newcombe (1975) who provided separate ad hoc weighting (likelihood) adjustments for the set of person fields and the set of household fields. Their ideas have been verified by Gill (1999) among others. As the EM algorithm is quite straightforward to convert to 3-classes, we make the appropriate algorithmic adjustments and choose appropriate starting probabilities. Winkler (1993b) provides details. Table 3 gives initial probabilities for a first class that we hope corresponds to person matches $M$ within a household, an in-between class I that we hope corresponds to nonmatches within the same household, and a class $\mathrm{O}_h$ that are pairs not agreeing on household fields. To get the final u-probabilities we combine the i-probabilities and oh-probabilities according to the proportions in classes 2 and 3. When we run the EM program, we get probabilities of being in the three classes of 0.0846, 0.1958, and 0.7196, respectively. The probability 0.0846 associated with the first class accurately corresponds to the known number of true matches (obtained via two levels of review and one level of adjudication). Notice that the starting i-probabilities are reasonable guesses for the probabilities of persons within the same household who are not matches.

If we apply the 3-class EM algorithm to the 1,354,457 pairs agreeing on block (but not block plus first character of last name) and use good initial guesses for the probabilities, then we get similarly "good" m-probabilities as we did in Table 3. This is true even though the estimated proportion of pairs in the first class is 0.0081. In general, when we begin with sets of pairs that are much too large, the EM algorithm will not converge to estimates that are not reasonable for separating matches from the other pairs. The EM algorithm when applied to the much larger set of pairs can be much more sensitive to the set of starting points.

If the EM algorithm is applied with care, then it will generally yield good parameter estimates with lists of individuals. It will not always yield reasonable lists with agriculture or business lists because of the (moderately) high proportion of truly matching pairs that disagree on names or on addresses. The EM algorithm was used for production matching in the 1990 Decennial Census (Winkler and Thibaudeau, 1991) because Winkler (1989a) had been able to demonstrate that matching probabilities (particularly m-probabilities) varied significantly (say between a suburban area and an adjacent urban area). If we think of $1 - P(A_i|M)$ as crudely representing the average typographical error in the $i$th field, then the variation of parameters is understandable because lists associated with urban areas often contain more typographical error.

Winkler (1988, 1989a) showed the EM algorithm yielded "optimal parameters" in the sense of effective local maxima of the likelihood. The 2-class and 3-class EM algorithms under condition (CI) are quite robust. If starting points are varied substantially, the EM converges to the same limiting values where the limiting values are determined by characteristics of the files A and B. The 2-class algorithm will outperform the 3-class algorithm in situations where there is typically only one entity at an address (or telephone number). In those situations, the address can be considered an identifier of the individual entity.

During 1990 production matching, the EM algorithm showed its flexibility. In three regions among a number of regions processed in 1 week, clerical review became much larger with the EM parameters than was expected. Upon quick review, we discovered that two keypunchers had managed to bypass edits on the year of birth. All records from these keypunchers disagreed on the computed age. The clerical review became much larger because first name and the age were the main fields for separating persons within a household.

More generally, we may wish to account for dependencies directly using appropriate loglinear models (Bishop et al., 1975). Winkler (1993b) provides a general EMH algorithm that accounts for the general interactions between fields and allows convex constraints to predispose certain estimated probabilities into regions based on a priori information used in similar matching projects. The EMH algorithm is a form of MCECM algorithm (Meng and Rubin, 1993) that additionally allows convex constraints. The interaction EM can yield parameters that yield slight improvements in matching efficacy. It is much more difficult to apply because of its sensitivity to moderate changes in the set of interactions. Winkler (1993b) and Larsen and Rubin (2001) demonstrated that effective sets of interactions can be selected based on experience. The starting point for the interaction EM is the set of parameters from the CI EM.

### 2.3. String comparators

In most matching situations, we will get poor matching performance when we compare two strings exactly (character-by-character) because of typographical error. Dealing with typographical error via approximate string comparison has been a major research project in computer science (see e.g., Hall and Dowling, 1980; Navarro, 2001). In record linkage, we need to have a function that represents approximate agreement, with agreement being represented by 1 and degrees of partial agreement being represented by numbers between 0 and 1. We also need to adjust the likelihood ratios (1) according to the partial agreement values. Having such methods is crucial to matching. For instance, in a major census application for measuring undercount, more than 25% of matches would not have been found via exact character-by-character matching. Three geographic regions (St. Louis, urban; Columbia, MO, suburban; and Washington, suburban/rural) are considered in Table 4. The function Φ represents exact agreement when it takes value 1 and represents partial agreement when it takes values less than 1. In the St. Louis region, for instance, 25% of first names and 15% of last names did not agree character-by-character among pairs that are matches.

Jaro (1989) introduced a string comparator that accounts for insertions, deletions, and transpositions. The basic Jaro algorithm has three components: (1) compute the string lengths, (2) find the number of common characters in the two strings, and (3) find the number of transpositions. The definition of common is that the agreeing character must be within half the length of the shorter string. The definition of transposition is that the character from one string is out of order with the corresponding common character from the other string. The string comparator value (rescaled for consistency with the practice in computer science) is:

$$\Phi_j(s_1, s_2) = 1/3(N_\mathrm{C}/\mathrm{len}_{s1} + N_\mathrm{C}/\mathrm{len}_{s2} + 0.5N_\mathrm{t}/N_\mathrm{C}) \tag{9}$$

where $s_1$ and $s_2$ are the strings with lengths $\mathrm{len}_{s1}$ and $\mathrm{len}_{s2}$, respectively, $N_C$ is the number of common characters between strings $s_1$ and $s_2$ where the distance for common is half of the minimum length of $s_1$ and $s_2$, and $N_t$ is the number of transpositions. The number of transpositions $N_t$ is computed somewhat differently from the obvious manner.

Using truth data sets, Winkler (1990) introduced methods for modeling how the different values of the string comparator affect the likelihood (1) in the Fellegi–Sunter decision rule. Winkler (1990) also showed how a variant of the Jaro string comparator Φ dramatically improves matching efficacy in comparison to situations when string

Table 4
Proportional agreement by string comparator values among matches

| | Key Fields by Geography | | |
|---|---|---|---|
| | StL | Col | Wash |
| First | | | |
| $\Phi = 1.0$ | 0.75 | 0.82 | 0.75 |
| $\Phi \geq 0.6$ | 0.93 | 0.94 | 0.93 |
| Last | | | |
| $\Phi = 1.0$ | 0.85 | 0.88 | 0.86 |
| $\Phi \geq 0.6$ | 0.95 | 0.96 | 0.96 |

comparators are not used. The Winkler variant employs some ideas of Pollock and Zamora (1984) in a large study for the Chemical Abstracts Service. They provided empirical evidence that quantified how the probability of keypunch errors increased as the character position in a string moved from the left to the right. The Winkler variant, referred to as the *Jaro–Winkler string comparator*, is widely used in computer science.

Work by Cohen et al. (2003a,b) provides empirical evidence that the new string comparators can perform favorably in comparison to Bigrams and Edit Distance. Edit Distance uses dynamic programming to determine the minimum number of insertions, deletions, and substitutions to get from one string to another. The Bigram metric counts the number of consecutive pairs of characters that agree between two strings. A generalization of bigrams is $q$-grams where $q$ can be greater than 2. Cohen et al. (2003a,b) provided additional string comparators that they demonstrated slightly outperformed the Jaro–Winkler string comparator with several small test decks but not with a test deck similar to Census data. Yancey (2005), in a rather exhaustive study, also demonstrated Jaro–Winkler string comparator outperformed new string comparators of Cohen et al. (2003a,b) with large census test decks. Yancey introduced several hybrid string comparators that used both the Jaro–Winkler string comparator and variants of edit distance.

Cohen et al. (2003a,b) observed that the computational algorithm for edit distance is 10 times as slow as the corresponding algorithm for the Jaro–Winkler string comparator. The speed of the string comparator dramatically affects the speed of matching software. It is fairly typical for matching software with the Jaro–Winkler string comparator to expend 30–70% of the CPU cycles in the string comparator subroutine.

Table 5 compares the values of the Jaro, Winkler, Bigram, and Edit-Distance values for selected first names and last names. Bigram and Edit Distance are normalized to be

Table 5
Comparison of string comparators using last names and first names

| Two Strings | | String Comparator Values | | | |
|---|---|---|---|---|---|
| | | Jaro | Winkler | Bigram | Edit |
| Shackleford | Shackelford | 0.970 | 0.982 | 0.800 | 0.818 |
| Dunningham | Cunnigham | 0.867 | 0.867 | 0.917 | 0.889 |
| Nichleson | Nichulson | 0.926 | 0.956 | 0.667 | 0.889 |
| Jones | Johnson | 0.867 | 0.893 | 0.167 | 0.667 |
| Massey | Massie | 0.889 | 0.933 | 0.600 | 0.667 |
| Abroms | Abrams | 0.889 | 0.922 | 0.600 | 0.833 |
| Hardin | Martinez | 0.778 | 0.778 | 0.286 | 0.143 |
| Itman | Smith | 0.467 | 0.467 | 0.200 | 0.000 |
| Jeraldine | Geraldine | 0.926 | 0.926 | 0.875 | 0.889 |
| Marhta | Martha | 0.944 | 0.961 | 0.400 | 0.667 |
| Michelle | Michael | 0.833 | 0.900 | 0.500 | 0.625 |
| Julies | Julius | 0.889 | 0.933 | 0.800 | 0.833 |
| Tanya | Tonya | 0.867 | 0.880 | 0.500 | 0.800 |
| Dwayne | Duane | 0.778 | 0.800 | 0.200 | 0.500 |
| Sean | Susan | 0.667 | 0.667 | 0.200 | 0.400 |
| Jon | John | 0.778 | 0.822 | 0.333 | 0.750 |
| Jon | Jan | 0.778 | 0.800 | 0.000 | 0.667 |

between 0 and 1. All string comparators take value 1 when the strings agree character by character.

### 2.4. An empirical example

In the following, we compare different matching procedures on the data that were used for the initial EM analyses (Tables 2 and 3). Although we also demonstrated very similar results with several alternative pairs of files, we do not present the additional results here (see Winkler, 1990). The results are based only on pairs that agree on block identification code and first character of the last name.

The procedures that we use are as follows. The simplest procedure, *crude*, merely uses an ad hoc (but knowledgeable) guess for matching parameters and does not use string comparators. The next, *param*, does not use string comparators but does estimate the m- and u-probabilities. Such probabilities are estimated through an iterative procedure that involves manual review of matching results and successive reuse of reestimated parameters. Such iterative-refinement procedures are a feature of Statistics Canada's CANLINK system.

The third type, *param2*, uses the same probabilities as *param* and the basic Jaro string comparator. The fourth type, *em*, uses the EM algorithm for estimating parameters and the Jaro string comparator. The fifth type, *em2*, uses the EM algorithm for estimating parameters and the Winkler variant of the string comparator that performs an upward adjustment based on the amount of agreement in the first four characters in the string.

In Table 6, the cutoff between designated matches is determined by a 0.002 false match rate. The *crude* and *param* types are allowed to rise slightly above the 0.002 level because they generally have higher error levels. In each pair of columns (designated matches and designated clerical pairs), we break out the counts into true matches and true nonmatches. In the designated matches, true nonmatches are false matches.

By examining the table, we observe that a dramatic improvement in matches can occur when string comparators are first used (from *param* to *param2*). The reason is that disagreements (on a character-by-character basis) are replaced by partial agreements and adjustment of the likelihood ratios (see Winkler 1990). The improvement

Table 6
Matching results via matching strategies

| | Designated Computer Match | Designated Clerical Pair |
|---|---|---|
| Truth | Match/Nonmatch | Match/Nonmatch |
| *Crude* | 310/ 1 | 9344/794 |
| *param* | 7899/ 16 | 1863/198 |
| *param2* | 9276/ 23 | 545/191 |
| *em* | 9587/ 23 | 271/192 |
| *em2* | 9639/ 24 | 215/189 |

*Note*: 0.2% false matches among designated matches.

due to the Winkler variant of the string comparator (from *em* to *em2*) is quite minor. The *param* method is essentially the same as a traditional method used by Statistics Canada. After a review of nine string comparator methods (Budzinsky, 1991), Statistics Canada provided options for three string comparators in CANLINK software with the Jaro–Winkler comparator being the default.

The improvement between *param2* and *em2* is not quite as dramatic because it is much more difficult to show improvements among "hard-to-match" pairs and because of the differences in the parameter-estimation methods. Iterative refinement is used for *param2* (a standard method in CANLINK software) in which pairs are reviewed, reclassified, and parameters reestimated. This method is a type of (partially) supervised learning and is time-consuming.

The improvement due to the parameters from *em2* can be explained because the parameters are slightly more general than those obtained under CI. If $A_i^x$ represents agreement or disagreement on the $i$th field, then our CI assumption yields

$$P(A_1^x \cap A_2^x \cdots \cap A_k^x | D) = \prod_{i=1}^{k} P(A_i^x | D) \tag{10}$$

where $D$ is either $M$ or $U$. Superficially, the EM considers different orderings of the form

$$P(A_{\rho,1}^x \cap \cdots \cap A_{\rho,k}^x | D) = \prod_{i=1}^{k} P(A_{\rho,i}^x | A_{\rho,i=1}^x, \cdots, A_{\rho,1}^x, D) \tag{11}$$

where $\rho, i$ represents the $i$th entry in a permutation $\rho$ of the integers 1 thru $k$. The greater generality of (11) in comparison to (10) can yield better fits to the data. We can reasonably assume that the EM algorithm under the CI assumption (as the actual computational methods work) simultaneously chooses the best permutation $\rho$ and the best parameters.

In this section, we have demonstrated that very dramatic improvement in record linkage efficacy through advancing from seemingly reasonable ad hoc procedures to procedures that use modern computerized record linkage procedures. The issue that affects statistical agencies is whether their survey frames are well-maintained using effective procedures. Upgrading matching procedures is often as straightforward as replacing a subroutine that uses ad hoc methods with another subroutine. It is crucial to never assume that moderately sophisticated record linkage procedures are being used as the following situation demonstrates.

Maintenance of state voter registration lists is a situation where efficacy could be enhanced by moving from ad hoc to modern record linkage procedures. There have been two U.S. Federal laws (in 1993 and 2002) allocating money and mandating requirements on list maintenance. The voter registration lists are compared to department of motor vehicle lists, social services lists, and other lists including the main U.S. Social Security Administration list. Each list is searched internally for duplicates. All the states (Levitt el al., 2005) appear to be using ad hoc matching procedures that were originally developed for matching department of motor vehicle lists. The efficacy of the state ad hoc computer matching procedures in many situations may be between the worst two methods (*crude* and *param*) in Table 6.

### 2.5. *Training data*

Representative training data are seldom available for getting the parameters for record linkage classification rules. If training data are available, then it is possible to get the parameters by adding appropriate quantities to yield the probabilities in (1) and (2). In fact, with sufficient training data, it is straightforward to estimate probabilities in (1) that account for the dependencies between different matching fields and to estimate error rates.

Winkler (1989a) showed that optimal record linkage parameters vary significantly in different geographic regions. For the 1990 U.S. Decennial Census, training data would have been needed for the 457 regions where matching was performed. The amount of time needed to obtain the training data in the 457 regions would have substantially exceeded the 3 weeks that was allotted for the computer matching. In more than 20 years of record linkage at the Census Bureau, there have never been training data. In more than 30 years in maintaining the National Health Files and performing other large matching projects at Oxford University, Gill (2000, private communication) never had training data.

## 3. Data preparation

In matching projects, putting the data from two files A and B into consistent forms so that the data can be run through record linkage software often requires more work (3–12 months with a moderate or large staff) than the actual matching operations (1–3 weeks with one individual). Inability or lack of time and resources for cleaning up files in preparation of matching are often the main reasons that matching projects fail. We provide details of file acquisition, preparation, and standardization in the next sections.

### 3.1. *Description of a matching project*

Constructing a frame or administrative list entities for an entire country or a large region of a country involves many steps. The construction methods also hold pairs of lists or for the situation of finding duplicates within a given list.

(1) Identify existing lists that can be used in creating the main list. In this situation, it is important to concentrate on 10 or fewer lists. It is practically infeasible to consider thousands of lists.

(2) With each list, obtain an annotated layout. The annotation should include the locations of different fields and the potential values that different fields can assume. For instance, a given list may have several status codes associated with whether the entity is still in business or alive. With lists of businesses, it may have additional status codes denoted whether the record is associated with another entity as a subsidiary or duplicate. If the annotated layout is not available, then reject the list. If the list is on an incompatible computer system or in an incompatible format such as a typed list or microfiche, then reject the list.

(3) Obtain the lists to begin putting them in a standard format that will be used by the duplicate-detection and updating programs. If the list will not pass through name and address standardization programs, then reject it. If some or many records in the list cannot be standardized, then consider rejecting the list or only use records that can be standardized. The standard format should include a field for the source of a list and the date of the list. If possible, it is a good idea to also have a date for the individual record in the list.

(4) If resources permit, greater accuracy may be obtained by matching each potential update source against the main list sequentially. Matching each list in a sequential manner allows more accurate clerical clean-up of duplicates. If the clerical clean-up cannot be done in an efficient manner, then duplicates in the main list will yield more and more additional duplicates as the main list is successively updated. If it appears that an individual list is causing too many duplicates to be erroneously added to the main list, then reject the list as an update source. If a large subset of the update source does not yield a sufficiently large number of new entities in the main list, then it might also be excluded.

(5) After the initial matching, additional computerized and clerical procedures should be systematically applied for further identifying duplicates in the main list. A very useful procedure is to assure that the representations of names and addresses associated with an entity are in the most useful form and free of typographical errors. These extra improvement procedures should be used continuously. If updates and clean-ups of lists containing many small businesses are only done annually, then the overall quality of the list can deteriorate in an additive fashion during each subsequent update. In the U.S., it is known that the yearly turnover (going in and out of business or substantial changes in name and address information that make updating very difficult) can exceed 10% with small businesses.

Many matching projects fail because groups cannot even get through the first 1–2 steps mentioned above. Maintaining lists can be difficult. In the U.S., the Postal Change of Address files for individuals represent 16% of the population per year. Some individuals may move more than once. With lists of small business (such as petroleum retailers), the change of name or address can exceed 10% per year. In maintaining a large national health file or national death index, 1–3% net error per year can yield substantial error after several years.

### 3.2. *Initial file preparation*

In obtaining the files, the first issue is to determine whether the files reside in sequential (standard flat) files, databases, or in SAS files. As most record linkage software is designed for only sequential files, files in other formats will need to have copies that are in sequential formats. Some groups that do record linkage with many files will have a standard format and procedures so that the files are in the most compatible form for record linkage. An annotated layout will give the descriptions of individual fields that might be compared. For instance, a sex code might be broken out into Sex1 (male = M, female = F, missing = b where b represents blank) or Sex2 (male = 1, female = 2, missing = 0). Simple programs can have tables that are used in converting from one set of codes to another set of codes.

It is very typical for well-maintained files to carry status codes indicating whether an entity is still alive or in business and whether information such as an address or telephone number is current. If a file has status codes indicating that certain records are out-of-scope, then in most matching applications the out-of-scope records should be dropped before using the file for updating or merging. In some files, it may be difficult to determine out-of-scopes. For instance, electric utilities have very good address information that individuals might wish to use in updating a list of residences. Unfortunately, electric utilities typically include small commercial establishments with residential customers because they maintain their lists by flow-rate categories. If the electric utility list is used to update a list of households, many "out-of-scope" commercial addresses will be added.

It may be necessary to review various fields across two files. For instance, if one file has addresses that are almost entirely of the form house number and street name and another file has a substantial portion of the addresses in the form PO Box, then it may be difficult to match to two files using name and address information. With lists of businesses, it may be necessary to have auxiliary information that allows separating headquarters from subsidiaries. With many businesses, headquarters fill out survey forms. If a survey form is sent to the subsidiary and returned, then the survey organization may double-count the information from the subsidiary that is also reported in the totals from the headquarters.

In the following, we provide summaries of various procedures that can be used for the preliminary cleaning of files and can often be in straightforward computer routines. These consistency checks and clean-up procedures prior to running files through a matching program are referred to as *standardization*.

(1) Replacing spelling variants with a common consistent spelling is referred to as *spelling standardization.*

    (a) Replace 'Doctor', 'Dr.' with 'Dr'
    (b) Replace nicknames such as 'Bob', 'Bill' with 'Robert' and 'William'
    (c) Replace words such as 'Company', 'Cmpny', 'Co.' with 'Co'

    Note: The third example is application dependent because 'Co' can refer to county or Colorado.

(2) Replacing inconsistent codes is referred to as assuring *code consistency.*

    (a) Replace Sex Sex (male='1', female='2', missing='0') with (male='M', female='F', missing='')
    (b) Replace 'January 11, 1999' and '11 January, 1999' with MMDDYYYY='01111999' or YYYYMMDD='19990111'

    Code consistency is sometimes referred to as making the value-states of variables (or fields) consistent.

    In record linkage, a variable (or field) is typically a character string such as a complete name, complete address, or a sub-component such as first name or last name.

(3) Identifying the starting and ending positions of the individual components of a free form string such as a name or address is referred to as *parsing.*

(a) Identify locations of first name, middle initial, and last name in ‘Mr John A Smith Jr’ and ‘John Alexander Smith’
(b) Identify locations of house number and street name in ‘123 East Main Street’ and ‘123 E. Main St. Apt. 16’

The idea of parsing is to allow the comparison of fields (variables) that should be consistent and reasonably easy to compare. It is not easy to compare free-form names and addresses except possibly manually. The above three ideas of standardization are often preliminary to situations when free-form names and addresses are broken (parsed) into components. We cover general name and address standardization in the next two sections.

### *3.3. Name standardization and parsing*

Standardization consists of replacing various spelling of words with a single spelling. For instance, different spellings and abbreviations of “Incorporated” might be replaced with the single standardized spelling “Inc.” The standardization component of software might separate a general string such as a complete name or address into words (i.e., sets of characters that are separated by spaces and other delimiters). Each word is then compared lookup tables to get standard spelling. The first half of the following table shows various commonly occurring words that are replaced by standardized spellings (given in capital letters). After standardization, the name string is parsed into components (second half of the following table) that can be compared (Table 7). The examples are produced by general name standardization software (Winkler 1993a) for the U.S. Census of Agriculture matching system. Because the software does well with business lists and person matching, it has been used for additional matching applications at the Census Bureau and other agencies. At present, it is not clear that there is any commercial software for name standardization. Promising new methods based on hidden Markov models (Borkar et al., 2001; Christen et al., 2002; Churches et al., 2002) may improve over the rule-based name standardization in Winkler (1993a). Although the methods clearly improve over more conventional address standardization methods (see following section) for difficult situations such as Asian or Indian addresses, they did not perform as well as more conventional methods of name standardization. Bilmes (1998) provides a tutorial on EM-type algorithms that show that hidden Markov methods are slight

Table 7
Examples of name parsing

| Standardized | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 1. | DR John J Smith MD | | | | | | | |
| 2. | Smith DRY FRM | | | | | | | |
| 3. | Smith & Son ENTP | | | | | | | |
| Parsed | | | | | | | | |
| | Pre | First | Mid | Last | Post1 | Post2 | Bus1 | Bus2 |
| 1. | DR | John | J | Smith | MD | | | |
| 2. | | | | Smith | | | DRY | FRM |
| 3. | | | | Smith | | Son | ENTP | |

Table 8
Examples of address parsing

Standardized

1. 16 W Main ST APT 16
2. RR 2 BX 215
3. Fuller Bldg Suite 405
4. 14588 HWY 16 W

Parsed

| | Pre2 | Hsnm | Stnm | RR | Box | Post1 | Post2 | Unit1 | Unit2 | Bldg |
|---|---|---|---|---|---|---|---|---|---|---|
| 1. | W | 16 | Main | | | ST | | 16 | | |
| 2. | | | | 2 | 215 | | | | | |
| 3. | | | | | | | | | 405 | Fuller |
| 4. | | 14588 | HWY | 16 | | | W | | | |

generalizations of the simplest EM methods. Among mathematical statisticians, hidden Markov is referred to as the Baum-Welsh algorithm.

### 3.4. *Address standardization and parsing*

Table 8 illustrates address standardization with a proprietary package developed by the Geography Division at the U.S. Census Bureau. In testing in 1994, the software significantly outperformed the best U.S. commercial packages in terms of standardization rates while producing comparably accurate standardizations. The first half of the table shows a few addresses that have been standardized. In standardization, commonly occurring words such as “Street” are replaced by an appropriate abbreviation such as “St” that can be considered a standard spelling that may account for some spelling errors. The second half of the table represents components of addresses produced by the parsing. The general software produces approximately 50 components. The general name and address standardization software that we make available with the matching software only outputs the most important components of the addresses.

### 3.5. *Summarizing comments on preprocessing*

Many files cannot be sufficiently preprocessed to clean-up much of the data. Examples include legacy files that contain considerable missing data such as date of birth and high typographical error rate in other fields. In situations of reasonably high-quality data, preprocessing can yield a greater improvement in matching efficacy than string comparators and “optimized” parameters. In some situations, 90% of the improvement in matching efficacy may be due to preprocessing. The results of Table 6 show that appropriate string comparators can yield greater improvement than better record linkage parameters.

## 4. More advanced methods

Nearly all of the recent record linkage work is in the computer science literature and is based on statistical models (Winkler 2006a). The three most important research problems involve: (1) estimating error rates, (2) adjusting statistical analyses of merged files for linkage error, and (3) speeding up the record linkage process.

### 4.1. Error rate estimation

With any matching project, we are concerned with false match rates among the set of pairs among designated matches above the cutoff score $T_\mu$ in (2) and the false nonmatch rates among designated nonmatches below the cutoff score $T_\lambda$ in (2). Very few matching projects estimate these rates although valid estimates are crucial to understanding the usefulness of any files obtained via the record linkage procedures. Sometimes reasonable upper bounds for the estimated error rates can be obtained via experienced practitioners and the error rates are validated during follow-up studies (Winkler 1995). If a moderately large amount of training data is available, then it may be possible to get valid estimates of the error rates.

If a small amount of training data is available, then it may be possible to get improved record linkage and good estimates of error rates. Larsen and Rubin (2001) combined small amounts of (labeled) training data with large amounts of unlabeled data to estimate error rates using an MCMC procedure. In machine learning (Winkler 2000), the procedures are referred to as *semisupervised learning*. In ordinary machine learning, the procedures to get parameters are "supervised" by the training data that is labeled with the true classes into which later records (or pairs) will be classified. Winkler (2002) also used semisupervised learning with a variant of the general EM algorithm. Both the Larsen and Rubin (2001) and Winkler (2002) methods were effective because they accounted for interactions between the fields and were able to use labeled training data that was concentrated between the lower cutoff $T_\lambda$ and the upper cutoff $T_\mu$.

Belin and Rubin (1995) were the first to provide an unsupervised method for obtaining estimates of false match rates. The method proceeded by estimating Box-Cox transforms that would cause a mixture of two transformed normal distributions to closely approximate two well separated curves such as given in Fig. 1. They cautioned that their methods might not be robust to matching situations. Winkler (1995) observed that their algorithms would typically not work with business lists, agriculture lists, and low-quality person lists where the curves of nonmatches were not well separated from the curves of matches. Scheuren and Winkler (1993), who had the Belin–Rubin EM-based fitting software, observed that the Belin–Rubin methods did work reasonably well with a number of well-separated person lists.

Because the EM-based methods of this section serve as a template of other EM-based methods, we provide details of the unsupervised learning methods of Winkler (2006b) that are used for estimating false match rates. The basic model is that of semisupervised learning in which we combine a small proportion of labeled (true or pseudotrue matching status) pairs of records with a very large amount of unlabeled data. The CI model corresponds to the naive Bayesian network formulization of Nigam et al. (2000). The more general formulization of Winkler (2000, 2002) allows interactions between agreements (but is not used in this chapter).

Our development is similar theoretically to that of Nigam et al. (2000). The notation differs very slightly because it deals more with the representational framework of record linkage. Let $\gamma_i$ be the agreement pattern associated with pair $p_i$. Classes $C_j$ are an arbitrary partition of the set of pairs $D$ in $\mathbf{A} \times \mathbf{B}$. Later, we will assume that some of the $C_j$ will be subsets of $M$ and the remaining $C_j$ are subsets of $U$. For coherence and clarity Eqs. (12) and (13) repeat the earlier equations but use slightly different notation. Unlike general text classification in which every document may have a unique agreement pattern, in record linkage, some agreement patterns $\gamma_i$ may have many pairs $p_{i(l)}$ associated with them. Specifically,

$$P(\gamma_i|\Theta) = \sum_{i}^{|C|} P(\gamma_i|C_j;\Theta)P(C_j;\Theta) \tag{12}$$

where $\gamma_i$ is a specific pair, $C_j$ is a specific class, and the sum is over the set of classes. Under the Naive Bayes or CI, we have

$$P(\gamma_i|C_j;\Theta) = \Pi_k P(\gamma_{i,k}|C_j;\Theta) \tag{13}$$

where the product is over the $k$th individual field agreement $\gamma_{ik}$ in pair agreement pattern $\gamma_i$. In some situations, we use a Dirichlet prior

$$P(\Theta) = \Pi_j(\Theta_{Cj})^{\alpha-1}\Pi_k(\Theta\gamma_{i,k|Cj})^{\alpha-1} \tag{14}$$

where the first product is over the classes $C_j$ and the second product is over the fields. We use Du to denote unlabeled pairs and Dl to denote labeled pairs. Given the set $D$ of all labeled and unlabeled pairs, the log likelihood is given by

$$\begin{aligned} l_c(\Theta|D;z) = &\log(P(\Theta)) \\ &+(1-\lambda)\sum_{i\in \mathrm{Du}}\sum_{j} z_{ij}\log(P(\gamma_i|C_j;\Theta)P(C_j;\Theta)) \\ &+\lambda\sum_{i\in \mathrm{Dl}}\sum_{j} z_{ij}\log(P(\gamma_i|C_j;\Theta)P(C_j;\Theta)). \end{aligned} \tag{15}$$

where $0 \le \lambda \le 1$. The first sum is over the unlabeled pairs and the second sum is over the labeled pairs. In the third terms Eq. (15), we sum over the observed $z_{ij}$. In the second term, we put in expected values for the $z_{ij}$ based on the initial estimates P($\gamma_i|C_j;\Theta$) and P($C_j;\Theta$). After reestimating the parameters $P(\gamma_i|C_j;\Theta)$ and P($C_j;\Theta$)) during the M-step (that is in closed form under condition (CI)), we put in new expected values and repeat the M-step. The computer algorithms are easily monitored by checking that the likelihood increases after each combination of E- and M-steps and by checking that the sum of the probabilities add to 1.0. We observe that if λ is 1, then we only use training data and our methods correspond to naive Bayes methods in which training data are available. If λ is 0, then we are in the unsupervised learning situations of Winkler (1993b). Winkler (2000, 2002) provides more details of the computational algorithms.

We create "pseudotruth" data sets in which matches are those unlabeled pairs above a certain high cutoff and nonmatches are those unlabeled pairs below a certain low cutoff. Figure 1 illustrates the situation using actual 1990 Decennial Census data in which we plot log of the probability ratio (1) against the log of frequency. With the datasets of this chapter, we choose high and low cutoffs in a similar manner so that we do not include in-between pairs in our designated pseudotruth data sets. We use these "designated" pseudotruth data sets in a semisupervised learning procedure that is nearly identical to the semisupervised procedure where we have actual truth data. A key

Table 9
Pseudotruth data with actual error rates

| | Matches | Nonmatches | Other |
|---|---|---|---|
| **A x B** pairs | 8817 (0.008) | 98257 (0.001) | 9231 (0.136) |

difference from the corresponding procedure with actual truth data is that the sample of labeled pairs is concentrated in the difficult-to-classify in-between region where, in the pseudotruth situation, we have no way to designate comparable labeled pairs. The sizes of the pseudotruth data is given in Table 9. The errors associated with the artificial pseudotruth are given in parentheses following the counts. The *Other* class gives counts of the pairs and proportions of true matches that are not included in the pseudotruth set of pairs. In the *Other* class, the proportions of matches vary somewhat and would be difficult to determine without training data.

We determine how accurately we can estimate the lower cumulative distributions of matches and the upper cumulative distribution of nonmatches. This corresponds to the overlap region of the curves of matches and nonmatches. If we can accurately estimate these two tails of distributions, then we can accurately estimate error rates at differing levels. Our comparisons consist of a set of figures in which we compare a plot of the cumulative distribution of estimates of matches versus the true cumulative distribution with the truth represented by the 45° line. We also do this for nonmatches. As the plots get closer to the 45° lines, the estimates get closer to the truth.

Our primary results are from using the CI model and "semisupervised" methods of this chapter with the CI model and actual semisupervised methods of Winkler (2002). With our pseudotruth data, we obtain the best sets of estimates of the bottom 30% tails of the curve of matches and the top 5% tails of nonmatches with CI and $\lambda = 0.2$. Figures 2A and 2B illustrate the set of curves that provide quite accurate fits. The 45° line represents the truth whereas the curve represents the cumulative estimates of matches and nonmatches for the left and right tails, respectively. Although we looked at results for $\lambda = 0.1, 0.5$, and 0.8 and various interactions models, the results under CI were the best with $\lambda = 0.2$. We also looked at several different ways of constructing the pseudotruth data. Additionally, we considered other pairs of files in which all of the error-rates estimates were better (closer to the 45° line) than those for the pair of files given in Fig. 2a.

We can use the model given in this section (essentially the same as in Winkler 2000; 2002) and the associated EM software to obtain all of the EM estimates that are used in this chapter. In each situation, the inputs will vary significantly.

### 4.2. *Adjusting analyses for linkage error*

Adjusting statistical analyses for linkage error is clearly a statistical problem. We briefly provide background and describe issues. We wish to match two files $A = (a_{ij})$ and $B = (b_{kl})$ using name, address, and other quasi-identifying information. We would like to examine joint relationships between a-variables and b-variables on $A \cap B$. We can examine the joint relationships if matching error is very low or we have a model for adjusting for matching error. We consider the simplest situation of ordinary regression

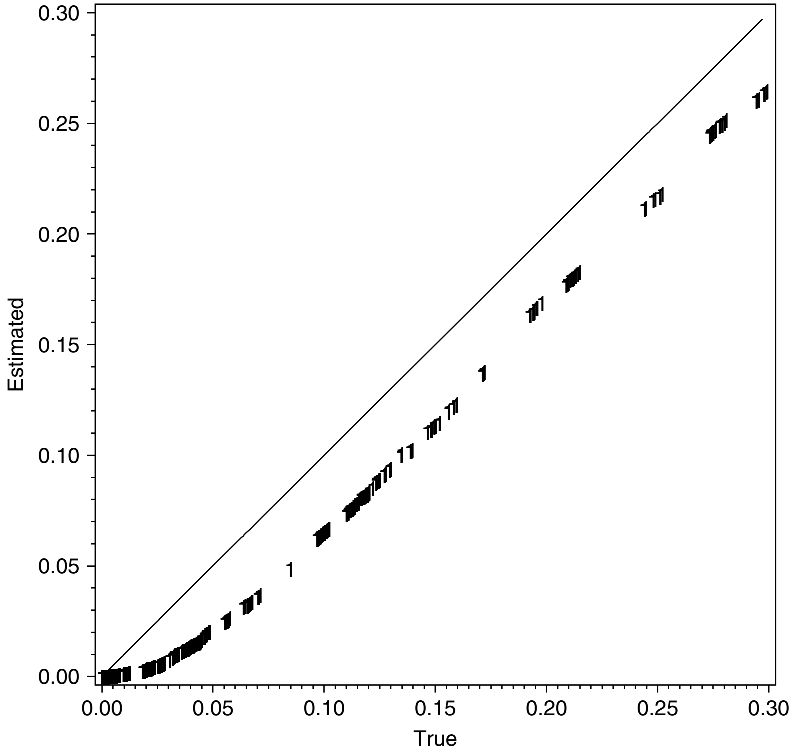


Fig. 2a. Estimates versus truth, file A cumulative matches, tail of distribution independent EM, lambda = 0.2.

where one file provides an independent $x$-variable and the other file provides the dependent $y$-variable associated with the model

$$Y = \beta X + \varepsilon \tag{16}$$

where $\varepsilon$ is suitable normal noise with mean 0 and constant variance $\sigma^2$. We assume that we use all $n$ A records and all pairs to which it can be linked. We wish to use $(X_i, Y_i)$ but must use $(X_i, Z_i)$ where $Z_i$ is the observed $y$-variable that may or may not be from the correct B record. For $i = 1, \ldots, n$,

$$Z_i = \begin{cases} Y_i \text{ with probability } p_i, \\ Y_j \text{ with probability } q_{ij} \text{ for } j \neq i, \end{cases}$$

where $p_i + \Sigma_j q_{ij} = 1$.

The probability $p_i$ of matching the correct record may be 0 or 1. We define $h_i = 1 - p_i$ and divide the set of pairs into $n$ mutually exclusive classes. The classes are determined by the records from one of the files. Each class consists of the independent $x$-variable $X_i$, the true value of the dependent $y$-variable, the values of the $y$-variables from the second file to which the record in the first file containing $X_i$ have been paired, and the

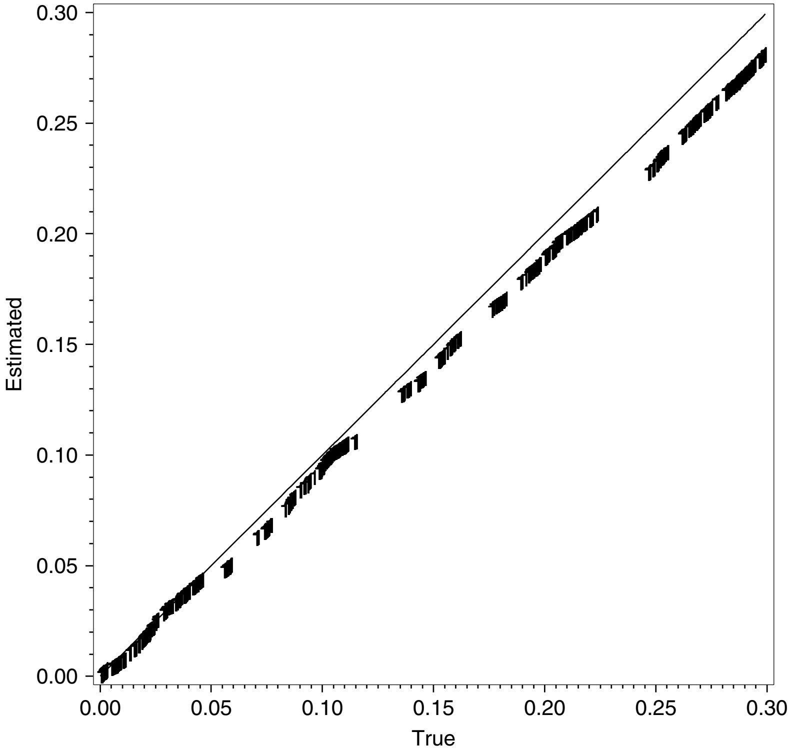


Fig. 2b. Estimates versus truth, file A cumulative nonmatches, tail of distribution independent EM, lambda = 0.2.

computer matching weights (scores). Under an assumption of one-to-one matching, for each $i = 1, \ldots, n$, there exists at most one $j$ such that $q_{ij} > 0$. We let $\varphi$ be define by $\varphi(i) = j$.

Under the model, we observe

$$\begin{aligned} E(Z) &= (1/n)\Sigma_i E(Z|i) = (1/n)\Sigma_i(Y_i p_i + \Sigma_j Y_j q_{ij}) \\ &= (1/n)\Sigma_i Y_i + (1/n)\Sigma_i(Y_i(-h_i) + Y_{\varphi(i)} h_i) \\ &= \overline{Y} + B. \end{aligned} \tag{17}$$

As each $X_i, i = 1, \ldots, n$, can be paired with either $Y_i$ or $Y_{\varphi(i)}$, the second equality in (17) represents $2n$ points. Similarly, we can represent $\sigma_{zy}$ in terms of $\sigma_{xy}$ and a bias term $B_{xy}$ and $\sigma_z^2$ in terms of $\sigma_y^2$ and a bias terms $B_{yy}$. We neither assume that the bias terms have expectation zero nor that they are uncorrelated with the observed data. The advantage of the adjustments using equations like (17) is that the estimated coefficient $\hat{\beta}_{xy}$ between $y$ and $x$ is increased to a value close to the true value $\beta$.

In the simulations of Scheuren and Winkler (1993), having reasonable estimates of the probabilities $p_i$ and $q_{ij}$ from the methods and software of Belin and Rubin (1995) were crucial to the application. Scheuren and Winkler showed that their adjustment methods could work well in a few situations when the probabilities $p_i$ and $q_{ij}$ were known

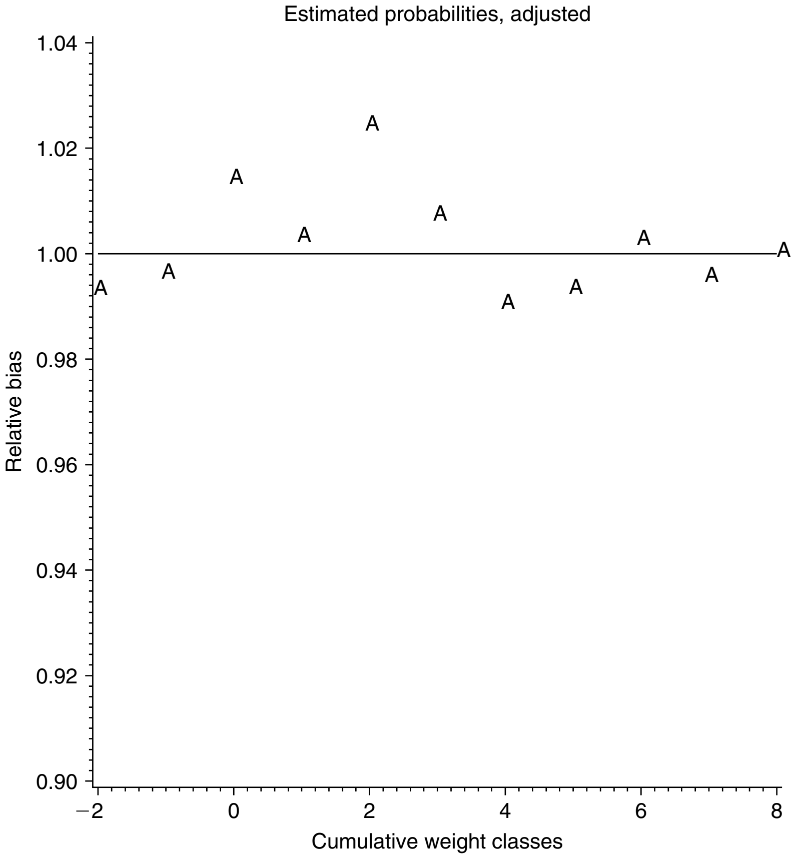


Fig. 3a. Relative bias of beta coefficients using the Scheuren–Winkler adjustment procedure.

and noted that, in many situations, the methods of Belin and Rubin could not supply reasonable estimates. Figure 3 illustrates the relative bias of the regression coefficient (i.e., relative bias $= \hat{\beta}_{xy}/\beta$) using the adjustment procedures and with the observed data, respectively. The quasi-identifying data are similar to the data of the example of Section 2.4. The curves of nonmatches and matches, however, are much closer together than with the curves of Fig. 1. Each plot uses all points above a certain matching weight. As we move from right to left, more pairs are associated with each plot. As the matching error increases with lower matching weights, we expect the plots from the observed data to show progressively more bias. Although we do not show the corresponding plots when the true probabilities are used, they are similar to those given in Fig. 3. The similarity indicates the reasonable quality of the estimates of the probabilities $p_i$ and $q_{ij}$ given by the Belin–Rubin procedure.

Lahiri and Larsen (2005) extended the methods of Scheuren and Winkler under the assumption that the values of the probabilities $p_i$ and $q_{ij}$ were known. In their simulations, they were able to show that their methods substantially outperformed the methods of Scheuren and Winkler. A crucial issue related to applying the methods of Lahiri and

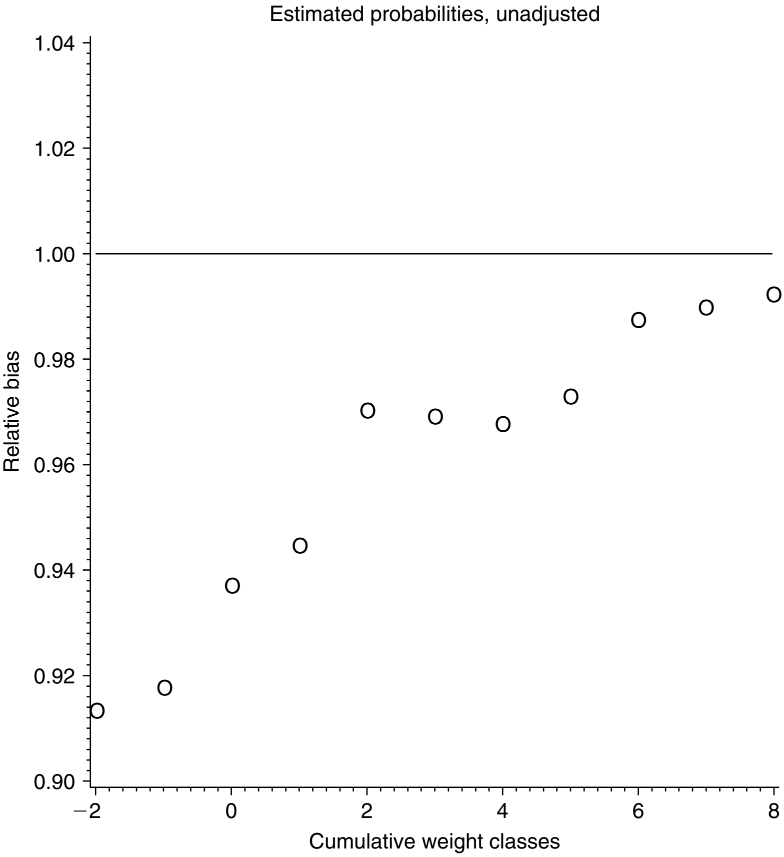


Fig. 3b. Relative bias of beta coefficients without adjustment.

Larsen (2005) and Scheuren and Winkler (1993) is having suitable methods of estimating the probabilities $p_i$ and $q_{ij}$. Winkler (2006b) provides an alternative method for estimating the probabilities $p_i$ and $q_{ij}$ that should hold in more situations than the method of Belin and Rubin (1995). None of the methods for estimating the probabilities $p_i$ and $q_{ij}$ will work in situations where the curves of matches and nonmatches (analogous to Fig. 1) overlap substantially. The curves overlap with poor quality person lists and with almost all agriculture and business lists.

Scheuren and Winkler (1997) provided methods for both improving the matching (i.e., causing the curves of matches and nonmatches to pull apart) in a situation similar to that of Scheuren and Winkler (1993). In their situation, the name and address matching had far more typographical error than in the earlier work and curves similar to Fig. 1 almost totally overlapped. Using high-weight pairs from the initial match (approximately 0.5% of all pairs above a certain matching weight), they were able to get an initial guess $y = \hat{\beta}x$ for the relationship between $x$ and $y$ using the ideas of Scheuren and Winkler (1993). With this initial guess for the regression model, they created a new variable

$\text{pred}(y) = \hat{\beta}x$ that they put in the file containing the $x$-variable. They then developed an additional metric for comparing the $y$-variable with the predicted $y$-variable pred($y$) that they added to the matching software. When they repeated the matching with name, address, and ($y$, pred($y$)), they observed a substantial improvement.

Figure 4 illustrates the improvement. In the first matching pass, determining the regression relationship with all pairs above a certain point is effectively impossible. The false matches (given by "o") overwhelm the true matches (given by "*"). After addition of ($y$, pred($y$)) comparison, the matching improves as shown in Fig. 4B where there is substantially smaller proportion of false matches. The true underlying beta coefficient is approximately 5.

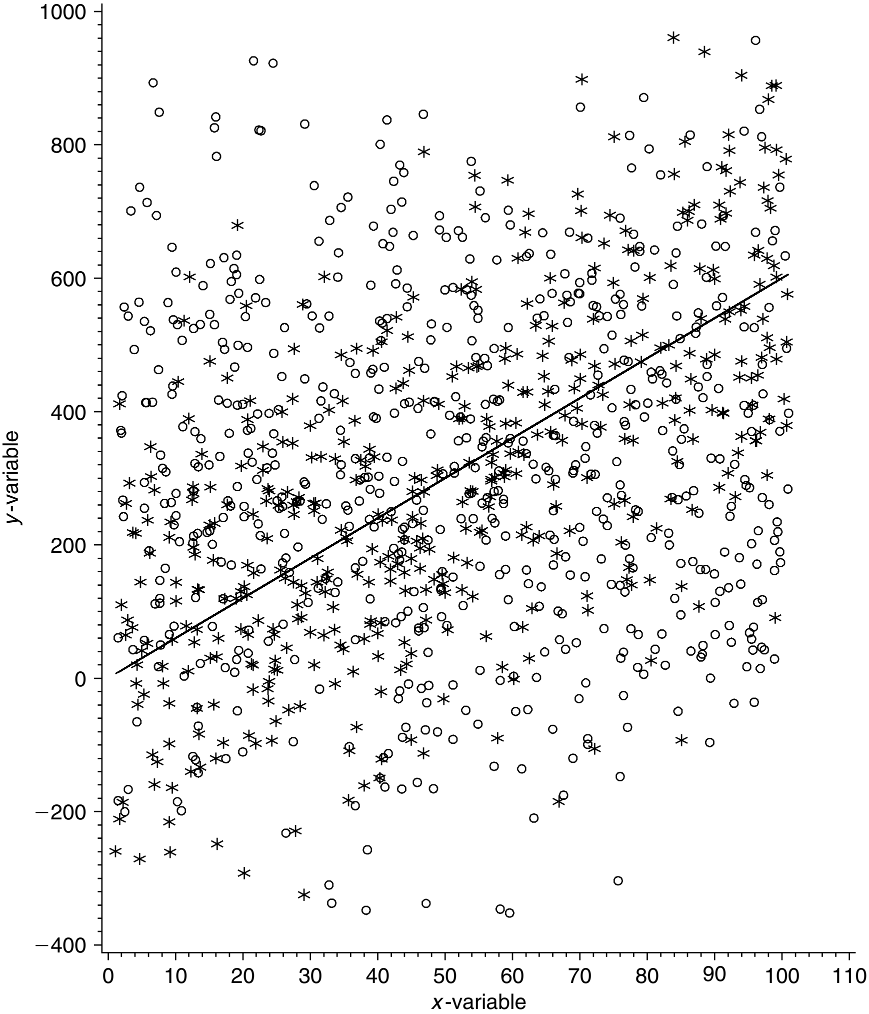


Fig. 4a. (A) Poor matching scenario, 1st pass. All false & 5% true matches, observed data, highoverlap 1104 points, beta = 2.47, R-square = 0.07.

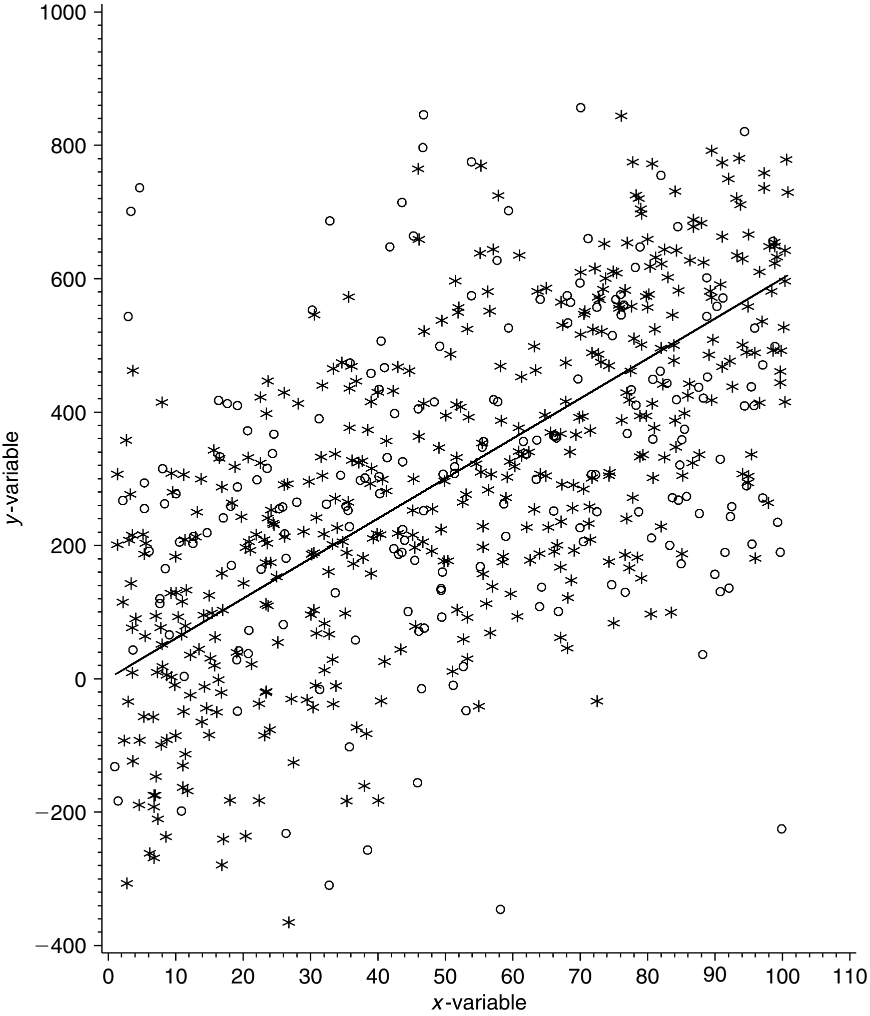


Fig. 4b. (B) Poor matching scenario, 2nd pass. All false & 5% true matches, observed data, highoverlap 650 points, beta = 4.75, R-square = 0.33.

### 4.3. *Speeding up record linkage*

The issue of speed affects larger matching problems in which files with 10 million or more records are compared. At present, the fastest methods are BigMatch (Winkler et al., 2008) that have been production tested for use in the 2010 Decennial Census. Details of the BigMatch technology are given in Winkler (2006a) or Yancey (2004). There are two key improvements. First, the larger file of the two files being matched is never sorted. In traditional record linkage, pairs of files are successively sorted and matched according to different blocking criteria. If 10 blocking passes are used, then BigMatch eliminates 10 sorts of the larger file. Second, BigMatch uses a very efficient retrieval/comparison mechanism for comparing pairs of records according to the different blocking criteria.

The retrieval/comparison mechanism is three times as fast as inverted index procedures that are widely used in computer science.

BigMatch software matches $10^{17}$ pairs (300 million × 300 million) records in 63 hours using 40 processors on a 64-processor SGI Linux machine. Using seven blocking criteria, the software performs detailed computation on only $10^{12}$ pairs. Winkler (2004) provides methods for estimating the number of matches that are missed by a set of blocking criteria. BigMatch software is 40–50 times as fast as recent parallel software (Kawai et al., 2006; Kim and Lee, 2007), at least 10 times as fast as other sequential software being researched (e.g., Chaudhuri et al., 2003), and possibly 80 times as fast as commercial Vality-suite software from IBM. It nearly maintains the accuracy of earlier software (Winkler and Thibaudeau, 1991).

## 5. Concluding remarks

This chapter covers modern computerized record linkage procedures that are used for removing duplicates from lists and for improving coverage in a list by updating it with external lists. It is somewhat remarkable how much these modern methods improve accuracy of lists in comparison to ad hoc methods that are still in wide-spread use. As observed by Herzog et al. (2007) and others, duplication and lack of coverage in the list frame can bias estimates (both from sampling and censuses) more than any other source of error.

*Sample Surveys: Design, Methods and Applications, Vol. 29A*
ISSN: 0169-7161
© 2009 Elsevier B.V. All rights reserved
DOI: 10.1016/S0169-7161(08)00015-1

15

# Statistical Disclosure Control for Survey Data

*Chris Skinner*

## 1. Introduction

### *1.1. The problem of statistical disclosure control*

Survey respondents are usually provided with an assurance that their responses will be treated confidentially. These assurances may relate to the way their responses will be handled within the agency conducting the survey or they may relate to the nature of the statistical outputs of the survey as, for example, in the "confidentiality guarantee" in the United Kingdom (U.K.) National Statistics Code of Practice (National Statistics, 2004, p. 7) that "no statistics will be produced that are likely to identify an individual." This chapter is concerned with methods for ensuring that the latter kinds of assurances are met. Thus, in the context of this chapter, *statistical disclosure control* (SDC) refers to the methodology used, in the design of the statistical outputs from the survey, for protecting the confidentiality of respondents' answers. Methods relating to the first kind of assurance, for example, computer security and staff protocols for the management of data within the survey agency, fall outside the scope of this chapter.

There are various kinds of *statistical outputs* from surveys. The most traditional are tables of descriptive estimates, such as totals, means, and proportions. The release of such estimates from surveys of households and individuals have typically not been considered to represent a major threat to confidentiality, in particular because of the protection provided by sampling. Tabular outputs from the kinds of establishment surveys conducted by government have, however, long been deemed risky, especially because of the threat of disclosure of information about large businesses in cells of tables which are sampled with a 100% sampling fraction. SDC methods for such tables have a long history and will be outlined in Section 2.

Although the traditional model of delivering all the estimates from a survey in a single report continues to meet certain needs, there has been increasing demand for more flexible survey outputs, often for multiple users, where the set of population parameters of interest is not prespecified. There are several reasons why it may not be possible to prespecify all the parameters. Data analysis is an iterative process, and what analyses are of most interest may only become clear after initial exploratory analyses of the data. Moreover, given the considerable expense of running surveys, it is natural for many commissioners of surveys to seek to facilitate the use of the data by multiple users. But it is usually impossible to prespecify all possible users and their needs in advance. A natural way to provide flexible outputs from a survey to address such needs is to make the survey microdata available, so that users can carry out the statistical analyses that interest them.

However, the release of such microdata raises serious confidentiality protection issues. Of course, statistical analyses of survey data do not require that the identities of the survey units are known. Names, addresses, and contact information for individuals or establishment can be stripped from the data to form an *anonymized* microdata file. The problem, however, is that such basic anonymization is often insufficient to protect confidentiality, and therefore, it is necessary to use one of a range of alternative approaches to SDC and this will be discussed further in Section 3.

### *1.2. Concepts of confidentiality, disclosure, and disclosure risk*

To be precise about what is meant by "protecting confidentiality" requires discussion of definitions. These usually involve the notion of a hypothetical *intruder* who might seek to breach confidentiality. There are thus three key parties: (1) the *respondent* who provides the data, (2) the *agency* which collects the data, releases statistical outputs, and designs the SDC strategy, and (3) the hypothetical *intruder* who has access to these outputs and seeks to use them to disclose information about the respondent. One important notion of disclosure is *identity disclosure* or *identification,* which would occur if the intruder linked a known individual (or other unit) to an individual microdata record or other element of the statistical output. Another important notion is *attribute disclosure*, which would occur if the intruder could determine the value of some survey variable for an identified individual (or other unit) using the statistical output. More generally, *prediction disclosure* would occur if the intruder could predict the value of some survey variable for an identified individual with some uncertainty. When assessing the potential for disclosure for a particular statistical output, it is usual to refer to the *disclosure risk.* This might be defined as the probability of disclosure with respect to specified sources of uncertainty. Or the term might be used loosely to emphasize not only the uncertainty about potential disclosure but also the potential harm that might arise from disclosure (Lambert, 1993). The *confidentiality* of the answers provided by a respondent might be said to be protected if the disclosure risk for this respondent and the respondent's answers is sufficiently low. In this chapter, disclosure risk is discussed in more detail in Sections 2 and 3. For further discussion of definitions of disclosure, see Duncan and Lambert (1986, 1989) and Skinner (1992).

### *1.3. Approaches to protecting confidentiality*

If the disclosure risk is not deemed to be sufficiently low, then it will be necessary to use some method to reduce the risk. There are broadly two approaches, which are referred to here as *safe setting* and *safe data* (Marsh et al., 1994). The safe setting approach imposes restrictions on the set of possible users of the statistical output and/or on the ways that the output can be used. For example, users might be required to sign a licensing agreement or might only be able to access microdata by visiting a secure laboratory or by submitting requests remotely (National Research Council, 2005). The safe data approach, on the other hand, involves some modification to the statistical output. For example, the degree of geographical detail in a microdata file from a national social survey might be limited so that no area containing less than 100,000 households is identified. In this chapter, we focus on the safe data approach and generally refer to methods for modifying the statistical output as SDC methods.

### *1.4. SDC methods, utility, and data quality*

SDC methods vary according to the form of the statistical output. Some simple approaches are as follows:

- *Reduction of detail*, for example, the number of categories of a categorical variable might be reduced in a cross-classified table or in microdata.
- *Suppression*, for example, the entry in a table might be replaced by an asterisk, indicating that the entry has been suppressed for confidentiality reasons.

In each of these cases, the SDC method will lead to some *loss of information* for the user of the statistical output. Thus, the method will reduce the number of population parameters for which a user can obtain survey estimates. Other kinds of SDC methods might not affect the number of parameters which can be estimated but may affect the *quality* of the estimates that can be produced. For example, if *random noise* is added to an income variable to protect confidentiality, then this may induce bias or variance inflation in associated survey estimates. The general term *utility* may be used to cover both the information provided by the statistical outputs, for example, the range of estimates or analyses which can be produced, and the quality of this information, for example, the extent of errors in these estimates. It should, of course, be recognized that survey data are subject to many sources of error, even prior to the application of SDC methods, and the impact of SDC methods on data quality therefore needs to be considered in this context.

Generally, utility needs to be considered from the perspective of a *user* of the statistical outputs, who represents a key fourth party to add to the three parties referred to earlier: the respondent, the agency, and the intruder.

### *1.5. SDC as an optimization problem: the risk-utility trade-off*

The key challenge in SDC is how to deal with the trade-off between disclosure risk and utility. In general, the more the disclosure risk is reduced by an SDC method, the lower will be the expected utility of the output. This trade-off may be formulated as an optimization problem. Let $D$ be the (anonymized) survey data and let $f(D)$ be the statistical output, resulting from the use of an SDC method. Let $R[f(D)]$ be a measure of the disclosure risk of the output, and let $U[f(D)]$ be a measure of the utility of the output. Then, the basic challenge of SDC might be represented as the constrained optimization problem:

$$\text{for given } D \text{ and } \varepsilon\text{, find an SDC method, } f(.)\text{, which}$$
$$\text{maximizes } U[f(D)]\text{, subject to } R[f(D)] < \varepsilon.$$

The elements of this problem need some clarification:

$f(.)$ : the *SDC method*—a wide variety of these have been proposed and we shall refer to some of these in this chapter;

$R(.)$ : the *disclosure risk function*—we shall discuss ways in which this function may be defined; this is certainly not straightforward, for example, because of its dependence on assumptions about the intruder and because of the challenge of combining the threats of disclosure for multiple respondents into a scalar function;

$U(.)$ : the *utility function*—this will also not be straightforward to specify as a scalar function, given the potential multiple uses of the output;

$\varepsilon$ : the *maximum acceptable risk*—in principle, one might expect the agency to provide this value in the light of its assurances to respondents. However, in practice, agencies find it very difficult to specify a value of $\varepsilon$, other than zero, that is, no disclosure risk. Unfortunately, for most definitions of disclosure risk, the only way to achieve no disclosure risk is by not releasing any output and this is rarely a solution of interest!

Given these difficulties in specifying $R(.)$ and $U(.)$ as scalar functions and in specifying a value for $\varepsilon$, the above optimization problem serves mainly as conceptual motivation. In practice, different SDC methods can be evaluated and compared by considering the values of alternative measures of risk and utility. For given measures of each, it can sometimes be useful to construct an RU map (Duncan et al., 2001), where a measure of risk is plotted against a measure of utility for a set of candidate SDC methods. The points on this map are expected to display a general positive relationship between risk and utility, but one might still find that, for given values of risk, some methods have greater utility than others and thus are to be preferred. This approach avoids having to assume a single value of $\varepsilon$.

## 2. Tabular outputs

### *2.1. Disclosure risk in social surveys and the protection provided by sampling*

The main developments in SDC methods for tabular outputs have been motivated by the potential risks of disclosure arising when 100% sampling has been used, such as in censuses or in administrative data. Frequency tables based upon such data sources may often include small counts, as low as zero or one, for example, in tables of numbers of deaths by area by cause of death. Such tables might lead to identity disclosure, for example, if it is public knowledge that someone has died, then it might be possible to identify that person as a count of one in a table of deaths using some known characteristics of that person. Attribute disclosure might also occur. For example, it might be possible to find out the cause of the person's death if the table cross-classifies this cause by other variables potentially known to an intruder.

In social surveys, however, the use of sampling greatly reduces the risks of such kinds of disclosure for two reasons. First, the presence of sampling requires different kinds of statistical outputs. Thus, the entries in tables for categorical variables tend to be weighted proportions (possibly within domains defined by rows or columns) and not unweighted sample counts. Even if a user of the table could work out the cell counts (e.g., because the survey uses equal weights and the sample base has been provided), the survey agency will often ensure that the published cells do not contain very small counts, where the estimates would be deemed too unreliable due to sampling error. For example, the agency might suppress cell entries where the sample count in the cell falls below some threshold, for example, 50 persons in a national social survey. This should prevent the kinds of situations of most concern with 100% data. Sometimes, agencies use techniques of small area estimation (see Chapters 31 and 32) in domains with small sample counts and these techniques may also act to reduce disclosure risk.

Second, the presence of sampling should reduce the precision with which an intruder could achieve predictive disclosure. For example, suppose that an intruder could find out from a survey table that, among 100 respondents falling into a certain domain, 99 of them have a certain attribute and suppose that the intruder knows someone in the population who falls into this domain. Then, the intruder cannot predict that this person has the attribute with probability 0.99, since this person need not be a respondent and prediction is subject to sampling uncertainty. This conclusion depends, however, on the identities of the survey respondents being kept confidential by the agency, preventing the intruder knowing whether the known person is a respondent, referred to as *response knowledge* by Bethlehem et al. (1990). In general, it seems very important that agencies do adopt this practice since it greatly reduces disclosure risk while not affecting the statistical utility of the outputs. In some exceptional cases, it may be difficult to achieve this completely. For example, in a survey of children it will usually be necessary to obtain the consent of a child's parent (or other adult) in order for the child to take part in the survey. The child might be assured that their responses will be kept confidential from their parent. However, when examining the outputs of the survey, the parent (as intruder) would know that their child was a respondent.

For the reasons given above, disclosure will not generally be of concern in the release of tables of estimates from social surveys, where the sample inclusion probabilities are small (say never exceeding 0.1). See also Federal Committee on Statistical Methodology (2005, pp. 12–14).

## *2.2. Disclosure risk in establishment surveys*

A common form of output from an establishment survey consists of a table of estimated totals, cross-classified by characteristics of the establishment. Each estimate takes the form $\hat{Y}_c = \sum_s w_i I_{ci} y_i$, where $w_i$ is the survey weight, $I_{ci}$ is a 0–1 indicator for cell $c$ in the cross-classification, and $y_i$ is the survey variable for the $i$th establishment in the sample $s$. For example, $y_i$ might be a measure of output and the cells might be formed by cross-classifying industrial activity and a measure of size.

The relevant definition of disclosure in such a setting will often be a form of prediction disclosure. Prediction disclosure for a specific cell $c$ might be defined under the following set-up and assumptions:

- the intruder is one of the establishments in the cell which has the aim of predicting the value $y_i$ for one of the other establishments in the cell or, more generally, the intruder consists of a *coalition* of $m$ of the $N_c$ establishments in the cell with the same predictive aim;

- the intruder knows the identities of all establishments within the cell (since, e.g., they might represent businesses competing in a similar market).

Given such assumptions, prediction disclosure might be said to occur if the intruder is able to predict the value $y_i$ with a specified degree of precision. To clarify the notion of precision, we focus in the next subsection on the important case where the units in the cell all fall within completely enumerated strata. Thus, $w_i = 1$ when $I_{ci} = 1$ so that $\hat{Y}_c = \sum_{U_c} y_i$, where $U_c$ is the set of all establishments in cell $c$ and $N_c$ is the size of $U_c$. In this case, the intruder faces no uncertainty due to sampling and this might, therefore, be treated as the worst case.

### 2.2.1. *Prediction disclosure in the absence of sampling*

In the absence of sampling, prediction is normally considered from a deterministic perspective and is represented by an interval (between an upper and lower bound) within which the intruder knows that a value $y_i$ must lie. The precision of prediction is represented by the difference between the true value and one of the bounds. It is supposed that the intruder undertakes prediction by combining prior information with the reported value $\hat{Y}_c$.

One approach to specifying the prior information is used in the *prior-posterior rule* (Willenborg and de Waal, 2001), also called the *pq* rule, which depends upon two constants, $p$ and $q$, set by the agency. The constant $q$ is used to specify the precision of prediction based upon the prior information alone. Under the *pq* rule, it is assumed that intruder can infer the $y_i$ value for each establishment in the cell to within $q\%$. Thus, the agency assumes that, prior to the table being published, the intruder could know that a value $y_i$ falls within the interval $[(1 - q/100)y_i, (1 + q/100)y_i]$. The combination of this prior information with the output $\hat{Y}_c = \sum_{U_c} y_i$ can then be used by the intruder to obtain sharper bounds on a true value. For example, let $y_{(1)} \leq y_{(2)} \leq \ldots \leq y_{(N_c)}$ be the order statistics and suppose that the intruder is the establishment with the second largest value, $y_{(N_c-1)}$. Then, this intruder can determine an upper bound for the largest value $y_{(N_c)}$ by subtracting its own value $y_{(N_c-1)}$ together with the sum of the lower bounds for $y_{(1)}, \ldots, y_{(N_c-2)}$ from $\hat{Y}_c$. The precision of prediction using this upper bound is given by the difference between this upper bound and the true value $y_{(N_c)}$, which is $(q/100) \sum_{i=1}^{N_c-2} y_{(i)}$. This cell would be called *sensitive* under the *pq* rule, that is, judged *disclosive*, if this difference was less than $p\%$ of the true value, that is, if

$$(p/100)y_{(N_c)} - (q/100) \sum_{i=1}^{N_c-2} y_{(i)} > 0. \tag{1}$$

The expression on the left-hand side of (1) is a special case of a *linear sensitivity measure*, which more generally takes the form $R_c = \sum_{i=1}^{N_c} a_i y_{(i)}$, where the $a_i$ are specified weights. The cell is said to be sensitive if $R_c > 0$. In this case, prediction disclosure would be deemed to occur. A widely used special case of the *pq* rule is the *p% rule*, which arises from setting $q = 100$, that is, no prior information is assumed. Another commonly used linear sensitivity measure arises with the $(n, k)$ or *dominance rule*. See Willenborg and de Waal (2001), Cox (2001), Giessing (2001), and Federal Committee on Statistical Methodology (2005) for further discussion.

### *2.2.2. Prediction disclosure in the presence of sampling*

More generally, all cell units may not be completely enumerated. In this case, $\hat{Y}_c$ will be subject to sampling error and, in general, this will lead to additional disclosure protection, provided that the intruder does not know whether other establishments (other than those in the coalition) are sampled or not. The definition of risk in this setting appears to need further research. Willenborg and de Waal (2001, Section 6.2.5) presented some ideas. An alternative model-based stochastic approach might assume that before the release of the table, the prior information about the $y_i$ can be represented by a linear regression model depending upon publicly available covariate values $x_i$ with a specified residual variance. The predictive distribution of $y_i$ given $x_i$ could then be updated using the known value(s) of $y_i$ for the intruder and the reported $\hat{Y}_c$, which might be assumed to follow the distribution $\hat{Y}_c \sim N[Y_c, v(\hat{Y}_c)]$, where $v(\hat{Y}_c)$ is the reported variance estimate of $\hat{Y}_c$. Prediction disclosure could then be measured in terms of the resulting residual variance in the prediction of $y_i$.

## *2.3. SDC methods for tabular outputs*

If a cell in a table is deemed sensitive, that is, the cell value represents an unacceptably high disclosure risk, a number of SDC approaches may be used.

### *2.3.1. Redefinition of cells*

The cells are redefined to remove sensitive cells, for example, by combining sensitive cells with other cells or by combining categories of the cross-classified variables. This is also called *table redesign* (Willenborg and de Waal, 2001).

### *2.3.2. Cell suppression*

The value of a sensitive cell is suppressed. Depending upon the nature of the table and its published margins, it may also be necessary to suppress the values of "complementary" cells to prevent an intruder being able to deduce the value of the cell from other values in the table. There is a large literature on approaches to choosing complementary cells which ensure disclosure protection. See, for example, Willenborg and de Waal (2001), Cox (2001), and Giessing (2001) and references therein.

### *2.3.3. Cell modification*

The cell values may be modified in some way. It will generally be necessary to modify not only the values in the sensitive cells but also values in some complementary nonsensitive cells, for the same reason as in cell suppression. Modification may be deterministic, for example, Cox et al. (2004), or stochastic, for example, Willenborg and de Waal (2001, Section 9.2). A simple method is *rounding*, where the modified cell values are multiples of a given base integer (Willenborg and de Waal, 2001, Chapter 9). This method is more commonly applied to frequency tables derived from 100% data but can also be applied to tables of estimated totals from surveys, where the base integer may be chosen according to the magnitudes of the estimated totals. Instead of replacing the cell values by single safe values, it is also possible to replace the values by intervals, defined by lower and upper bounds (Salazar, 2003; Giessing and Dittrich, 2006). The method of *controlled tabular adjustment* (Cox et al., 2004) determines modified cell values within such bounds so that the table remains additive and certain safety and statistical properties are met.

### *2.3.4. Pretabular microdata modification*

Instead of modifying the cell values, the underlying microdata may be perturbed, for example, by adding noise, and then the table formed from the perturbed microdata (Evans et al., 1998; Massell et al., 2006).

The statistical output from a survey will typically include many tables. Although the above methods may be applied separately to each table, such an approach takes no account of the possible additional disclosure risks arising from the combination of information from different tables, in particular, from common margins. To protect against such additional risks raise new considerations for SDC. Moreover, the set of tables constituting the statistical output is not necessarily fixed, as in a traditional survey report. With developments in online dissemination, there is increasing demand for the generation of tables which can respond in a more flexible way to the needs of users. This implies the need to consider SDC methods which not only protect each table separately as above but also protect against the risk arising from alternative possible sequences of released tables (see, e.g., Dobra et al., 2003).

## 3. Microdata

### *3.1. Assessing disclosure risk*

We suppose the agency is considering releasing to researchers an anonymized microdata file, where the records of the file correspond to the basic analysis units and each record contains a series of survey variables. The record may also include identifiers for higher level analysis units, for example, household identifiers where the basic units are individuals, as well as information required for survey analysis such as survey weights and primary sampling unit (PSU) identifiers.

We suppose that the threat of concern is that an intruder may link a record in the file to some external data source of known units using some variables, which are included in both the microdata file and the external source. These variables are often called *key variables* or identifying variables. There are various ways of defining disclosure risk in this setting. See, for example, Paass (1988) and Duncan and Lambert (1989). A common approach, often motivated by the nature of the confidentiality pledge, is to consider a form of *identification risk* (Bethlehem et al., 1990; Reiter, 2005), concerned with the possibility that the intruder will be able to determine a correct link between a microdata record and a known unit. This definition of risk will only be appropriate if the records in the microdata can meaningfully be said to be associated with units in the population. When microdata is subject to some forms of SDC, this may not be the case (e.g., if the released records are obtained by combining original records) and in this case, it may be more appropriate to consider some definition of predictive disclosure (e.g., Fuller, 1993) although we do not pursue this further here.

A number of approaches to the assessment of identification risk are possible, but all depend importantly upon assumptions about the nature of the key variables. One approach is to conduct an empirical experiment, matching the proposed microdata against another data source, which is treated as a surrogate for the data source held by the intruder. Having made assumptions about the key variables, the agency can use record linkage methods (see Chapter 14), which it is plausible would be available to an intruder, to match units between the two data sets. Risk might then be measured in terms of the number of units for which matches are achieved together with a measure of the match quality (in terms of the proportions of false positives and negatives). Such an experiment, therefore, requires that the agency has information which enables it to establish precisely which units are in common between the two sources and which are not.

The key challenge in this approach is how to construct a realistic surrogate intruder data set, for which there is some overlap of units with the microdata and the nature of this overlap is known. On some occasions a suitable alternative data source may be available. Blien et al. (1992) provide one example of a data source listing people in certain occupations. Another possibility might be a different survey undertaken by the agency, although agencies often control samples to avoid such overlap. Even if there is overlap, say with a census, determining precisely which units are in common and which are not may be resource intensive. Thus, this approach is unlikely to be suitable for routine use.

In the absence of another data set, the agency may consider a reidentification experiment, in which the microdata file is matched against itself in a similar way, possibly after the application of some SDC method (Winkler, 2004). This approach has the advantage that it is not model-dependent, but it is possible that the reidentification risk is overestimated if the disclosure protection effects of sampling and measurement error are not allowed for in a realistic way.

In the remainder of Section 3, we consider a third approach, which again only requires data from the microdata file, but makes theoretical assumptions, especially of a modeling kind, to estimate identification risk. As for the reidentification experiment, this approach must make assumptions about how the key variables are measured in the microdata and by the intruder on known units using external information. A simplifying but "worst case" assumption is that the key variables are recorded in identical ways in the microdata and externally. We refer to this as the *no measurement error assumption*, since measurement error in either of the data sources may be expected to invalidate this assumption. If at least one of the key variables is continuous and the no measurement error assumption is made, then an intruder who observes an exact match between the values of the key variables in the microdata and on the known units could conclude with probability one that the match is correct, in other words, the identification risk would be one. If at least one of the key variables is continuous and it is supposed that measurement error may occur, then the risk will generally be below one. Moreover, an exact matching approach is not obviously sensible and a broader class of methods of record linkage might be considered. See Fuller (1993) for the assessment of disclosure risk under some measurement error model assumptions.

In practice, variables are rarely recorded in a continuous way in social survey microdata. For example, age would rarely be coded with more detail than 1 year bands. And from now on, we restrict attention to the case of categorical key variables. For simplicity, we restrict attention to the case of exact matching, although more general record linkage methods could be used. We focus on a microdata file, where the only SDC methods which have been applied are recoding of key variables or random (sub)sampling. We comment briefly on the impact of other SDC methods on risk in Section 3.4.

### *3.2. File-level measures of identification risk*

We consider a finite population $U$ of $N$ units (which will typically be individuals) and suppose the microdata file consists of records for a sample $s \subset U$ of size $n \leq N$. We assume that the possibility of statistical disclosure arises if an intruder gains access to the microdata and attempts to match a microdata record to external information on a known unit using the values of $m$ categorical key variables $X_1, \ldots, X_m$. (Note that $s$ and $X_1, \ldots, X_m$ are defined after the application of (sub)sampling or recoding, respectively, as SDC methods to the original microdata file.)

Let the variable formed by cross-classifying $X_1, \ldots, X_m$ be denoted by $X$, with values denoted $k = 1, \ldots, K$, where $K$ is the number of categories or key values of $X$. Each of these key values corresponds to a possible combination of categories of the key variables. Under the no measurement error assumption, identity disclosure is of particular concern if a record is unique in the population with respect to the key variables. A record with key value $k$ is said to be *population unique* if $F_k = 1$, where $F_k$ denotes the number of units in $U$ with key value $k$. If an intruder observes a match with a record with key value $k$, knows that the record is population unique and can make the no measurement error assumption then the intruder can infer that the match is correct.

As a simple measure of disclosure risk, we might therefore consider taking some summary of the extent of population uniqueness. In survey sampling, it is usual to define parameters of interest at the population level and this might lead us to define our measure as the population proportion $N_1/N$, where $N_r = \sum_k I(F_k = r)$ is the population frequencies of frequencies, $r = 1, 2, \ldots$. From a disclosure risk perspective, however, we are interested in the risk for a specific microdata file it is natural to allow the risk measure to be sample dependent. Thus, we might expect the risk to be higher if a sample is selected with a high proportion of unusual identifiable units than for a sample where this proportion is lower. Thus, a more natural file-level measure is the proportion of population uniques in the sample. Let the sample counterpart of $F_k$ be denoted by $f_k$, then this measure can be expressed as follows:

$$\Pr(PU) = \sum_k I(f_k = 1, F_k = 1)/n. \tag{2}$$

It could be argued, however, that the denominator of this proportion should be made even smaller, since the only records which might possibly be population unique are ones that are sample unique (since $f_k \leq F_k$), that is, have a key value $k$ such that $f_k = 1$. Thus, a more conservative measure would be to take

$$\Pr(PU|SU) = \sum_k I(f_k = 1, F_k = 1)/n_1, \tag{3}$$

where $n_1$ is the number of sample uniques and, more generally, $n_r = \sum_k I(f_k = r)$ is the sample frequencies of frequencies. For further consideration of the proportion of sample uniques that are population unique, see Fienberg and Makov (1998) and Samuels (1998).

It may be argued (e.g., Skinner and Elliot, 2002) that these measures may be overoptimistic, since they only capture the risk arising from population uniques and not from other records with $F_k \geq 2$. If an intruder observes a match on a key value with frequency $F_k$, then (subject to the no measurement error assumption) the probability that the match is correct is $1/F_k$ under the exchangeability assumption that the intruder is equally likely to have selected any of the $F_k$ units in the population. An alternative measure of risk is then obtained by extending this notion of probability of correct match across different key values. Again, on worst case grounds, it is natural to restrict attention to sample uniques. One measure arises from supposing that the intruder starts with the microdata, is equally likely to select any sample unique and then matches this sample unique to the population. The probability that the resulting match is correct is then the simple average of $1/F_k$ across sample uniques:

$$\theta_s = \left[\sum_k I(f_k = 1)/F_k]/n_1\right]. \tag{4}$$

Another measure is

$$\theta_U = \sum_k I(f_k = 1) \Big/ \sum_k F_k I(f_k = 1), \tag{5}$$

which is the probability of a correct match under a scenario where the intruder searches at random across the population and finds a match with a sample unique.

All the above four measures are functions of both the $f_k$ and the $F_k$. The agency conducting the survey will be able to determine the sample quantities $f_k$ from the microdata but the population quantities $F_k$ will generally be unknown. It is, therefore, of interest to be able to make inference about the measures from sample data.

Skinner and Elliot (2002) showed that, under Bernoulli sampling with inclusion probability $\pi$, a simple design-unbiased estimator of $\theta_U$ is $\hat{\theta}_U = n_1/[n_1 + 2(\pi^{-1} - 1)n_2]$. They also provided a design consistent estimator for the asymptotic variance of $\hat{\theta}_U - \theta_U$. Skinner and Carter (2003) showed that a design-consistent estimator of $\theta_U$ for an arbitrary complex design is $\hat{\theta}_U = n_1/[n_1 + 2(\overline{\pi}_2^{-1} - 1)n_2]$, where $\overline{\pi}_2^{-1}$ is the mean of the inverse inclusion probabilities $\pi_i^{-1}$ for units $i$ with key values for which $f_k = 2$. They also provided a design-consistent estimator of the asymptotic variance of $\hat{\theta}_U - \theta_U$ under Poisson sampling.

Such simple design-based inference does not seem to be possible for the other three measures in (2)–(4). Assuming a symmetric design, such as Bernoulli sampling, we might suppose that $n_1, n_2, \ldots$ represent sufficient statistics and seek design-based moment-based estimators of the measures by solving the equations:

$$E(n_r) = \sum_t N_t P_{rt}, \quad r = 1, 2, \ldots,$$

where the coefficients $P_{rt}$ are known for sampling schemes, such as simple random sampling or Bernoulli sampling (Goodman, 1949). The solution of these equations for $N_t$ with $E(n_r)$ replaced by $n_r$ gives unbiased estimators of $K$ and $N_1$ under apparently weak conditions (Goodman, 1949). Unfortunately, Goodman found that the estimator of $K$ can be "very unreasonable" and the same appears to be so for the corresponding estimator of $N_1$. Bunge and Fitzpatrick (1993) reviewed approaches to estimating $K$ and discussed these difficulties. Zayatz (1991) and Greenberg and Zayatz (1992) proposed an alternative "nonparametric" estimator of $N_1$ but this appears to be subject to serious upward bias for small sampling fractions (Chen and Keller-McNulty, 1998).

One way of addressing these estimation difficulties is by making stronger modeling assumptions, in particular by assuming that the $F_k$ are independently distributed as follows:

$$F_k|\lambda_k \sim \text{Po}(\lambda_k) \tag{6}$$

where the $\lambda_k$ are independently and identically distributed, that is, that the $F_k$ follow a compound Poisson distribution. A tractable choice for the distribution of $\lambda_k$ is the gamma distribution (Bethlehem et al., 1990) although it does not appear to fit well in some real data applications (e.g., Chen and Keller-McNulty, 1998; Skinner et al., 1994). A much better fit is provided by the log-normal (Skinner and Holmes, 1993). Samuels (1998) discussed estimation of $\Pr(PU|SU)$ based on a Poisson-Dirichlet model. A general conclusion seems to be that results can be somewhat sensitive to the choice of model, especially as the sampling fraction decreases, and that $\theta_U$ can be more robustly estimated than the other three measures.

### *3.3. Record-level measures of identification risk*

A concern with file-level measures is that the principles governing confidentiality protection often seek to avoid the identification of *any* individual, that is require the risk to be below a threshold for each record, and such aims may not adequately be addressed by aggregate measures of the form (2)–(5). To address this concern, it is more natural to consider record level measures, that is, measures which may take different values for each microdata record. Such measures may help identify those parts of the sample where risk is high and more protection is needed and may be aggregated to a file level measure in different ways if desired (Lambert, 1993). Although record level measures may provide greater flexibility and insight when assessing whether specified forms of microdata output are "disclosive," they are potentially more difficult to estimate than file level measures.

A number of approaches have been proposed for the estimation of record level measures. For continuous key variables, Fuller (1993) showed how to assess the record level probability of identification in the presence of added noise, under normality assumptions. See also Paass (1988) and Duncan and Lambert (1989). We now consider related methods for categorical variables, following Skinner and Holmes (1998) and Elamir and Skinner (2006).

Consider a microdata record with key value $X$. Suppose the record is sample unique, that is, with a key value $k$ for which $f_k = 1$, since such records may be expected to be most risky. Suppose the intruder observes an exact match between this record and a known unit in the population. We make the no measurement error assumption so that there will be $F_k$ units in the population which potentially match the record. We also assume no response knowledge (see Section 2.1). The probability that this observed match is correct is

$$\Pr(\text{correct match} \mid \text{exact match}, X = k, F_k) = 1/F_k, \tag{7}$$

where the probability distribution is with respect to the design under a symmetric sampling scheme, such as simple random sampling or Bernoulli sampling. (Alternatively, it could be with respect to a stochastic mechanism used by the intruder, which selects any of the $F_k$ units with equal probability). This probability is conditional on the key value $k$ and on $F_k$.

In practice, we only observe the sample frequencies $f_k$ and not the $F_k$. We, therefore, integrate out over the uncertainty about $F_k$ and write the measure as

$$\Pr(\text{correct match} \mid \text{exact match}, X = k, f_k) = E(1/F_k|k, f_k = 1). \tag{8}$$

This expectation is with respect to both the sampling scheme and a model generating the $F_k$, such as the compound Poisson model in (6). An alternative measure, focusing on the risk from population uniqueness, is

$$\Pr(F_k = 1|k, f_k = 1). \tag{9}$$

The expressions in (8) and (9) may be generalized for any record in the microdata with $f_k > 1$. A difference between the probabilities in (8) and (9) and those in the previous section is that here we condition on the record's key value $X = k$. Thus, although we might assume $F_k|\lambda_k \sim \text{Po}(\lambda_k)$, as in (6), we should like to condition on the particular key value $k$ when considering the distribution of $\lambda_k$. Otherwise, if the $\lambda_k$ is identically distributed as in the previous section, then we would obtain the same measure of risk for all (sample unique) records. A natural model is a log-linear model:

$$\log(\lambda_k) = z_k\beta, \tag{10}$$

where $z_k$ is a vector of indicator variables representing the main effects and the interactions between the key variables $X_1, \ldots, X_m$, and $\beta$ is a vector of unknown parameters.

Expressions for the risk measures in (8) and (9) in terms of $\beta$ are provided by Skinner and Holmes (1998) and Elamir and Skinner (2006). Assumptions about the sampling scheme are required to estimate $\beta$. Under Bernoulli sampling with inclusion probability $\pi$, it follows from (6) that $f_k|\lambda_k \sim \text{Po}(\pi\lambda_k)$. Assuming also (10), $\beta$ may be estimated by standard maximum likelihood methods. A simple extension of this argument also applies under Poisson sampling where the inclusion probability $\pi_k$ may vary with respect to the key variables, for example, if a stratifying variable is included among the key variables. In this case, we have $f_k|\lambda_k \sim \text{Po}(\pi_k\lambda_k)$. Skinner and Shlomo (2008) discussed methods for the specification of the model in (10). Skinner (2007) discussed the possible dependence of the measure on the search method used by the intruder.

## *3.4. SDC methods*

In this section, we summarize a number of SDC methods for survey microdata.

### *3.4.1. Transformation of variables to reduce detail*

Categorical key variables may be transformed, in particular, by combining categories. For example, the variable household size might be *top coded* by creating a single maximum category, such as 8+. Continuous key variables may be *banded* to form ordinal categorical variables by specifying a series of cut-points between which the intervals define categories. The protection provided by combining categories of key variables can be assessed following the methods in Sections 3.2 and 3.3. See also Reiter (2005). Provided the transformation is clear and explicit, this SDC method has the advantage that the reduction of utility is clear to the data user, who may suffer loss of information but the validity of analyses is not damaged.

### *3.4.2. Stochastic perturbation of variables*

The values of potential key variables are perturbed in a stochastic way. In the case of continuous variables, perturbation might involve the *addition of noise*, analogous to the addition of measurement error (Fuller, 1993; Sullivan and Fuller, 1989). In the case of categorical variables, perturbation may consist of misclassification, termed the *Postrandomization Method* (PRAM) by Gouweleeuw et al. (1998). Perturbation may be undertaken in a way to preserve specified features of the microdata, for example, the means and standard deviations of variables in the perturbed microdata may be the same as in the original microdata, but in practice there will inevitably be unspecified features of the microdata which are not reproduced. For example, the estimated correlation between a perturbed variable and an unperturbed variable will often be downwardly biased if an analyst uses the perturbed data but ignores the fact that perturbation has taken place. An alternative is to provide users with the precise details of the perturbation method, including parameter values, such as the standard deviation of the noise or the entries in the misclassification matrix, so that they may “undo” the impact of perturbation when undertaking their analyses. See, for example, Van den Hout and Van der Heijden (2002) in the case of PRAM or Fuller (1993) in the case of added noise. In principle, this may permit valid analyses although there will usually be a loss of precision and the practical disadvantages are significant.

### *3.4.3. Synthetic microdata*

This approach is similar to the previous approach, except that the aim is to avoid requiring special methods of analysis. Instead, the values of variables in the file are replaced by values generated from a model in a way that is designed for the analysis of the synthetic data, as if it were the true data, to generate consistent point estimates (under the assumption that the model is valid). The model is obtained from fitting to the original microdata. To enable valid standard errors as well as consistent point estimators, Raghunathan et al. (2003) proposed that multiple copies of the synthetic microdata are generated in such a way that multiple imputation methodology can be used. See Reiter (2002) for discussion of complex designs. Abowd and Lane (2004) discussed release strategies combining remote access to one or more such synthetic microdata files with much more restricted access to the original microdata in a safe setting.

### *3.4.4. Selective perturbation*

Often concern focuses only on records deemed to be risky and it may be expected that utility will be greater if only a subset of risk records is perturbed. In addition to creating stochastically perturbed or synthetic values for only targeted records, it is also possible just to create missing values in these records, called *local suppression* by Willenborg and de Waal (2001), or both to create missing values and to replace these by imputed values, called *blank and impute* by Federal Committee on Statistical Methodology (2005). A major problem with such methods is that they are likely to create biases if the targeted values are unusual. The data user will typically not be able to quantify these biases, especially when the records selected for blanking depend on the values of the variable(s) which are to made missing. Reiter (2003) discussed how valid inference may be conducted if multiple imputed values are generated in a specified way for the selected records. He referred to the resulting data as *partially synthetic microdata*.

### *3.4.5. Record swapping*

The previous methods focus on the perturbation of the values of the variables for all or a subset of records. The method of record swapping involves, instead, the values of one or more key variables being swapped between records. The choice of records between which values are swapped may be controlled so that certain bivariate or multivariate frequencies are maintained (Dalenius and Reiss, 1982) in particular by only swapping records sharing certain characteristics (Willenborg and de Waal, 2001, Section 5.6). In general, however, it will not be possible to control all multivariate relationships and record swapping may damage utility in an analogous way to misclassification (Skinner and Shlomo, 2007). Reiter (2005) discussed the impact of swapping on identification risk.

### *3.4.6. Microaggregation*

This method (Defays and Anwar, 1998) is relevant for continuous variables, such as in business survey microdata, and in its basic form consists of ordering the values of each variable and forming groups of a specified size $k$ (the first group contains the $k$ smallest values, the second group the next $k$ smallest values, and so on). The method replaces the values by their group means, separately for each variable. An advantage of the method is that the modification to the data will usually be greatest for outlying values, which might also be deemed the most risky. It is difficult, however, for the user to assess the biasing impact of the method on analyses.

SDC methods will generally be applied after the editing phase of the survey, during which data may be modified to meet certain edit constraints (see Chapter 9). The application of some SDC methods may, however, lead to failure of some of these constraints. Shlomo and de Waal (2006) discussed how SDC methods may be adapted to take account of editing considerations.

## *3.5. SDC for survey weights and other design information*

Survey weights and other complex design information are often released with survey microdata in order that valid analyses can be undertaken. It is possible, however, that such design information may contribute to disclosure risk. For example, suppose a survey is stratified by a categorical variable $X$ with different sampling fractions in different categories of $X$. Then, if the nature of the sampling design is published (as is common), it may be possible for the intruder to determine the categories of $X$ from the survey weight. Thus, the survey design variable may effectively become a key variable. See de Waal and Willenborg (1997) and Willenborg and de Waal (2001, Section 5.7) for further discussion of how survey weights may lead to design variables becoming key variables. Note that this does not imply that survey weights should not be released; it just means that disclosure risk assessments should take account of what information survey weights may convey. Willenborg and de Waal (2001, Section 5.7.3) and Mitra and Reiter (2006) proposed some approaches to adjusting weights to reduce risk.

In addition to the release of survey weights, it is common to release either stratum or PSU labels or replicate labels, to enable variances to be estimated. These labels will generally be arbitrary and will not, in themselves, convey any identifying information. Nevertheless, as for survey weights, the possibility that they could be used to convey information indirectly needs to be considered. For example, if the PSUs are defined by areas for which public information is available, for example, a property tax rate, and the microdata file includes area-level variables, then it is possible that these variables may enable a PSU to be linked to a known area. As another example, suppose that a PSU is an institution, such as a school, then school level variables on the microdata file, such as the school enrolment size, might enable the PSU to be linked to a known institution. Even for individual level microdata variables, it is possible that sample-based estimates of the total or mean of such variables for a stratum, say, could be matched to published values, allowing for sampling uncertainty.

A standard simple approach to avoiding releasing PSU or replicate identifiers is to provide information on design effects or generalized variance functions instead. Such methods are often inadequate, however, for the full range of uses of survey microdata (Yung, 1997). Some possible more sophisticated approaches include the use of adjusted bootstrap replicate weights (Yung, 1997), adjusted pseudoreplicates or pseudo PSU identifiers (Dohrmann et al., 2002), or combined stratum variance estimators (Lu et al., 2006).

## 4. Conclusion

The development of SDC methodology continues to be stimulated by a wide range of practical challenges and by ongoing innovations in the ways that survey data are used, with no signs of diminishing concerns about confidentiality. There has been a tendency for some SDC methods to be developed in somewhat ad hoc way to address specific problems, and one aim of this chapter has been to draw out some principles and general approaches which can guide a more unified methodological development. Statistical modeling has provided one important framework for this purpose. Other fields with the potential to influence the systematic development of SDC methodology in the future include data mining, in particular methods related to record linkage and approaches to privacy protection in computer science and database technology.

## Acknowledgments

I am grateful to Natalie Shlomo and a reviewer for comments on an earlier draft.

# Introduction to Part 3

Jack G. Gambino

National statistical offices (NSOs) conduct periodic population censuses and surveys of households, businesses, and agricultural operations. Four of the chapters in Part 3 are devoted to these four areas. Two of the other three chapters cover what are usually "private sector" endeavors, namely, opinion polls and marketing research. The remaining chapter looks at environmental surveys, which here are literally surveys of the environment, involving direct measurement, and not surveys of households and enterprises on environmental topics. The reader will find that, although there are substantial differences in the various types of surveys covered in these chapters, there is also a great deal of overlap in the underlying survey sampling methodology. In fact, the theory covered in Part 1 of this volume is the statistical foundation for the types of surveys discussed in Part 3. It is in this sense that Part 3 is on survey applications.

The complete survey process involves many more facets than we will cover in this introduction and in the chapters themselves. We do not discuss survey financing, the choice of variables of interest, concepts and definitions, and other "front-end" topics. Nor do we discuss back-end topics such as data capture and dissemination. Our range of topics goes from the choice of sampling frame to estimation, at least the parts where survey statisticians play a key role.

## 1. Frames and designs

A challenge common to the types of survey under consideration is the choice of sampling frame. Traditional list frames and area frames are used by most types of surveys, the only real difference being their prevalence: in many countries, business surveys and household surveys are much more likely to use, respectively, a list frame and an area frame. We see in Chapters 16, 20, and 22 that the use of telephone lists to create a frame is common in opinion polls, marketing surveys, and NSO-run household surveys. Chapter 22 by Francovic, Panagopoulos, and Shapiro includes some classic examples of the bias that can result when such frames have uneven coverage of the population. The use of more than one frame for a given survey, discussed in Part 1, is becoming more common, both because of deficiencies in list frames and the high costs often associated with the use of area frames.

The use of area frames is common for household surveys and for the types of environmental survey discussed in Chapter 19. For agricultural surveys, the use of area frames tends to decrease as the degree of consolidation of farm operations increases, that is, it is related to the relative importance of the traditional family farm. As Nusser and House note in Chapter 18, surveys of small farms may be similar in design to household surveys. Thus, in developing countries where family farms still dominate the agricultural sector, this is the case. Conversely, in many developed countries, there has been tremendous consolidation of farm operations, and as a result, a large proportion of agricultural production is now truly the result of business operations. These businesses appear in lists, such as business registers, and can therefore be surveyed like any other business. However, in terms of numbers, there are still very many family farms. Thus, it seems natural to use business survey methods for large agricultural enterprises and household survey methods for small operations. Nusser and House discuss the challenges due to this mix.

There is a relationship between frames and sampling units on the one hand and sample designs on the other. List frames are often associated with stratified simple random sampling (or something close to it); therefore, this is the type of design commonly used for business surveys. Area frames, which are often used for household surveys and some agricultural and environmental surveys, are often associated with multistage designs, typically with PPS sampling. We also see this type of design in Nirel and Glickman's discussion, in Chapter 21, of the design of surveys for estimating census coverage errors (over and undercounts), particularly those using a dual system estimator approach.

The choice of sampling units and sampling stages is strongly influenced by the nature of the frame and by collection costs. These include the cost of maintaining selected units (e.g., ensuring that the list of lower-level units within the selected higher-level ones is reasonably complete and kept up to date) and, for surveys that conduct interviews in person, the cost of interviewer travel. This is discussed in greater detail by Gambino and Silva in Chapter 16.

In Chapter 19, Marker and Stevens discuss both traditional area frames, where the boundaries of sampling units are defined using physical features such as roads and rivers, and frames that may be unfamiliar to some survey statisticians. These include the use of a grid to create a frame of units and frames based on data from a geographic information system (GIS). There are other features, such as spatial balance (defined formally in Chapter 19), that are important in some environmental surveys but not usually in other surveys. More generally, Marker and Stevens discuss some designs used for environmental surveys that are not usually seen elsewhere (e.g., random tessellation stratified (RTS) designs).

Because environmental and agricultural economists both have an interest in land use, watersheds, and the effect of fertilizers and pesticides on the environment, there are a number of issues common to environmental surveys and certain agricultural surveys, particularly those in which there is no respondent in the usual sense. These surveys involve some form of direct measurement, such as remote sensing (satellite or aerial imagery). More generally, the spatial context underlies both types of survey. In Chapter 18, Nusser and House discuss agricultural surveys of this type as well as agricultural surveys that collect information from farm operators using methods like those used in business and household surveys.

In Chapter 20 by Velu and Naidu, we see that the design of marketing surveys has had an evolution that parallels that of NSO-run household surveys: reacting to changes in society and technology, moving away from interviewing in the home to less expensive methods, and confronting increased concerns about privacy and confidentiality. In addition to designs familiar to most statisticians, such as designs that use random digit dialing (RDD) or frames based on telephone lists, Velu and Naidu discuss less familiar designs such as shopping center sampling (mall intercept interviews). They note that despite problems such as non-representativity of the sample for the population under study, such surveys are growing in popularity due to their advantages, particularly cost.

The traditional census is a survey in which all units are selected, of course, but sampling plays a role here as well. Therefore, frame and design issues arise in this context too. In Chapter 21, Nirel and Glickman begin with a brief review of traditional censuses (area based, single point in time, etc.) to contrast them with new approaches using administrative data or cumulated samples over time. The chapter then discusses at length three important census topics. The first is the measurement of coverage error using a survey, with a focus on the dual system estimator approach. This is followed by a discussion of the use of surveys cumulated over time in lieu of a traditional census, using the French approach to illustrate the method. Finally, an approach that combines features of the cumulation approach and the traditional census is discussed. In this approach, a traditional census with limited content is conducted periodically, but it is supplemented by a large, ongoing survey with rich content, which is cumulated over time to provide estimates for small areas. The authors use the American Community Survey to illustrate this approach.

## 2. Stratification, allocation and sampling

Most of the sampling methods discussed in Part 3 are well known and are covered in detail in Part 1. The chapters in Part 3 explain how these methods are used in practice and also describe some approaches to sample selection that are unique to a particular area such as marketing research or environmental surveys. Sampling methods used in opinion and election polling (Chapter 22) and in surveys designed to measure census coverage (Chapter 21) have a great deal in common with those used in household surveys. In all cases, the same factors lead to the use of multistage PPS sampling (e.g., cost considerations) or to simpler designs (e.g., the existence of an adequate list frame).

Stratification in list-based agricultural surveys is very similar to that in business surveys. For business surveys, strata are formed at the highest level using geography (usually large sub-national units such as provinces), type of industry, and unit size (e.g., size of establishment). For agricultural surveys, type of industry may be replaced by commodity as a stratification dimension. In Chapter 17, Hidiroglou and Lavallée describe methods for determining the boundary between take-all and take-some strata (a take-all stratum is one in which all units are sampled with probability 1). For business surveys, defining strata is not nearly as difficult as dealing with changes in the classification of units (e.g., change in industry or size) once the strata have been formed. This is also discussed in Chapter 17.

Like stratification methods, methods for sample allocation (i.e., deciding how much of the sample should be allotted to each stratum) are also well-developed. The total sample size and sample allocation are determined by cost, quality (related to variance), and time (how soon and how frequently we need results). The theory is described in Chapter 17. Current concerns are often related to the need to allocate the sample for multiple purposes and for various domains, particularly levels of geography. These concerns, and the compromises needed to address them, are mentioned in various places in Part 3, particularly in Chapter 16, 17 and 19. Frankovic et al. use the term allocation in a different sense. For example, they discuss the allocation of undecided respondents to the candidates in an election poll.

With the major exception of random digit dialing (RDD, described by both Velu and Naidu in Chapter 20 and in greater detail by Wolter et al. in Chapter 7 in Part 1), the sampling methods used in marketing research are often quite different from those described in the rest of Part 3. We have already mentioned shopping center interviewing. To this we add the use of consumer panels, which is described in detail by Velu and Naidu. They also include a timely discussion of internet surveys, by which they mean surveys that select potential respondents via the internet, as opposed to traditional surveys that may have the internet as a mode of response.

Environmental surveys, described by Marker and Stevens in Chapter 19, also present some unique sampling design challenges, because of their spatial context. Although it is true that for human populations, proximity can mean similarity (in incomes, say), the effect is much more pronounced for environmental variables. Marker and Stevens devote a large part of their chapter to describing the consequences of this fact on sample design.

Many surveys, be they business, household, agricultural, or environmental, are repeated with some predetermined frequency, because estimation of changes and trends is important for policy makers. Chapters 16 and 17, on household and business surveys, respectively, each devote a whole section to sample rotation. Much of what they present applies to any type of survey. The key idea, which is well known, is that maximizing sample overlap from period to period is optimal for estimating change. But the price to pay is response burden. Thus, a compromise is needed between estimating change efficiently and managing response burden. Gambino and Silva discuss the implications in Chapter 16.

For traditional censuses, of course, rotation does not come into play. However, it is at the heart of new "census-like" approaches such as the American Community Survey and the new French census, both of which implement variations on Kish's rolling samples idea. Both are described by Nirel and Glickman in Chapter 21.

## 3. Estimation

The primary output of any survey or census is a set of estimates. Comparing the estimation methods discussed in the seven chapters of Part 3, we notice that estimation methods have become much more sophisticated than simple expansion (Horvitz–Thompson) estimators. At the very least, the weights used to produce estimates are adjusted to make the sample "look like the population," using poststratification. If additional auxiliary information is available, a survey is likely to use some form of regression estimator or, more generally, calibration to improve the quality of its estimates. See Chapter 25 for details.

Each type of survey covered in Part 3 has its own special issues and concerns related to estimation. Household surveys often use complex, multistage design, and this is reflected in the estimation methods used. This is especially true of variance estimation, where replication methods (jackknife, bootstrap, etc.) are commonly used because of their relative ease of implementation for a wide variety of estimators. Business surveys, on the other hand, sometimes use two-phase designs, which present different challenges. Outliers are also a more severe problem for business surveys due to the skewed distributions of their populations and the sometimes dramatic change in the size of some population units. Among the interesting estimation challenges of agricultural surveys, we note the need to forecast crop harvests as quickly and as early as possible in the year. The surveys designed to estimate census coverage, described by Nirel and Glickman, have some distinct features, in part because their goal is to measure something that, one hopes, is small, namely errors in the coverage. The difficulty is exacerbated by the variability in the coverage error across regions, races, age groups, and so on. Nirel and Glickman also discuss rolling samples (censuses) that introduce other interesting estimation challenges such as how to combine different "vintages" of sample (e.g., whether older units should get the same weight as recent ones).

A requirement of many surveys, particularly business and agricultural surveys, is the reconciliation of estimates with data from external (administrative) sources. In addition, the estimates produced by various surveys (e.g., various industries, various farm types) must be brought together for the National Accounts, and so, the estimates must be coherent as much as possible. This is especially important because these estimates are often used as indicators of not only where the economy has been, but also where it is heading. The estimation of trends, especially the trade-off between trend and level estimation, is a theme covered throughout Nusser and House's chapter on agricultural surveys. Marker and Stevens also devote a part of their chapter on environmental surveys to the issue. Velu and Naidu note in Chapter 20 that to estimate trend, marketing researchers use panels of various kinds (consumer panels, store audits).

Every paper on small area estimation mentions the increasing demand for such estimates. The chapters in Part 3 all address the issue of small area or small domain estimation to some degree. We have already mentioned the rolling samples discussed by Nirel and Glickman. One of the primary motivations for having such samples is to accumulate units over time to make the estimation of variables for small domains feasible. Most surveys cannot be designed to produce good estimates for all small domains of interest. The statistician then turns to special estimation methods for such domains. Chapters 31 and 32 in Part 5 include detailed discussions of small area estimation methods.

## 4. Auxiliary information

Auxiliary data, typically from a census or administrative source, is used at various stages of survey design (stratification, sample allocation, unit formation, etc.). Most introductory texts on sampling explain how this is done in business and household surveys, but the same applies to the other types of survey covered in Part 3. For example, Marker and Stevens describe how an auxiliary variable is used to implement spatial balancing. Lavallée and Hidiroglou include a detailed discussion of the uses of auxiliary data, particularly administrative data, at various stages of the survey process in business surveys. Much of what they say applies to other surveys as well.

Auxiliary data is also essential in modern estimation methods, and we have already mentioned their use in regression and calibration estimators. Both Lavallée and

Hidiroglou (Chapter 17) and Gambino and Silva (Chapter 16) include sections with detailed discussion of these estimation methods and their use in their respective surveys.

A recent trend has been the increased use of auxiliary data, usually tax data, as a complement to, or replacement for, survey data. Gambino and Silva describe how income tax data are used to replace questions on income for some household surveys for respondents who consent to this. However, this is a very limited use of auxiliary data compared to how it is being used in some business surveys conducted by NSOs. Broadly speaking, they divide the population into three groups: the take-all units (the very biggest businesses, which are sampled with certainty), the take-some units (medium-sized businesses, which are sampled with some probability between 0 and 1), and the so-called take-none units (the smallest units, which are not sampled at all). The information for the take-none units is obtained from tax files. This is discussed briefly by Hidiroglou and Lavallée in the section on the uses of administrative data. They also explain methods for determining a boundary between take-all and take-some strata in greater detail.

## 5. Challenges

The challenges that face the various types of surveys discussed in Part 3 are surprisingly similar. The increasing appetite for information is certainly one of them. To satisfy it, the burden on respondents inevitably increases, except possibly for surveys that focus on the relatively limited set of variables that can be obtained from administrative sources. The need to manage the growth in response burden, particularly in business surveys, has led to the development of formal methods of sample coordination (discussed in Chapter 17). The increase in respondent burden coincides with what appears to be an across-the-board decrease in response rates. This phenomenon can be explained in part by the individual's perceived increase in burden, not necessarily from the specific survey or poll that is making the contact, but from the many solicitations we are all subjected to. Another part of the explanation is that there have been technological developments that make it easier to thwart contact efforts (discussed in several chapters in Part 3).

The rapidly increasing use of mobile (or cell) phones, especially for individuals who do not have a landline telephone, is another factor that makes it difficult to maintain high response rates because most telephone surveys have relied on landline telephones to select their sample or to conduct interviews. This is discussed in several chapters as well. In Chapter 22, Frankovic et al. discuss the challenges posed not only by cell phones but also by other technologies.

New technologies present positive challenges as well. The ease with which GPS (global positioning system) data can be obtained or recorded presents opportunities not only for agricultural and environmental surveys but also for censuses and household surveys. This is discussed in several chapters, but especially by Nusser and House in Chapter 18. Another positive development is the use of the internet as a mode for responding to surveys. Again, this cuts across the different types of surveys and is mentioned in several places in Part 3. A more difficult challenge is to use the internet not simply as a collection mode but as a frame from which to sample. A lot of thought is being given to how to do this properly, involving difficult challenges such as selection bias and coverage issues. Developments in this area are discussed by both Velu and Naidu in Chapter 20 and Frankovic et al. in Chapter 22.

One consequence of the various developments we have discussed is that for many surveys, multimode collection, where responses for a given survey can be obtained in a variety of ways (telephone, paper, internet, CAPI, etc.), may become the norm, in the hope that it will help to prevent response rates from decreasing further. However, it is well known that the mode of collection has an effect on response. This leads to the challenge of measuring and adjusting for this mode effect. This is not a new problem, and most chapters in Part 3 mention it, but the greater use of multiple modes and the increasing variety of modes available make this a topic worthy of greater attention. The same can be said more generally about the need to better measure and manage nonsampling errors because the changes we have described in the last few paragraphs have complicated an already difficult problem.

*Sample Surveys: Design, Methods and Applications, Vol. 29A*
ISSN: 0169-7161
© 2009 Published by Elsevier B.V.
DOI: 10.1016/S0169-7161(08)00016-3

# Sampling and Estimation in Household Surveys

*Jack G. Gambino and Pedro Luis do Nascimento Silva*

## 1. Introduction

A household survey is a particular type of social survey. In a household survey, we are interested in the characteristics of all or some members of the household. These characteristics typically include a subset of variables such as health, education, income, expenditure, employment status, use of various types of services, etc. Since they became common in the 1940s, a number of major trends in household surveys have been evident. Many of these trends are closely linked to technological advances both in statistical agencies and in society, and have accelerated following the spread of personal computers in the early 1980s. These trends include, but are not limited to, the following.

### *1.1. Simplification of sample designs*

A good example of simplification is the Canadian Labour Force Survey (LFS), which went from as many as four stages of sampling to two stages, and for which the feasibility of using a single-stage design, with an address register as a frame, is currently being studied. In the United Kingdom, the LFS already uses, and the new Integrated Household Survey will use, an unclustered design (see Office for National Statistics, 2004). In the United States, the American Community Survey (ACS) also adopted a stratified unclustered design (see U.S. Census Bureau, 2006b).

### *1.2. Increasingly complex estimation methods*

The increasing power and availability of computers has made it possible to use increasingly complex estimation procedures.

### *1.3. Increased use of telephone interviewing*

As the proportion of households with a telephone has increased, the proportion of interviews conducted in person (across surveys) has decreased. This trend accelerated with the introduction of computer-assisted interviewing. A related trend is the increased use of multiple modes of collection for the same survey. The latter trend is likely to continue as use of the internet as a medium for survey response becomes more popular.

### *1.4. Increasingly complex questionnaires*

The introduction of computer-assisted interviewing has made it possible to have questionnaires with complex skip patterns, built-in edits, and questions tailored to the respondent.

### *1.5. Increased availability of data on data collection (paradata)*

The use of computers for interviewing makes it possible to save data on various aspects of the interviewing process. In addition, interviews can be monitored, providing another source of data.

We discuss some of these trends throughout the chapter. There are a number of other important trends we do not discuss since they are covered in other chapters in this volume. They include increasingly elaborate editing and imputation procedures, the rising importance of confidentiality and privacy, questionnaire design and related research, and the advent of more sophisticated data analysis methods, particularly for data from complex surveys.

## 2. Survey designs

Much of the theory and practice of survey design was developed from the 1930s to the 1960s. In fact, many methods currently in use, particularly for area-based sampling, are already included in the classic text by Hansen et al. (1953). We will often refer to dwellings, which are sometimes referred to as dwelling units or housing units in other publications. A dwelling may be vacant (unoccupied) or occupied by a household.

### *2.1. Frames*

The traditional frames used for household surveys are area frames and list frames. Alternatives to these include the use of random-digit dialling (RDD) and the internet.

#### *2.1.1. List frames*

If a list of population units is available, then it can be used to sample directly, possibly after stratification of the units into more homogeneous groups. Examples include lists of dwellings, lists of households or families, lists of people, and lists of telephone numbers. A major advantage of list frames for surveys is that they are easy to use for sampling and usually lead to relatively straightforward weighting and estimation procedures. On the other hand, it is often difficult and expensive to keep a list up-to-date in light of individual changes such as moves, marriages and divorces, and births and deaths. Certain types of unit pose difficulties. For example, students and workers living temporarily at a work site may be listed twice or missed completely. There are many other situations that can lead to missed units (or undercoverage) or double-counted units (or overcoverage), as well as to units that do not belong to the list (also overcoverage). Nevertheless, the benefits of having an up-to-date list of units are sufficiently great that several countries have invested in the creation and maintenance of permanent lists, including population registers and address registers. For example, Scandinavian countries have made increasing use of population registers in their censuses, including complete replacement of traditional censuses in some cases (see Statistics Finland, 2004; Statistics Norway, 2005).

### *2.1.2. Area frames*

An area frame is obtained by dividing a country (or province, state, etc.) into many mutually exclusive and exhaustive smaller areas. In principle, therefore, an area frame has complete coverage. In practice, the area frame approach pushes the coverage problems of list frames down to a smaller geographical level since, at some point in the sampling process, a list of ultimate sampling units (dwellings, households, or people) will be needed. Obtaining such lists for many small areas can be very time-consuming and expensive.

The use of area frames has led naturally to multistage designs where clusters form the penultimate stage: the country may be divided into provinces or states, which are divided into counties, say, and so on, until we come to the smallest area units, which we will refer to as (geographical) clusters. We may then start the sample selection process at one of these geographical levels. The simplest case would be direct selection of clusters (possibly after stratification), followed by selection of units (typically dwellings) within the clusters that were selected at the first stage. This has the great advantage that we need a complete list of ultimate (or elementary) units only for clusters that are selected in the sample—in effect, we "localize" the list frame creation problem.

In practice, getting a complete list of units (dwellings and/or people) in a cluster can be difficult. The cluster may contain easy to miss dwellings (e.g., they may not be obvious from the street). There may be problems identifying the cluster boundary, especially if a part of the boundary is an imaginary line or if new construction has occurred in the area. As a result, in the field, it may not be clear who belongs to the cluster.

The use of an area frame with multistage sampling is very common in both developing and developed countries. One important benefit is the reduced travel costs for personal interviewing when dwellings are selected in compact geographical areas such as clusters and higher level sampling units (e.g., a village can be a sampling unit) since the interviewer drives to a cluster and contacts several dwellings in close proximity.

### *2.1.3. Apartment frames*

A list of apartment buildings (typically useful in metropolitan areas) is, in a sense, at the intersection of list frames and area frames: a survey may use an area frame, but whenever apartment buildings are found (because they are new or were missed earlier), they are "removed" from the area frame and put in a separately maintained list frame of apartment buildings. Each apartment building may then be treated as a cluster.

### *2.1.4. Telephone-based frames*

A list of telephone numbers is simply a list frame, as discussed above. An alternative way of using telephone numbers is via RDD. In RDD, telephone numbers are generated at random, avoiding the need for a list of numbers. In practice, the process is more sophisticated than simply generating a string of digits and expecting that the result will be a valid telephone number. Efforts are made to eliminate invalid or business numbers in advance. One can also vary the probability for certain sets of numbers. There is a vast literature on RDD, which we do not cover here. Nathan (2001) includes an extensive list of references on RDD and other telephone-based methods of data collection.

Until recently, the use of mobile (or cell) phones as either a frame or as a mode for conducting interviews has been avoided, but this may be changing in light of the increasing number of households with mobile phones (and without a traditional landline telephone). Problems related to the use of mobile phones for surveys have generated a great deal of interest recently; see, for example, the 2005 Cell Phone Sampling Summit and several sessions at the 2007 AAPOR conference. Blumberg and Luke (2007) present recent results on the rapid increase in the number of cell phone-only households in the United States using data from the most recent National Health Interview Survey. They also look at the demographic and health-related characteristics of these households. For example, they find that homeowners are much less likely to be in a cell phone-only household than renters.

In this section, we have noted that survey statisticians face a variety of problems in constructing and maintaining frames. Yansaneh (2005) discusses these further and presents possible solutions to some of the problems.

## *2.2. Units and stages*

We have already had occasion to mention sampling units and sampling stages in this chapter. We now discuss these more formally in the following.

### *2.2.1. Persons and households*

In the surveys under consideration, interest is usually in persons, families, or households. We usually get to these units via the dwelling (for our purposes, we define a dwelling as a set of living quarters; a formal definition can be quite involved—e.g., the formal definition of private dwelling on the Statistics Canada web site is more than 500 words long). As discussed above, we get to the dwelling via an area frame, a dwelling frame, or a telephone number. Although formal definitions of units may be quite involved, smooth implementation in the field may require simplifications to be practicable.

In some surveys, the ultimate unit of interest is not the household but one person within the household. This introduces a problem of representativity of the sample if persons within households are selected by a naive method: certain groups (e.g., age groups) may be over- or under-represented. For example, Beland et al. (2005) cite a Canadian example where a naive approach (an individual in the household is selected with equal probability) yields 8.2% of the sample in the 12–19 age range, whereas the percentage in the population is 12.4. For people aged 65 and older, the corresponding percentages were 21.5 and 14.5. Thus, young people would be under-represented and old people over-represented. Solutions to this problem are discussed in the study by Beland et al. (2005) and by Tambay and Mohl (1995). Clark and Steel (2007) discuss optimal choice of the number of persons to select from each household.

### *2.2.2. Clusters*

We already mentioned clusters when we discussed frames. A cluster is usually a compact geographical area containing a few dozen to a few hundred households. In urban areas, clusters are typically formed by combining contiguous block faces or blocks. In rural areas, many countries use census enumeration areas as clusters, but there may also be natural clusters such as villages that can be used. Sometimes these clusters vary widely in size, which is undesirable, and may require using sampling with probabilities proportional to size. Apartment buildings are sometimes used as clusters, especially in large metropolitan areas where a significant proportion of the population lives in apartment complexes.

#### 2.2.3. The role of census units

In most countries, the process of conducting a population census entails the formation of a hierarchy of geographical units. It is natural to consider these "ready-made" units when designing household surveys. We have mentioned the use of a census unit, namely an enumeration area, as a sampling unit in surveys. More generally, census units are used in surveys for a variety of purposes: the biggest census units may form geographical strata (examples are provinces and states), and the smallest census units may be useful building blocks for the formation of primary sampling units (PSUs) and (optimal) strata. In countries with big populations, PSUs can be very large (e.g., a whole city can be a PSU). In the United States, national household surveys often use counties or groups of counties as PSUs. Typically, the largest PSUs are self-representing, i.e., they are selected with probability 1, which means they are really strata rather than PSUs.

### 2.3. Stratification

Subnational geographical areas such as provinces, states, and regions form the highest level of stratification both because they have well-defined, stable boundaries and because they are often of interest for policy-making. In most cases, these subnational areas are too big and need to be divided into finer geographical strata. Once the lowest level of geographical stratification is reached, there may be enough PSUs (e.g., census enumeration areas) in some geographical strata to form optimal strata within the latter. Optimality is defined by some measure of homogeneity and the PSUs are grouped into final strata that are as homogeneous as possible.

In some cases, it is more convenient to use *implicit* stratification via ordering of the units—similar units are placed near each other. Then, some form of systematic sampling is used to select the sample to ensure that no major subpopulations are left out of the sample. A special case is geographical ordering to ensure that no major areas are left out of the sample and also to achieve approximately proportional allocation of the sample between areas.

Textbooks on survey sampling tend to devote little space to stratum formation and even then, they emphasize the case of a single variable. In practice, several variables are usually of interest, and a compromise stratification is needed. A common tool for stratum formation based on several variables is cluster analysis. For a recent approach to stratification using a spatial cluster analysis algorithm that minimizes distances between PSUs in a stratum on selected variables considering the spatial location of PSUs, see Palmieri Lage et al. (2001).

### 2.4. Sample size

The determination of sample size for household surveys is complicated by the fact that most surveys are interested in several variables, so the standard textbook formulas based on a single variable are not adequate. In addition, one must decide on a criterion: standard error versus coefficient of variation (CV), that is, to aim to control either absolute or relative error. Finally, if a clustered design will be used, the "IID" (independent and identically distributed) or "SRS" sample size formulas are inadequate—they will likely understate the required sample size. The design effect (deff) is a measure of this phenomenon. The deff is defined as the ratio of the variance of an estimator under the actual design to the variance of the estimator under simple random sampling (SRS), assuming that the sample size is the same for both designs. Thus,

$$\operatorname{deff}(\hat{\theta};\, p) = V_{p,n}(\hat{\theta}) / V_{\mathrm{SRS},n}(\hat{\theta}), \tag{1}$$

where $\hat{\theta}$ denotes an estimator of a parameter $\theta$, $p$ denotes a complex survey design, $V_{p,n}(\hat{\theta})$ and $V_{\mathrm{SRS},n}(\hat{\theta})$ denote the variances of $\hat{\theta}$ under the designs $p$ and SRS, respectively, with $n$ defined as the number of sampled households.

A common approach to sample size determination in complex surveys is to use information from similar surveys or census data to obtain (or assume) design effects and population variances for key variables, and use these *deffs* and variances to determine $n$. This follows from using (1) in the following manner. Suppose a sample size $n_0$ can be determined using the standard SRS formulas so that a specified variance $v$ is achieved for the estimator of a key parameter, that is, $n_0$ solves $V_{\mathrm{SRS},n_0}(\hat{\theta}) = v$. Then, if the same sample size $n_0$ was used with the complex design $p$, (1) implies that $V_{p,n_0}(\hat{\theta})/\operatorname{deff}(\hat{\theta};\, p) = v \Leftrightarrow V_{p,n_0}(\hat{\theta}) = v \times \operatorname{deff}(\hat{\theta};\, p)$. Hence, to obtain the same variance $v$ using a complex design $p$, we need to solve $V_{\mathrm{SRS},n}(\hat{\theta}) = v/\operatorname{deff}(\hat{\theta};\, p)$, which leads to the simple solution corresponding to multiplying the initial sample size $n_0$ by deff, that is,

$$n = n_0 \times \operatorname{deff}(\hat{\theta};\, p) \tag{2}$$

For surveys where proportions are the target parameters, the above solution is simple, since sample sizes under SRS can be determined easily using the fact that the variances of sample proportions are maximized when the population proportion is ½ (see, e.g., Cochran, 1977, Chapter 3). This is a conservative solution which is feasible even if little or no information is available about the possible range of the population proportions. In cases where the target proportions are far from ½, especially near 0 (rare subpopulations), it may be useful to consider using sample sizes that aim to provide specified levels of relative variance or CV of the estimated proportions. In either case, the theory for sample size determination under SRS is quite simple, and the adjustment (2) may be applied to determine sample sizes for a complex design.

In practice, one must make sure that the "right" design effects are used. For example, if regression estimators will be used for the actual survey, one should use the corresponding *deffs* and not *deffs* of simple means or totals; otherwise, the formulas may give incorrect sample size requirements. The numerator and denominator in the *deff* formula should agree both in terms of key design features (e.g., stratification) and choice of estimator. It is misleading to have, say, a ratio estimator in the numerator and a total in the denominator, or stratification in the numerator but not in the denominator. Finally, if $V_{\mathrm{SRS},n}(\hat{\theta})$ is to be estimated using data from a complex survey, the estimate of $V$ is not the usual SRS one since the actual complex design (e.g., clustering) needs to be taken into account. Two useful references on design effects are Lê and Verma (1997) and Park and Lee (2004).

An area where household survey practice needs to improve is in making estimates of design effects widely available. Such estimates are not regularly provided, especially in less developed survey organizations. Survey designers are then left with the task of having to estimate *deff*s from the survey microdata (when these are available and carry sufficient information about the design that variances can be correctly estimated), or alternatively, from published estimates of standard errors for certain parameter estimates (again, if available). In either case, this can be time-consuming for those unfamiliar with the survey or having limited access to detailed information about its design.

All surveys have nonresponse, which must be taken into account at the design stage. In addition to taking an anticipated rate of nonresponse into account, designers of household surveys that use the dwelling as a sampling unit also need to take the proportions of vacant and ineligible dwellings into account. These can be quite stable for large areas but may vary for smaller ones, even over short time periods. Attention must be paid to dwellings of temporary residence, such as those commonly found in beach and mountain resorts, where the resident population is sometimes smaller than the temporary population. Similar care is needed to account for addresses that are not residential if using an address frame where it is not possible to determine beforehand which ones are occupied by households. Two options for addressing these issues include: a) increasing sample size by dividing the initial sample size by the expected proportion of eligible and responding dwellings, in which case the selected sample is fixed but the effective sample is random; b) using a form of inverse sampling, where the required number of responding eligible dwellings is fixed, but the total number of selected dwellings is random. In both cases, weighting is required to compensate for the unequal observed eligibility and response rates. Such weighting requires precise tracking of eligibility and response indicators during fieldwork (see Chapters 8 and 9).

Allocation of the sample to strata, both geographical and optimal, requires a compromise among the important variables that the survey will measure. In addition, the designer must make compromises between different geographical levels (national, subnational, and so on). For many variables, simply allocating the sample proportional to population size is nearly optimal for national estimates. However, this allocation will likely be poor for subnational estimates. For example, if the country is divided into $R$ regions and the national sample size is $n$, then a good allocation for regional estimates is likely to be to give each region about $n/R$ units. Unless the regions have approximately equal populations, the two allocations (proportional and equal) are likely to be very different, and a compromise must be found. A common approach is to allocate the sample proportional to the square root (or some other power) of population size (see, e.g., Kish, 1988). Singh et al. (1994) describe a pragmatic solution to the specific problem of producing good estimates for both the nation and relatively small areas within it, which has been used for the Canadian LFS since the late 1980s. In this approach, most of the sample, say two-thirds, is allocated to produce the best possible national estimates. The remaining sample is then allocated disproportionately to some smaller areas to ensure a minimal level of quality in each area. As a result, large metropolitan areas get little or no sample in the second allocation round and, conversely, sparsely populated small areas get much of their sample in that round.

## 2.5. Sample selection

One aspect where household and business surveys tend to differ most is sample selection. The methods used by business surveys are discussed in Chapter 17. Household surveys that sample from a list can use methods similar to those. For example, the ACS and the U.K. Integrated Household Sample Survey have unclustered samples of addresses selected from address lists. One advantage of this type of frame is the ability to coordinate samples over time, either to ensure that adjacent survey waves have overlap or to avoid such overlap, when it is important to get a fresh sample of addresses in each wave.

However, for multistage household surveys, methods where the probability of selection of a PSU is proportional to its size are more common. These are referred to as probability proportional to size or PPS methods. They are discussed in most standard textbooks on sampling (see, e.g., Cochran, 1977, Chapter 9A). Under multistage PPS sampling, even though PSUs are selected with unequal probabilities, we can have a *self-weighting* design, in which all ultimate sampling units in a stratum have the same final design weight (see Section 5.1). To illustrate this, consider a two-stage design and suppose PSUs are selected using PPS sampling, with size defined as the number of second-stage units in a PSU. Thus, if $M_i$ is the size of the $i$th PSU and $M$ is the total size of all the PSUs in the stratum, then the probability that PSU $i$ is selected is $p_i = nM_i/M$, where $n$ is the number of PSUs selected in the stratum. Now, select the same number $m$ of second-stage units from each sampled PSU using SRS. Then the second-stage inclusion probability is $p_{2ij} = m/M_i$ for all $j$ in PSU $i$. Hence, the overall inclusion probability for each second-stage unit is $p_i p_{2ij} = nM_i/M \times m/M_i = nm/M = 1/d$. The design is then *self-weighting* because all units in the sample have the same design weight $d$.

In practice, this textbook procedure is often not useful since unit sizes $M_i$ become out of date quickly. In fact, since the sizes are typically based on a recent census or an administrative source, they are likely to be out of date as soon as they become available. To preserve self-weighting in this more realistic situation, an alternative is to use systematic sampling at the second stage instead of SRS and fix the sampling interval over time. For example, if we should select every $K$th unit according to the census counts, then we continue to select every $K$th unit thereafter. One undesirable consequence of this procedure is that the sample size in each PSU is no longer constant. If a PSU has grown by 10%, then its sample will also grow by 10% and conversely for decreases in size. Since populations tend to grow over time, the former is the more serious problem. At the national level, this implies that the total sample size, and therefore costs, will gradually increase. One way to deal with this growth is to randomly drop enough units from the sample to keep the total sample size stable. This is the approach used by the Canadian LFS and the U.S. Current Population Survey (CPS).

An alternative is to design a sample which is self-weighting, but to allow the weights to vary over time for households selected from different PSUs. This weight variation would happen anyway if nonresponse varies between PSUs and if simple weight adjustments are applied at the PSU level. This design will be slightly less efficient than the corresponding self-weighting design but will not suffer from the cost-increase problem described above. Its main disadvantage is that varying household weights lead to more complex estimation procedures but this disadvantage is less important with the increased availability of modern computer facilities and software.

Regardless of the approach used, in practice, having a pure self-weighting design may not be attainable. For example, even if we implement the fixed sampling interval method described above, there will almost certainly be PSUs in the sample that have grown to such an extent that it will be necessary to subsample from them to control costs and balance interviewer workloads.

## 3. Repeated household surveys

### *3.1. Repeated versus longitudinal surveys*

Many household surveys are repeated over time with the same or very similar content and methodology to produce repeated measurements of key indicators that are used to assess how demographic, economic, and social conditions evolve. In fact, many descriptive analyses reported in household survey publications discuss how major indicators changed in comparison to previous survey rounds. The idea that a household survey is to be repeated introduces a number of interesting aspects of survey design and estimation, which we consider in this section. We start by establishing a distinction between repeated and longitudinal surveys.

*Longitudinal surveys* require that a sample of elementary survey units (say households or individuals) is followed over time, with the same units observed in at least two survey data collection rounds or waves. Observation of the selected units continues for a specified length of time, a number of waves, or until a well-specified event takes place (e.g., the person reaches a certain age). Longitudinal designs are essential if the survey must provide estimates of parameters that involve measures of change at the individual level.

*Repeated surveys* collect data from a specified target population at certain (regular) intervals using the same (or at least comparable) methodology. They do not require that the same elementary units should be followed over time but are often designed such that there is some overlap of units in successive survey waves. They also include surveys for which the samples on different occasions are deliberately non-overlapping or even completely independent.

When a repeated survey uses samples that are at least partially overlapping at the elementary unit level, it includes a longitudinal component, which may or may not be exploited for analysis. In such cases, the distinction between longitudinal and repeated surveys becomes blurred and the key to separating them is the main set of outcomes required from the survey. If the main parameters to be estimated require pairing measurements on the same elementary units from at least two survey waves, we classify the survey as longitudinal. Otherwise, we call it a repeated survey.

Longitudinal surveys are discussed in Chapters 5 and 34. In this section, we focus on some design and estimation issues regarding *repeated household surveys*. A more detailed classification of surveys in terms of how their samples evolve in time is provided by Duncan and Kalton (1987) and Kalton and Citro (1993).

The traditional design of household surveys requires specifying a sample selection procedure coupled with an estimation procedure that provides adequate precision for key parameters. Sample sizes are determined by taking account of the survey budget, cost functions describing the relative costs of including additional primary and elementary sampling units, and design effects if available. If the survey is to be repeated, the process by which the sample evolves in time is an additional element that must be designed for. We refer to this process as the *rotation scheme* or *rotation design* of the survey.

### *3.2. Objectives, rotation design, and frequency for repeated household surveys*

The key to efficient design for repeated household surveys is to match the sample selection mechanism, survey frequency, rotation scheme, and estimation procedures to satisfy the survey objectives at minimum cost for a fixed precision or maximum precision for a fixed cost. Key references on this topic are Binder and Hidiroglou (1988) and Duncan and Kalton (1987). For a good example of an in-depth discussion of how the survey objectives affect the rotation design in the case of a household labor force survey, see Steel (1997).

We first consider the problem of specifying what the key objectives of inference are for a repeated household survey. They may include the following:

(a) Estimating level: estimating specified population parameters at each time point;
(b) Estimating change: estimating (net) change in parameters between survey waves;
(c) Estimating averages: estimating the average value over several survey waves;
(d) Cumulating samples of rare populations or for small domains over time.

Kalton and Citro (1993) list several other objectives that require a survey to be repeated over time, but these require the longitudinal component of the survey to be of primary interest. Here, we focus only on repeated surveys, where the main objectives do not require the longitudinal component.

Considering objective (d), the best possible rotation design is to have completely non-overlapping samples in the various survey waves. This design maximizes the speed with which new observations from the rare target population can be found. For example, consider a survey which needs to cumulate observations of people who migrated from a foreign country to their current place of residence during the five years preceding interview time. To be able to have a sufficiently large sample of this subpopulation, it may be necessary to use either a cross-sectional screening survey with a very large sample size, or alternatively, a repeated survey with smaller samples at each wave, screening for this subpopulation, thus providing a sample of the intended size after a number of waves has been completed.

Note, however, that this would often be a secondary objective for a repeated survey because the alternative of using a large cross-sectional survey would probably be more cost-effective than the non-overlapping repeated survey option described above. Nevertheless, given an existing repeated survey (overlapping or not), it may be cost-effective to include the screening questions and additional survey modules as required for the measurement of this "rare" target subpopulation. Advantages and limitations of these two competing approaches must be carefully considered before choosing the survey design for any particular application.

It follows that the main objectives leading to a repeated survey design are likely to be (a), (b), (c), or a combination of these. For these objectives, alternative rotation designs affect the precision of estimators for each type of target parameter. Let $U_t$ denote the target population at time $t$ and let $\theta_t$ denote the value of a target parameter at time $t$, where $t$ could refer to years, quarters, months, weeks, etc. We assume that the definition of the target population is fixed over time, for example, all adults living in private dwellings. However, the size and composition of this target population may change over time, because people die, migrate, or reach the age limit to be included in the survey from a given time point onwards. Changes in the parameter $\theta_t$ over time can thus be caused by changes in both the composition of the population and in the values of the underlying characteristics of members of the population.

A repeated survey may have as key objective the estimation of the series of values of $\theta_1, \theta_2, \ldots, \theta_t, \ldots$, that is, the main goal is to get the best possible estimates for the *level* $\theta_t$ at each point in time (objective (a)). Alternatively, the target parameter may be the *change* between times $t$ and $t-1$, defined as $\theta_t - \theta_{t-1}$ (objective (b)). In some situations, the target parameter may be an average of the values of the parameters at different time points (objective (c)), and a simple example of this is $\bar{\theta}_{t,2} = (\theta_{t-1} + \theta_t)/2$, the (moving) average of the parameter values at two successive time points.

Denoting by $\hat{\theta}_t$, an (approximately) unbiased estimator of $\theta_t$, we have the following results on the precision of estimators of these types of target parameters. For the estimation of *change*, the variance of the simple (approximately) unbiased estimator $\hat{\theta}_t - \hat{\theta}_{t-1}$ is given by

$$V(\hat{\theta}_t - \hat{\theta}_{t-1}) = V(\hat{\theta}_t) + V(\hat{\theta}_{t-1}) - 2\mathrm{COV}(\hat{\theta}_{t-1}; \hat{\theta}_t). \tag{3}$$

If successive measurements on the same unit for the survey variable defining the parameters $\theta_t$ are positively correlated over time, then with some degree of overlap between the samples at times $t-1$ and $t$, the covariance term in the right-hand side of (3) would be positive. In this case, the estimation of the change would be more efficient with overlapping samples than with completely independent samples at times $t$ and $t-1$. Independent samples lead to complete independence between $\hat{\theta}_t$ and $\hat{\theta}_{t-1}$, in which case the variance of the difference is

$$V(\hat{\theta}_t - \hat{\theta}_{t-1}) = V(\hat{\theta}_t) + V(\hat{\theta}_{t-1}). \tag{4}$$

So for the estimation of change, some overlap of samples in successive waves increases the precision of the estimator when the underlying characteristic is positively correlated for measurements on successive occasions.

For the estimation of averages over time, the variance of the simple (approximately) unbiased estimator $\hat{\bar{\theta}}_{t,2} = (\hat{\theta}_t + \hat{\theta}_{t-1})/2$ is given by

$$V\left[\frac{1}{2}(\hat{\theta}_t + \hat{\theta}_{t-1})\right] = \frac{1}{4}\left[V(\hat{\theta}_t) + V(\hat{\theta}_{t-1}) + 2\mathrm{COV}(\hat{\theta}_{t-1}; \hat{\theta}_t)\right]. \tag{5}$$

Here, the positive correlation of successive measurements of the underlying survey characteristic would lead to reduced precision with overlapping surveys, and having independent or non-overlapping samples would be more efficient. This discussion illustrates the importance of regularly publishing estimates of the correlations over time between key measurements such that these are available to inform survey design or redesign.

These two examples illustrate the need to specify clearly what the inferential objectives are for the survey; otherwise, one may end up with an inefficient rotation design. In addition, they also indicate that if in a given situation, both changes and averages over time are required, overlapping samples increase efficiency for change but reduce it for averaging. This poses a problem to the survey designer and calls for an explicit assessment of the relative importance of the different survey objectives so that decisions regarding the rotation design are not misguided. The estimation of change is usually the harder of these two objectives, often requiring larger sample sizes than would be needed for the estimation of level itself. Hence, the survey could be designed to achieve the required level of precision for estimating change both in terms of sample design, sample size, and rotation design, and still be able to provide estimates of acceptable accuracy for averages over time.

Another important design parameter of a repeated survey is the frequency of the survey, which again is closely linked with the survey objectives. Surveys having a short interval between waves (say monthly or quarterly) are better for tracing respondents over time, provide better recall of information between surveys because successive interviews constitute useful benchmarking events, and are generally capable of providing more frequently updated estimates for the target parameters. Also, shorter survey intervals are better for monitoring more volatile target parameters. On the other hand, the shorter the interval between surveys, the larger the burden on respondents with overlapping surveys, which may increase nonresponse and lead to response conditioning, a well-known source of bias in repeated surveys. Longer survey intervals may suffice for less volatile parameters.

Labor force surveys provide an example where the key outputs are required monthly or quarterly, given the need to assess how employment and unemployment totals and rates evolve in the short term. In the European Union, member states are required to carry out continuous labor force surveys, that is, surveys which measure the labor force status of the people every week during the year, to report the results of such surveys at least quarterly (see the Council of the European Union, 1998; the European Parliament and Council of the European Union, 2002). It is interesting to note that the key concept in an LFS requires establishing each person's economic activity status in a specified reference week. For this reason, several countries conduct their LFS in a fixed or prespecified week every month, with the reference week being the week before the interview (see Table 1 below). This however places a heavy burden on the statistical agency, which then must have a workforce capable of handling all the required data collection of a country's LFS within a single (or sometimes two) week(s). Other countries, while still aiming to measure the same concept, use moving reference weeks to be able to spread the data collection activities over a longer period of time. In the United Kingdom, the sample for a quarter is split into 13 weekly assignments to create an efficient fieldwork design, and hence the sample size for every week is about 7.7% of the total sample size for a quarter. However, this choice has implications for both the estimation and analysis of the resulting indicators, which we do not discuss here (for further details, see Steel, 1997).

At the other end of the spectrum, demographic and health surveys are carried out with intervals of up to five years between successive survey waves because the main parameters of interest in these surveys are expected to vary slowly. The same is true for the case of household income and expenditure surveys in many countries—see, for example, the Seventeenth International Conference of Labour Statisticians (2002)—although there is a trend to increasing the frequency of such surveys in other countries.

A quick note on retrospective versus prospective data collection designs: for most repeated surveys, prospective data collection designs are adopted. We call a design

Table 1
Summary of characteristics of selected LFS rotation designs

| Item | Survey | | | | | |
|---|---|---|---|---|---|---|
| | U.S. CPS | Canadian LFS | Australian LFS | U.K. LFS | Japan LFS | Brazilian LFS |
| Frequency | Monthly | Monthly | Monthly | Quarterly | Monthly | Monthly |
| Reference week | Fixed | Fixed | Fixed | Moving | Fixed | Moving |
| Collection period | 1 week | 1 week | 2 weeks | 13 weeks | 2 weeks | 4 weeks |
| Rotation design | 4-8(2) | 6-0(1) | 8-0(1) | 1-2(5) | 2-10(2) | 4-8(2) |
| Monthly overlap (%) | 75 | 83 | 88 | 0 | 50 | 75 |
| Quarterly overlap (%) | 25 | 50 | 63 | 80 | 0 | 25 |
| Yearly overlap (%) | 50 | 0 | 0 | 20 | 50 | 50 |
| Sample size (HH) per month | 60,000 | 53,000 | 29,000 | 20,000 | 40,000 | 41,600 |

prospective when the data will be collected for a period similar to the current reference period every time a household is sampled. A design is called retrospective if in any given survey wave (say the first one), the household is asked to provide data for several past reference periods (say if this is a month, data are required for at least two previous months, which may or may not be the latest ones). Retrospective designs can be useful if there is a need to limit the total number of visits to a household, and yet information needs to be available on an individual basis for different time periods so that some form of longitudinal analysis is possible. However, caution is required given the well-known adverse effects of respondent recall of information for periods not too close to the time of interview or for information regarding events that may not be easily remembered.

### 3.3. *Estimation strategies for some basic objectives and rotation schemes*

Cochran (1977) in Chapter 12 considered the case when $\theta_t = \overline{Y}_t$ is the population mean of a survey variable $y$, that is, the target parameter is the *level at each time point.* Under SRS from the population at each time period, there are gains to be made from using samples with some overlap in adjacent survey waves, but these gains are modest unless the correlation $\rho$ of the measurements of the survey variable $y$ in two successive time periods is high (say bigger than 0.7). The gains in efficiency are made by using an estimator that combines, in an optimal way, the mean of the unmatched portion of the sample at time $t$ with a regression estimator of the mean based on the portion of the sample at time $t$, which is matched to units in the sample at time $t-1$. In this case, the optimal proportion of sample overlap between two successive survey waves would not exceed 50%, and this would be the limiting proportion of overlap required to maximize efficiency gains.

If the target parameter is the *change* between times $t$ and $t-1$, then the best rotation design requires matching larger proportions of sampling units in successive survey waves. Using more than 50% overlap would not be optimal for level estimation but would not result in substantial losses in efficiency compared with the estimators of level under the optimal overlap for level. Hence, in a survey where both estimation of level and change are important, sample overlap more than 50% may be used as a compromise to obtain large efficiency gains for the estimates of change and modest efficiency gains for estimates of level.

If cost considerations are added, in household surveys it is often the case that the cost of the first interview of a selected household is higher than in subsequent occasions. For example, in many countries, labor force surveys have most or all of their first interviews conducted in person, and most subsequent interviews are carried out over the telephone (see the U.S. CPS, the Canadian LFS, the U.K. LFS, the Australian LFS, etc.). In such cases, cost considerations would suggest retaining the largest possible portion of the sample in two successive survey waves. However, one has to consider the added burden of keeping the same respondents in the sample for several waves and the impact this may have on nonsampling errors, such as potential increases to nonresponse and attrition rates, as well as other more subtle forms of errors such as "panel conditioning"—see Chapter 5 for definitions of technical terms.

Many commonly used rotation designs have the number of times in sample equal to at most 8, which means a maximum overlap of samples in successive waves of 87.5% (as is the case in the Australian LFS, which is a monthly survey using a rotation scheme called in-for-8, where a selected household is in the sample for eight consecutive months).

Now consider a repeated survey where the target parameter is an average level over three waves, represented here by $\bar{\theta}_{t,3} = (\theta_{t-1} + \theta_t + \theta_{t+1})/3$. Here, the best rotation design is selecting independent samples every time, because with this design, the variance of the average is simply 1/3 of the average of the variances of the estimates for the individual survey periods. An example of a survey where the prime target is estimating averages is the ACS, designed to replace the "long form" sample in the decennial census in the United States. The idea is that survey data for periods of five years should provide equivalent data to those formerly obtained using the decennial census sample.

The above discussion reveals that precise knowledge of the key survey inference objectives is required for an efficient rotation design to be selected. Estimating averages over time (objective (c)) requires independent or non-overlapping samples at each survey wave, whereas estimating change (objective (b)) requires samples with high overlap in the survey comparison periods (base and current). Estimating level (objective (a)) suggests that a moderate amount of overlap is required but it retains some efficiency gains compared with independent samples even if the overlap is somewhat bigger than the optimal.

The advice above is based on variance considerations only. Repeated large-scale household surveys must often satisfy several of these objectives and for different survey characteristics. Hence, design choices are more complex and less likely to be based only on variance efficiencies. There are several reasons for not using completely independent repetition of cross-sectional designs across time. First, sample preparation costs and time are likely to be substantially bigger. For example, in many household surveys, there are substantial costs associated with selection of new PSUs, such as listing or frame updating costs, staff hiring or relocating, infrastructure, etc. Second, if a survey is repeated over time, it is not unlikely that users will use the survey results for comparisons over time and completely independent surveys would be very inefficient for this purpose. In addition, nonsampling error considerations must also play an important part in specifying rotation designs such that respondent burden, attrition and measurement error are kept under control.

### 3.4. Examples of non-overlapping repeated surveys

In this section, some alternative rotation designs used by some major household surveys around the world are highlighted to illustrate how different objectives lead to different design options. We start by describing perhaps the largest repeated cross-sectional survey (i.e., no overlap of samples in adjacent survey waves) in existence: the ACS, see U.S. Census Bureau, 2006b.

The ACS selects a stratified systematic sample of addresses. The survey has a five-year cycle, and each address sampled in a given year is deliberately excluded from samples selected in the four subsequent years (negative coordination of samples in successive years). This is achieved by randomly allocating each address in the Master Address File used as the frame for the survey into one of five subframes, each containing 20% of the addresses in the frame. Addresses from only one of the subframes are eligible to be in the ACS sample in each given year, and a subframe can be used only once in every five years. New addresses are randomly allocated to one of the subframes.

The main objective of the ACS is to provide estimates for small areas, replacing the previous approach of using a long form questionnaire for a large sample of households collected during the Decennial Censuses in the United States. It was designed to provide for sample accumulation over periods of up to five years, after which the sample for each small area would be of similar size to what would have been obtained in the Decennial Census. The survey data are then used to estimate parameters that can be seen as moving averages of five years, with five years of survey data being used to provide estimates for the smallest areas for which results are published, and fewer years being used to provide estimates for broader geographies. Currently, single-year estimates are published annually for areas with a population of 65,000 or more. Multi-year estimates based on three successive years of ACS samples are published for areas with populations of 20,000 or more. Multi-year estimates based on five successive years of ACS samples will be published for all legal, administrative, and statistical areas down to the block-group level regardless of population size.

Another very large survey using repeated non-overlapping cross-sectional samples is the French Population Census. From 2004 onwards, the "census" of France's resident population started using a new approach, which replaced the traditional enumeration previously conducted every eight or nine years. The 1999 general population census of France was the last one to provide simultaneous and exhaustive coverage of the entire population. The new "census" in fact uses a large sample stratified by area size and requires cumulating data over five years to provide national coverage. Data are collected in two months every year (during January–February). All small municipalities (those with fewer than 10,000 inhabitants) are allocated to one of five "balanced" groups. For each group of small municipalities, a comprehensive census (no subsampling of dwellings or households) is carried out once every five years. Large municipalities (those with 10,000 or more inhabitants) carry out a sample survey of about 8% of their population every year. So at the end of a five-year cycle, every small municipality has carried out a census and, in the large municipalities, a sample of around 40% of the households will be available. This comprehensive sample is then used to replace the previous census for all purposes. Once the system is fully in place, rolling periods of five years may be used to provide census-like results, which were previously updated only once every eight or nine years.

Household income and expenditure surveys in many countries provide another important example of repeated cross-sectional surveys that use non-overlapping designs. The current recommendations issued by the International Labour Organization (ILO) on this type of survey (see the Seventeenth International Conference of Labour Statisticians, 2002) specify that such surveys should be conducted with intervals not exceeding five years. The recommendations do not specify that non-overlapping designs are required, but in many countries, this is the preferred method, given the considerable burden that such surveys place on participating households.

Another type of repeated survey design implemented around the world is the use of a panel sample of PSUs, with complete refreshment of the list of households sampled in the selected PSUs. The series of Demographic and Health Surveys (DHS) adopts this design whenever possible. Here, the gains from retaining the same set of PSUs are not as important in terms of variance reduction for estimates of change as if the same households were retained, but there are potential advantages in terms of costs of survey taking and also perhaps less volatile estimates of change between successive waves of this survey in a given country or region. For a more detailed discussion, see Macro International (1996, p. 29).

### *3.5. Rotation designs in labor force surveys*

Labor force surveys conducted in most countries provide the most prominent application of overlapping repeated survey designs. For such surveys, intervals between survey waves are usually very short (months or quarters in most countries). In some countries, data collection is continuous throughout the year and publication periods may again be monthly or quarterly. In the United States and Canada, monthly surveys are used, with a single reference week every month. Rotation designs for these surveys are 6-0(1)[1] for the Canadian LFS, and 4-8(2) for the U.S. CPS. In both surveys the estimation of change in labor force indicators between adjacent months is a prominent survey objective. In both countries, some form of composite estimation (see section 5.3) is used to estimate the indicators of interest. Seasonally adjusted estimates derived from the time series of the composite estimates are also published, and are prominent in the analysis of survey results contained in monthly press releases issued by the corresponding statistical agencies.

In Australia, the LFS uses the 8-0(1) rotation design, and the key estimates highlighted in the publications are the estimates of the trend derived from the time series of the sample estimates. This is a unique example of a survey where the major indicators are based on time series modeling of the basic survey estimates. Because the targets for inference here are not simply the values of the unknown parameters, but of rather complex functions of these (the trend of the corresponding time series), this brings in some interesting design issues. McLaren and Steel (2001) studied options for designs for surveys where the key objective is trend estimation and concluded that monthly surveys using rotation designs 1-2(m) are the best. This study illustrates quite clearly the impact

[1] We use the convention *in-out* (*times*) to denote the number of waves that a household is included in the sample, then the number of waves that it is left *out* of the sample, and the number of *times* that this pattern is repeated. A similar convention was proposed by McLaren and Steel (2001).

that the choice of objectives has on rotation designs: if the trend is the main target, monthly surveys need not have monthly overlap. The same is not true, though, if the target parameter is the simple difference of the relevant indicators, without any reference to the underlying trend. The U.K. LFS is a quarterly survey, originally motivated by the aim of measuring quarter-on-quarter change, using a rotation design which may be described as equivalent to a monthly 1-2(5), whereby the sampled households enter the survey with an interview on a single month of the quarter, rest for two months, then return for another four successive quarters. Interestingly, this same rotation pattern (more precisely, a 1-2(8) pattern) is used by the Canadian LFS in Canada's three northern territories, where three-month estimates are published.

Table 1 displays some information on key aspects of rotation designs used in labor force surveys around the world. The U.K. LFS is a model close to the LFS design adopted throughout the European Union. The use of this model for LFSs has spread beyond the European Union. In 2005–2006, Statistics South Africa started an ambitious project to replace its semiannual LFS with a quarterly survey with a rotation design in 2008. A similar project is under way in Brazil, where an integrated household survey using a 1-2(5) rotation design will replace the current annual national household survey and the monthly LFS in 2009.

### *3.6. Some guidance on efficiency gains from overlapping repeated surveys*

Once a decision has been taken that the survey has to have some sample overlap over time, it becomes important to decide how much and which methods to use to control how the sample evolves over successive survey occasions. Cochran (1977, Section 12.11) provides some useful guidance on the choice of how much overlap to have. His results are all based on an assumed SRS design. Most household surveys use more complex sample designs. However, the sample design structure (stratification, clustering, sample sizes, selection probabilities, and estimator) is often held fixed over time. We can express the variance of survey estimates at each time point in terms of the product

$$V_p(\hat{\theta}_t) = V_{\mathrm{SRS}}(\hat{\theta}_t) \times \mathrm{deff}_t, \tag{6}$$

where $V_p(\hat{\theta}_t)$ is the variance of the survey estimator under the complex survey design adopted to carry out the survey, $V_{\mathrm{SRS}}(\hat{\theta}_t)$ is the variance that the survey estimator would have under a simple random sample design with the same sample size, and $\mathrm{deff}_t$ is the corresponding design effect. If we assume that the design effect is approximately constant over time (i.e., $\mathrm{deff}_t = \mathrm{deff}\ \forall t$), then the advice provided for SRS is relevant to compare the relative merits of complex surveys for alternative rotation designs.

Suppose that a SRS of size $n$ is used on two successive occasions ($t$ and $t+1$) and that the population is assumed fixed (no changes due to births or deaths), but the measurements may change. On the second occasion, a SRS of $m < n$ units sampled at $t$ are retained (overlap part) and $n - m$ units are replaced by newly selected ones, also sampled using SRS, from the units not sampled at $t$. Under this scenario and assuming that the finite population correction can be ignored, the variance of the sample mean $\overline{y}_t$, the simplest estimator of the population mean $\overline{Y}_t$ (the level) of a survey variable $y$ at each time point $t$, is given by

$$V_{\mathrm{SRS}}(\overline{y}_t) = S_t^2 / n, \tag{7}$$

where $S_t^2$ is the population variance of the survey variable $y$ at time $t$. Clearly, the variances of the simple estimates of level do not depend on $m$, the size of the matched or overlapping portion of the sample at time $t+1$. However, assuming that the variance of the survey variable is constant over time, that is, $S_t^2 = S_{t+1}^2 = S^2$, it follows that the variance of the estimate of change, namely, the difference in the population means, is given by

$$V_{\mathrm{SRS}}(\overline{y}_{t+1} - \overline{y}_t) = 2\frac{S^2}{n}\left(1 - \frac{m}{n}\rho\right), \tag{8}$$

where $\rho$ is the correlation of observations of the survey variable in two adjacent time periods.

As $\rho$ is often positive, it becomes clear that no overlap ($m = 0$) is the least efficient strategy for estimation of change and that a panel survey (complete overlap or $m = n$) is the most efficient, with a reduction factor of $1 - \rho$. The overlap fraction $m/n$ provides an attenuation of the variance reduction when a rotation design with less than 100% overlap is adopted.

Now with some sample overlap, there are alternative estimators of the level on the second (and subsequent) occasion(s). If an optimal estimator (see Cochran, 1977, eq. 12.73) is used with optimal weight and optimum matching proportion of $m/n = \sqrt{1-\rho^2}\big/\big(1+\sqrt{1-\rho^2}\,\big)$, its variance would be given approximately by

$$V_{\mathrm{SRS}}\left(\overline{y}_{t+1}^{\mathrm{opt}}\right) = \frac{S^2}{n}\frac{\left(1+\sqrt{1-\rho^2}\right)}{2}. \tag{9}$$

For values of $\rho$ above 0.7, the gains are noticeable, and the optimum matching proportion is never bigger than 50%.

The above discussion demonstrates a clear link between the selection of an estimator and the choice of a rotation design, particularly in terms of the proportion of sample overlap, given a specified survey objective (in the above, estimation of the current population mean). This discussion is at the heart of substantial developments in the literature on estimation from repeated surveys, reviewed in Binder and Hidiroglou (1988), and subsequently, in Silva and Cruz (2002). Some large-scale repeated surveys make use of composite estimators (see Section 5.3), and the prime examples are again the U.S. CPS (see U.S. Census Bureau and U.S. Bureau of Labor Statistics, 2002), and the Canadian LFS (see Gambino et al., 2001).

After reviewing efficiency gains for estimators of both level and change under the simplified SRS scenario discussed above, Cochran (1977, p. 354) suggests that "retention of 2/3, 3/4, or 4/5 from one occasion to the next may be a good practical policy if current estimates and estimates of change are both important." But this large overlap of successive surveys will only be advantageous if the estimators utilized are capable of exploiting the survey overlap as would the "optimal" estimator discussed above.

We conclude this section by pointing out that in addition to considerations of sampling error, it is essential that designers of repeated surveys consider the implications of alternative rotation designs in terms of nonsampling errors. The longer the households are retained in the sample, the more likely they are to drop out (attrition/nonresponse), as well as to start providing conditioned responses (measurement error). After a certain point, the combined adverse effects of nonsampling errors are more likely to overshadow any marginal gains in efficiency, so it is vital not to extend the length of survey participation beyond this point.

## 4. Data collection

The traditional modes of data collection for household surveys are personal interview, telephone interview, and questionnaire mail out followed usually by self-completion by the respondent. Recently, a variety of new methods have started to be used. These include use of the internet (html questionnaires), the telephone (where the respondent enters his replies using the telephone keypad), and self-completion using a computer (the respondent either enters his responses using the keyboard or gives them orally, and the computer records them).

Gradually, the use of paper questionnaires is diminishing and being replaced by computer-based questionnaires. The latter have several advantages, including the possibility of having built-in edits that are processed during each interview and the elimination of the data capture step needed for paper questionnaires. An important effect of these changes is that the file of survey responses is relatively clean from the outset. Computer-based questionnaires also make it possible to have very complex questionnaires with elaborate skip patterns. Even in some very large-scale household surveys, such as national population censuses, computers are now being used instead of traditional paper questionnaires: this has been the case in Colombia and Brazil, where handheld computers were used for population censuses in 2006 and 2007. The Brazilian case illustrates the potential for such devices to affect data collection because for the first time, the population census and an agricultural census have been integrated into a single field operation, with households in the rural area providing both the population and agricultural census information in a single interview (for details see the web site of the Brazilian official statistics agency IBGE).

One benefit of using computer-assisted interviewing that is receiving increased attention is the wealth of information about the data collection process that it makes available. Such data about the data collection process, and more generally about other aspects of the survey process, are referred to as paradata (see Scheuren, 2005). This information can be invaluable to improving the collection process. It provides answers to questions such as: which parts of the questionnaire are taking the most time? Which questions are being corrected most often? Which are triggering the most edit failures? This information can then be used to review concepts and definitions, improve the questionnaire, improve interviewer training, and so on. Granquist and Kovar (1997) advocate the use of such information as one of the primary objectives of survey data editing, but one that is often not so vigorously pursued in practice. The ultimate goal is to improve the survey process in subsequent surveys or survey waves.

A major change in field operations in developed countries has been the increase in the proportion of interviews conducted by telephone rather than in person. This has had an impact on both the types of people who conduct interviews and the way their workload is organized. For example, when most interviews were conducted in person, often in the daytime, the interviewer needed a car. With the introduction of computer-assisted interviewing from a central facility, the interviewer no longer needs a car and the number of evening interviews can increase since it is possible to have an evening shift that conducts interviews in different time zones. As a result of factors such as these, the demographic characteristics of interviewers have changed (e.g., the number of university students working part-time as interviewers has increased in Canada).

The introduction of computer-assisted interviewing from a central facility also makes it possible to monitor interviews as they happen. In some statistical agencies, elaborate quality assurance programs based on monitoring have been introduced to improve the data collection process (identify problems, target interviewer training needs, etc.).

Monitoring interviews conducted in a central facility has benefits, but until recently, the computers used for personal interviewing, namely, laptop computers and handheld devices, were not powerful enough to implement something similar to monitoring for personal interviews. However, recording of personal interviews (on the same computer used to enter responses to survey questions) has now become feasible. Biemer et al. (2000) discuss the application of computer audio-recorded interviewing (CARI) to the National (United States) Survey of Child and Adolescent Well-being. Another form of interview monitoring that has recently become feasible is to use computers equipped with a Global Positioning System device, which can be used to record the coordinates of dwellings visited for interview at the time of arrival or at the start of the interview. This enables survey organizations to supervise work in ways that were not previously feasible with paper-and-pencil type interviews. Devices like these were used to carry out a mid-decade population census in Brazil and are being considered for the redesigned South African LFS.

The use of computers in survey sampling extends well beyond computer-assisted interviewing. Since the whole survey process can be monitored using software tools, this opens new possibilities for improving data collection. In addition to the tools already mentioned in this section, such as live monitoring of interviews, computers make it easier to keep track of progress on many fronts, such as response rates by various categories (geographical, age group, etc.). Hunter and Carbonneau (2005) provide a high-level overview of what they refer to as active management.

The increased use of telephone interviewing is motivated by cost considerations, but it also introduces problems. We have already mentioned problems associated with the use of mobile telephones in Section 2.1 in the context of frame coverage. There are other aspects of telephone interviewing that make it more difficult to get a response. These include the use of answering machines and call display (caller ID) to screen calls and the apparently greater difficulty for the interviewer to establish a rapport with the respondent by telephone than in person. The recent Second International Conference on Telephone Survey Methodology, held in 2006, was devoted to the subject. A monograph containing selected papers from the conference will be published.

In recent years, there has been increased interest in finding better ways to survey rare populations (or, more accurately, small groups within larger populations) and groups that are difficult to survey, such as the homeless and nomadic populations. Statistics Canada (2004) devoted a methodology symposium to the topic and some of the latest research in this area is covered in the proceedings of that conference.

### *4.1. Combining data from different sources*

Another way to reduce survey costs is to try to make use of existing data and, in particular, data from administrative sources. We focus on the combined use of administrative and survey data, although there are cases where administrative data can be used on their own. In addition to its low cost (since it was collected for some other purpose and therefore already "paid for"), a great benefit of using administrative data is the reduction in burden on survey respondents. In addition, if the concepts used by the survey (e.g., to define income) are close to those on which the administrative data are based, then the administrative data may be more accurate than the same data obtained via a survey. We give two examples of the integration of survey and administrative data. The Canadian Survey of Labour and Income Dynamics (SLID) asks survey respondents whether they prefer to answer several questions on income or, alternatively, to give permission to Statistics Canada to access the information from their income tax records. In recent surveys, 85% of respondents have chosen the latter option. Because of this success, the approach is being extended to other surveys such as the Survey of Household Spending and even the census of population.

A second example involves the long-standing longitudinal Survey of Income and Program Participation (SIPP) conducted by the U.S. Census Bureau, which is being replaced by the Dynamics of Economic Wellbeing System (DEWS). The DEWS will make extensive use of administrative data files to augment survey data (see U.S. Census Bureau, 2006a). A key motivating factor behind this change is the reduction of both costs and response burden. By using administrative data files over time, the new approach also avoids the problem of attrition common to all longitudinal surveys. More generally, data can be combined across multiple surveys and administrative sources. This is often the only way to obtain adequate estimates for small domains. Two sessions at Statistics Canada's 2006 methodology symposium included papers on this topic (see Statistics Canada, 2006).

Merkouris (2004) presented a regression-based method for combining information from several surveys. This method is essentially an extended calibration procedure whereby comparable estimates from various surveys are calibrated to each other, accounting for differences in effective sample sizes. The method has been applied successfully: data from the Canadian SLID was used to improve estimates for the much smaller Survey of Financial Security. Merkouris (2006) adapted the procedure to small domain estimation. A related approach was adopted in the Netherlands (Statistics Netherlands, 2004) to compile a whole "virtual census" using data from several sources, such as a population register, some other administrative records, and selected household surveys. The methodology, called *repeated weighting*, is described in detail in Houbiers et al. (2003).

## 5. Weighting and estimation

### *5.1. Simple estimation of totals, means, ratios, and proportions*

Estimation in household sample surveys is often started using "standard" weighting procedures. Assuming that

- a two stage stratified sampling design was used to select a sample of households,
- every member in each selected household was included in the survey, and
- the sample response was complete,

then the standard design-weighted estimator for the population total $Y = \sum_h \sum_{i \in U_h} \sum_{j \in U_{hi}} y_{hij}$ of a survey variable $y$ has the general form

$$\hat{Y} = \sum_{h} \sum_{i \in s_h} \sum_{j \in s_{hi}} d_{hij} y_{hij}, \tag{10}$$

where $d_{hij}$ is the design weight for household $j$ of PSU $i$ in stratum $h$, $y_{hij}$ is the corresponding value for the survey variable $y$, $U_h$ and $s_h$ are the population and sample sets of PSUs in stratum $h$, respectively, of sizes $N_h$ and $n_h$, $U_{hi}$ and $s_{hi}$ are the population and sample sets of households in PSU $i$ of stratum $h$, having sizes $M_{hi}$ and $m_{hi}$, respectively.

The design weight $d_{hij}$ is the reciprocal of the inclusion probability of household $j$ of PSU $i$ in stratum $h$, which can be calculated as the product of the inclusion probability $\pi_{hi}$ for PSU $hi$ and the conditional probability $\pi_{j|hi}$ of selecting household $j$ given that PSU $hi$ is selected. Design weights for multistage designs having more than two stages of selection can be computed using similar recursion algorithms where each additional stage requires computing an additional set of inclusion probabilities conditional on selection in preceding stages.

Although the design weight is simply the reciprocal of a unit's inclusion probability, in practice, its computation can be quite involved. For example, in the relatively simple case of the Canadian LFS, the design weight is the product of the following: the first-stage (PSU) inclusion probability, the second-stage (dwelling) inclusion probability, the cluster weight (a factor, usually equal to 1, that accounts for subsampling in PSUs whose population has increased significantly since the last redesign), and the stabilization weight (a factor to account for the high-level subsampling that the LFS uses to keep the national sample size stable over time; see Section 2.5 in this chapter). For some surveys, the computation of design weights can be much more complex than this, particularly for longitudinal surveys. In addition, there are further adjustments to the design weight needed to account for nonresponse and, in some cases, for coverage errors, unknown eligibility (in RDD surveys, for example), and so on.

Estimators of totals similar to (10) are available for designs having any number of stages of selection. The estimator of the population mean $\overline{Y} = \sum_h \sum_{i \in U_h} \sum_{j \in U_{hi}} y_{hij} \Big/ \sum_h \sum_{i \in U_h} M_{hi}$ would be obtained simply by substituting the design-weighted estimators of the totals in the numerator and denominator leading to

$$\overline{y}_d = \sum_{h} \sum_{i \in s_h} \sum_{j \in s_{hi}} d_{hij} y_{hij} \Big/ \sum_{h} \sum_{i \in s_h} \sum_{j \in s_{hi}} d_{hij}. \tag{11}$$

For most household surveys, the overall population size $M_0 = \sum_h \sum_{i \in U_h} M_{hi}$ is not known and the estimator that could be obtained from (14) by replacing the estimated population size in the denominator by $M_0$ is not available. However, even if this alternative estimator was available, (11) would still be the usual choice because for many survey situations encountered in practice, it would have smaller variance. Note that (11) is a special case of the estimator

$$\hat{R}_d = \sum_{h} \sum_{i \in s_h} \sum_{j \in s_{hi}} d_{hij} y_{hij} \Big/ \sum_{h} \sum_{i \in s_h} \sum_{j \in s_{hi}} d_{hij} x_{hij} \tag{12}$$

for the ratio of population totals $R = Y/X = \sum_h \sum_{i \in U_h} \sum_{j \in U_{hi}} y_{hij} \Big/ \sum_h \sum_{i \in U_h} \sum_{j \in U_{hi}} x_{hij}$, where the variable $x$ in the denominator is equal to 1 for every household. Another special case of interest occurs when the survey variable is simply an indicator variable. In this case, its population mean is simply a population proportion, but the estimator (11) is still the estimator applied to the sample observations of the corresponding indicator variable.

The above estimators would be used also to obtain estimates on characteristics of persons simply by making the corresponding $y$ and $x$ variables represent the sum of the values observed for all members of sampled households. This assumes that survey measurements are taken for every member of sampled households, which is a common situation in practice. However, if subsampling of household members takes place, there would be an additional level of weighting involved, but the above general approach to estimation would still apply.

### 5.2. Calibration estimation in household surveys

Despite their simplicity, such design-weighted estimators are not the ones most commonly used in the practice of household surveys. Instead, various forms of calibration estimators (Deville and Särndal, 1992) are now commonly used. Calibration estimators of totals (see Chapter 25) are defined as

$$\hat{Y}_{\mathrm{C}} = \sum_h \sum_{i \in s_h} \sum_{j \in s_{hi}} w_{hij} y_{hij}, \tag{13}$$

where the weights $w_{hij}$ are such that they minimize a distance function

$$F = \sum_h \sum_{i \in s_h} \sum_{j \in s_{hi}} G\big(w_{hij}, d_{hij}\big) \tag{14}$$

and satisfy the calibration equations

$$\sum_h \sum_{i \in s_h} \sum_{j \in s_{hi}} w_{hij} \mathbf{x}_{hij} = \mathbf{X}, \tag{15}$$

where $\mathbf{x}_{hij}$ is a vector of auxiliary variables observed for each sampled household, $\mathbf{X}$ is a vector of population totals for these auxiliary variables, assumed known, and $G(w, d)$ is a distance function satisfying some specified regularity conditions. A popular choice of distance function is the standard "chi-square" type distance defined as

$$G\left(w_{hij}, d_{hij}\right) = \left(w_{hij} - d_{hij}\right)^2 / q_{hij} d_{hij}, \tag{16}$$

where the $q_{hij}$ are known constants to be specified.

Calibration estimators for ratios and means (as well as proportions) follow directly from using the weights $w_{hij}$ in place of the design weights $d_{hij}$ in the expressions (11) and (12).

Calibration estimators have some desirable properties. First, weights satisfying (15) provide sample "estimates" for the totals of the auxiliary variables in $\mathbf{x}$ that match exactly the known population totals for these variables. If the population totals of the auxiliary variables have been published before the survey results are produced, then using calibration estimators for the survey would guarantee that the survey estimates are coherent with those already in the public domain. This property, although not essential from an estimation point of view, is one of the dominant reasons why calibration is so often used in household surveys. It appeals to survey practitioners in many instances as a way of enforcing agreement between their survey and publicly available totals for key demographic variables. Särndal calls this the cosmetic property of calibration estimators.

The second desirable property is *simplicity*, namely the fact that given the weights $w_{hij}$ calibration estimates are linear in $y$. This means that each survey record can carry a single weight to estimate all survey variables. Calculation of the estimates for totals, means, ratios, and many other parameters is straightforward using standard statistical software, after the calibration weights have been obtained and stored with each household survey record. In the case of some commonly used distance functions, the calibrated weights are given in a closed form expression and are easy to compute using a range of available software (e.g., CALMAR, GES, BASCULA, G-CALIB-S, R survey package, etc.).

The third property of such calibration estimators is their *flexibility* to incorporate auxiliary information that can include continuous, discrete, or both types of benchmark variables at the same time. If the auxiliary totals represent counts of the numbers of population units in certain classes of categorical (discrete) variables, then the values of the corresponding **x** variables are simply indicators of the units being members of the corresponding classes. Cross-classification of two or more categorical variables can also be easily accommodated by defining indicator variables for the corresponding combinations of categories.

Calibration estimators also yield some degree of *integration* in the sense that some widely used estimators are special cases, for example, ratio, regression, and poststratification estimators (Särndal et al., 1992, Chapter 7) as well as incomplete multiway poststratification (Bethlehem and Keller, 1987).

In addition, if the calibration is performed at the level of the household, all members of the same household will have a common calibration weight $w_{hij}$, which is a “natural” property since this is the case for the original design weights $d_{hij}$. If there are auxiliary variables referring to persons, such as age and sex, the calibration at the household level is still possible, provided the auxiliary variables **x** include the counts of household members in the specified age-sex groups for which population auxiliary information is available. This is the approach called “integrated household (family) weighting” by Lemaitre and Dufour (1987).

These are powerful arguments for using calibration estimators. However, when doing so, users must be aware of some difficulties that may be encountered as well. Some of the issues that should be of concern when performing calibration estimation in practice include as follows:

- Samples are often small in certain weighting classes;
- Large numbers of “model groups” and/or survey variables;
- Negative, small (less than 1) or extreme (large) weights;
- Large number of auxiliary variables;
- Nonresponse;
- Measurement error.

The last issue in this list (measurement errors and their effect on calibration) is discussed in Skinner (1999). All the other issues are considered in Silva (2003).

Calibration estimators may offer some protection against nonresponse bias. Poststratification and regression estimation, both special cases of calibration estimators, are widely used techniques to attempt to reduce nonresponse bias in sample surveys. Särndal and Lundström (2005) even suggest "calibration as a standard method for treatment of nonresponse." Calibration estimators are approximately design unbiased if there is complete response for any fixed choice of auxiliary variables. Under nonresponse bias, however, calibration estimators may be biased even in large samples. Skinner (1999) examined the impact of nonresponse on calibration estimators. His conclusions are as follows:

- "the presence of nonresponse may be expected to lead to negative weights much more frequently";
- "the calibration weights will not converge to the original design weights as the sample size increases";
- "the variance of the calibration estimator will be dependent on the distance functions $G(w, d)$ and revised methods of variance estimation need to be considered."

The intended bias reduction by calibration will only be achieved, however, if the combined nonresponse and sampling mechanisms are ignorable given the $\mathbf{x}$ variables considered for calibration. This suggests that the choice of $\mathbf{x}$ variables has to take account of the likely effects of nonresponse, and in particular, should aim to incorporate all $\mathbf{x}$ variables for which auxiliary population data is available that carry information about the unknown probabilities of responding to the survey.

The bias of the calibration estimator will be approximately zero if $y_{hij} = \boldsymbol{\beta}'\mathbf{x}_{hij}$ for every unit in the population, with a nonrandom vector $\boldsymbol{\beta}$ not dependent on the units. (e.g., see Bethlehem, 1988; Särndal and Lundström, 2005, and also Chapter 15 of Särndal et al., 1992). In household surveys, this is an unlikely scenario, and even under models of the form $y_{hij} = \boldsymbol{\beta}'\mathbf{x}_{hij} + \varepsilon_{hij}$, the residuals may not be sufficiently small to guarantee absence of bias due to nonresponse. Särndal and Lundström (2005, Section 9.5) examine additional conditions under which calibration estimators are nearly unbiased and show that if the reciprocals of the response probabilities are linearly related to the auxiliary variables used for calibration, then the calibration estimators will have zero "near bias."

Hence, the key to successfully reduce nonresponse bias in estimating for household surveys is to apply calibration estimation using auxiliary variables that are good linear predictors of the reciprocals of the response probabilities.

Gambino (1999) warns that "nonresponse adjustment can, in fact, increase bias rather than decreasing it," and consequently, that "the choice of variables to use for nonresponse adjustment should be studied even more carefully in the calibration approach than in the traditional approach" for nonresponse compensation.

## *5.3. Composite estimation for repeated household surveys*

For repeated surveys with partial overlap of sample over time, we can use information for the common (matching) sample between periods to improve estimates for the current period $t$, as we saw in Section 3.6. The common units can be used to obtain a good estimate of change $\hat{\Delta}_{t-1,t}$ between periods $t-1$ and $t$, which can then be added to the estimate $\hat{\theta}_{t-1}$ to produce an alternative estimate to $\hat{\theta}_t$. An optimal linear combination of these two estimates of $\theta_t$ is referred to as a composite estimate. The U.S. CPS has used such estimates since the 1950s. Initially, the CPS used the $K$-composite estimator

$$\hat{\theta}'_t = (1-K)\hat{\theta}_t + K(\hat{\theta}'_{t-1} + \hat{\Delta}_{t-1,t})$$

with $K = 1/2$. This was later replaced by the $AK$-composite estimator

$$\hat{\theta}'_t = (1-K)\hat{\theta}_t + K(\hat{\theta}'_{t-1} + \hat{\Delta}_{t-1,t}) + A(\hat{\theta}_{u,t} - \hat{\theta}_{m,t})$$

with $A = 0.2$ and $K = 0.4$, where $m$ and $u$ denote the matched and unmatched portions of the sample (see Cantwell and Ernst, 1992). Note that the term on the far right involves the difference between estimates for the current time point based on the current unmatched and matched samples, respectively. One drawback to using the K- and AK-composite estimators is that the optimal values of $A$ and $K$ depend on the variable of interest. Using different values for different variables will lead to inconsistencies in the sense that parts will not add up to totals (e.g., labor force $\neq$ employed $+$ unemployed). One solution to this problem, called composite weighting, was introduced into the CPS in 1998. Coefficients of $A = 0.4$ and $K = 0.7$ are used for employed and $A = 0.3$ and $K = 0.4$ are used for unemployed, with Not-in-Labour Force being used as a residual category to ensure additivity. Then, a final stage of raking is used to rake to control totals based on composited estimates (see Lent et al., 1999).

The Canadian LFS introduced a regression (GREG) approach, called regression composite estimation, that does not have the consistency problem and has other benefits as well (see Fuller and Rao, 2001; Gambino et al., 2001; Singh et al., 2001).

To implement regression composite estimation, the $X$ matrix used in regression is augmented by columns associated with *last* month's composite estimates for key variables, that is, some of last month's composite estimates are used as control totals. Thus, the elements of the added columns are defined in such a way that, when the final weights of this month are applied to each new column, the total is a composite estimate from the previous month. Therefore, the final calibration weights will respect both these new control totals and the ones corresponding to the original columns of $X$ (typically, age-sex and geographical area population totals).

There are several ways to define the new columns, depending on one's objectives. In the Canadian LFS, a typical new column corresponds to employment in some industry, such as agriculture. If one is primarily interested in estimates of level, the following way of forming columns produces good results. For person $i$ and times $t-1$ and $t$, let $y_{i,t-1}$ and $y_{i,t}$ be indicator variables that equal 1 whenever the person was employed in agriculture, and 0 otherwise. Then let

$$x_i^{(L)} = \begin{cases} \bar{y}'_{t-1} & \text{if } i \in u \\ y_{i,t-1} & \text{if } i \in m, \end{cases}$$

where $\overline{y}'_{t-1}$ is last month's composite estimate of the proportion of people employed in agriculture. The corresponding control total is last month's estimate of the number of people employed in agriculture, that is, $\hat{Y}'_{t-1}$.Thus, applying the final (regression) weights to the elements of the new column and summing will produce last month's estimate. The superscript $L$ is used as a reminder that the goal here is to improve estimates of level.

If estimates of change are of primary interest, the following produces good results:

$$x_i^{(C)} = \begin{cases} y_{i,t} & \text{if } i \in u \\ y_{i,t} + R(y_{i,t-1} - y_{i,t}) & \text{if } i \in m, \end{cases}$$

where $R = \sum w_i / \sum_m w_i$ and $1/R$ is (approximately) the fraction of the sample that is common between successive occasions.

Using the $L$ controls produces better estimates of level for the variables added to the $X$ matrix as controls. Similarly, adding $C$ controls produces better estimates of change for the variables that are added. Singh et al. (2001) present efficiency gains for $C$-based estimates of level and change and refer to earlier results on $L$-based estimates.

Although we can add both $L$ and $C$ controls to the regression, this would result in a large number of columns in the $X$ matrix, which can have undesirable consequences. Fuller and Rao (2001) proposed an alternative that allows the inclusion of the industries of greatest interest while allowing a compromise between improving estimates of level and improving estimates of change. They proposed taking a linear combination of the $L$ column and the $C$ column for an industry and using it as the new column in the $X$ matrix, that is, use

$$x_i = (1 - \alpha)x_i^{(L)} + \alpha x_i^{(C)}.$$

This is the method currently used by the Canadian LFS. For a discussion on the choice of $\alpha$, see Gambino et al. (2001). They also discuss some of the subtleties involved in implementing the above approach that we have not considered here. Note the importance of having good tracking or matching information for the survey units and also the need to apply composite estimation in line with the periods defining rotation of the sample (i.e., monthly rotation leads to monthly composite estimators, etc.).

### *5.4. Variance estimation*

Variance estimation for multistage household surveys is often done using approximate methods. This happens because sampling fractions are very small, exact design unbiased variance estimators are complex or unavailable (e.g., when systematic sampling is used at some stage), or estimators are not linear. There are two main alternative approaches, which are as follows:

- Approximate the variance and then estimate the approximation;
- Use some kind of resampling or replication methodology.

Wolter (2007) provides a detailed discussion on variance estimation in complex surveys, and many of the examples discussed in the book come from multistage household surveys. Skinner et al. (1989) also discuss in detail variance estimation under complex sampling designs, not only for standard estimators of totals, means, ratios, and proportions but also for parameters in models commonly fitted to survey data.

In the first approach, which we refer to as the "approximation approach," we approximate the variance of the estimator under the complex design assuming that the selection of PSUs (within strata) had taken place with replacement, even though this was not actually the case. If the estimator of the target parameter is linear, this is the only approximation required to obtain a simpler variance expression and then use the sample to estimate this approximate variance. This is the so-called *ultimate cluster* approach introduced by Hansen et al. (1953). If, in addition, the estimators are nonlinear, but may be written as smooth functions of linear estimators (such as estimators of totals), Taylor series methods are used to approximate their variance using functions of variances of these linear estimators obtained under the assumption of with-replacement sampling of PSUs. Obtaining design-unbiased variance estimators for these variance approximations simplifies considerably, and for a large set of designs and estimators, the corresponding variance estimators are available in explicit form and have been incorporated in statistical software. Such software includes special modules in general statistical packages like SAS, SPSS, STATA, and R (see the "survey" package—Lumley, 2004). It also includes specialized packages such as SUDAAN, PC-CARP, and EPI-INFO.

In contrast, resampling methods start from a completely different perspective. They rely on repeatedly sampling from the observed sample to generate "pseudoestimates" of the target parameter, which are subsequently used to estimate the variance of the original estimator. Let $\hat{\theta}$ denote the estimator of a vector target parameter $\theta$, obtained using the "original" survey weights $w_i$. Then a resampling estimator of the variance of $\hat{\theta}$ is of the form

$$\hat{V}_R(\hat{\theta}) = \sum_{r=1}^{R} K_r(\hat{\theta}^{(r)} - \hat{\theta})(\hat{\theta}^{(r)} - \hat{\theta})' \tag{17}$$

for some specified coefficients $K_r$, where $r$ denotes a particular replicate sample selected from $s$, $R$ denotes the total number of replicates used, and $\hat{\theta}^{(r)}$ denotes the pseudoestimate of $\theta$ based on the $r$th replicate sample. These replicate samples may be identified in the main sample data set by adding a single column containing revised weights corresponding to each sample replicate, and for each of these columns, having zero weights for units excluded from each particular replicate sample. The constants $K_r$ vary according to the method used to obtain the replicate samples. Three alternative approaches are popular: *jackknife*, *bootstrap*, and *balanced repeated replication*. For details, see Wolter (2007).

Approximation-based methods are relatively cheap in terms of computation time, provided the survey design and the target parameters are amenable to the approximations, and one has the required software to do the calculations. Their main disadvantages are the need to develop new approximations and variance estimators whenever new estimators are employed and the somewhat complex expressions required for some cases, especially for nonlinear estimators.

Resampling methods are reasonably simple to compute, provided the survey data contains the necessary replication weights. They are, however, more costly in terms of computation time, a disadvantage which is becoming less important with the increase in computer power. In addition, the methods are quite general and may apply to novel situations without much effort from a secondary analyst. The burden here lies mostly on the survey organization to compute and store replicate survey weights with each data record.

Modern computer software is available for survey data analysis that is capable of computing variance estimates without much effort, provided the user has access to the required information on the survey design (Lumley, 2004).

For domain estimates, when the sample within a domain is sufficiently large to warrant direct inference from the observed sample, the general approaches discussed above can be applied directly as well. However, many emerging applications require more sophisticated methods to estimate for small domains (small areas). This topic is covered in a large and growing literature and will not be treated here. We note only that when small area estimation methods are used, the variance estimation becomes more complex. The reader is referred to Rao (2003) for a comprehensive review of this topic. See also Chapters 31 and 32.

## 6. Nonsampling errors in household surveys

This section considers briefly some issues regarding nonsampling errors in household surveys, a topic which requires, and has started to receive, more attention from survey statisticians. See Chapters 8–12 for more detailed treatment of nonsampling errors such as nonresponse, and measurement and processing errors. We identify some factors that make it difficult to pay greater attention to the measurement and control of nonsampling errors in household surveys, in comparison to the measurement and control of sampling errors, and point to some recent initiatives that might help to improve the situation.

Data quality issues in sample surveys have received increased attention in recent years, with a number of initiatives and publications addressing the topic, including several international conferences (see the list at the end of the chapter). Unfortunately, the discussion is still predominantly restricted to developed countries, with little participation and contribution of experiences coming from developing countries. We reach this conclusion after examining the proceedings and publications issued after these various conferences and initiatives.

After over 50 years of widespread dissemination of (sample) surveys as a key observation instrument in social science, the concept of sampling errors and their control, measurement and interpretation has reached a certain level of maturity. Treatment of nonsampling errors in household surveys is not as well developed, especially in developing and transition countries. Lack of a widely accepted unifying theory (see Lyberg et al., 1997, p. xiii; Platek and Särndal, 2001; and subsequent discussion), lack of standard methods for compiling information about and estimating parameters of the nonsampling error components, and lack of a culture that recognizes these errors as important to measure, assess, and report on imply that nonsampling errors, their measurement and assessment receive less attention in many household surveys carried out in developing or transition countries. This is not to say that these surveys are of low quality but rather to stress that little is known about their quality levels.

This has not happened by chance. The problem of nonsampling errors is a difficult one. Such errors come from many sources in a survey. Efforts to counter one type of error often result in increased errors of another kind. Prevention methods depend not only on technology and methodology but also on culture and environment, making it harder to generalize and propagate successful experiences. Compensation methods are usually complex and expensive to implement properly. Measurement and assessment are hard to perform. For example, how does one measure the degree to which a respondent misunderstands or misinterprets the questions asked in a survey (or, more precisely, the impact of such problems on survey estimates)? In addition, surveys are often carried out with very limited budgets, with publication deadlines that are becoming tighter in order to satisfy the increasing demands of information-hungry societies. In this context, it is correct for priority to be given to prevention rather than measurement and compensation, but this leaves little room for assessing how successful prevention efforts were, thereby reducing the prospects for future improvement.

Even if the situation is not good, some new developments are encouraging. The recent attention given to the subject of data quality by several leading statistical agencies, statistical and survey academic associations, and even multilateral government organizations, is a welcome development. The main initiatives that we shall refer to here are the General Data Dissemination System (GDDS) and the Special Data Dissemination Standard (SDDS) of the International Monetary Fund (IMF, 2001), which are trying to promote standardization of reporting about the quality of statistical data by means of voluntary adherence of countries to either of these two initiatives. According to the IMF, "particular attention is paid to the needs of users, which are addressed through guidelines relating to the quality and integrity of the data and access by the public to the data." These initiatives provide countries with a framework for data quality (see http://dsbb.imf.org/dqrsindex.htm) that helps to identify key problem areas and targets for quality improvement. Over 60 countries have now subscribed to the SDDS, having satisfied a set of tighter controls and criteria for the assessment of the quality of their statistical output.

A detailed discussion of the data quality standards promoted by the IMF or other organizations is beyond the scope of this chapter, but readers are encouraged to pursue the matter with the references cited here. Statistical agencies or other survey agencies in developing countries can use the available standards as starting points (if nothing similar is available locally) to promote greater quality awareness both among their members and staff, and perhaps also within their user communities.

Initiatives like these are essential to support statistical agencies in developing countries to improve their position: their statistics may be of good quality but they often do not know how good they are. International cooperation from developed towards developing countries and also among the latter is essential for progress towards better measurement and reporting about nonsampling survey errors and other aspects of survey data quality. A good example of such cooperation was the production of the volumes *Household Sample Surveys in Developing and Transition Countries* and *Designing Household Survey Samples: Practical Guidelines* by the United Nations Statistics Division (see United Nations, 2005a,b).

## 7. Integration of household surveys

The integration of household surveys can mean a variety of things, including

- Content harmonization, that is, the use of common concepts, definitions, and questions across surveys;
- Integration of fieldwork, including the ability to move cases among interviewers using different modes of collection, both for the same survey and possibly across surveys;

- Master sample, that is, the selection of a common sample that is divided among surveys, possibly using more than one phase of sampling;
- The use of common systems (collection, processing, estimation, and so on).

In the past twenty years or so, there have been efforts in several national statistical agencies to create general systems for data collection, sampling, estimation, etc. that would be sufficiently flexible that most surveys conducted by a given agency could use these systems. These efforts have met with mixed success—it appears inevitable that some surveys will make the claim that their requirements are unique.

The United Kingdom's Office for National Statistics (ONS) is currently in the process of moving its major household surveys into an Integrated Household Survey. This endeavor includes the integration of fieldwork: interviewers, and the systems they use, will be able to work on several surveys during the same time period. The integration extends to all steps in the survey process. All respondents will be asked the same core questions and different subsets of respondents will be asked additional questions from modules on a variety of topics. Details are available at the ONS web site.

At Statistics Canada, different approaches to creating a master sample for household surveys have been studied. One option, to have a distinct first-phase sample in which all respondents get a small, core set of questions, was rejected due partly to the substantial additional cost of a separate first phase. In addition, the benefits of using first-phase information to select more efficient second-phase samples only accrue to surveys that target subpopulations (e.g., travelers are important in the Canadian context), whereas most major surveys, such as the LFS and health surveys, are interested in the population as a whole. To make the first phase more useful, the core content would have to be increased to the point where it affects response burden and jeopardizes response rates. The preferred option at Statistics Canada is to make the "front end" of major surveys such as the LFS the same (corresponding to the core content of phase one of the two-phase approach) and then to pool the samples of all these surveys to create a master sample for subsampling.

In parallel with the study of design options, Statistics Canada is working on content harmonization for key variables. The objective is not only to harmonize questions on important variables such as income and education but also to create well-tested software modules that new surveys can use without needing to develop them themselves. The goal is to have different versions of certain modules, and each survey would choose a version depending on its requirements. For example, for income, there would be a short set of questions and a long set. A survey where income is of primary interest would select the long set and, conversely, most other surveys would select the short set.

Several developing countries, such as South Africa and Vietnam, have developed master samples. Pettersson (2005) discusses the issues and challenges faced by developing countries in the creation of a master sample. These include the availability of maps for PSUs, the accuracy of information, such as population counts, about such units and how to deal with regions that are difficult to access. Of course, many of these challenges are also faced by developed countries, but usually not to the same degree. The development of integrated household survey programs in developing countries has been a United Nations priority for some time. A discussion of efforts in this area, as well as further references, can be found in United Nations (2005b).

## 8. Survey redesign

Major ongoing surveys such as labor force surveys need to be redesigned periodically. Redesigns are necessary for several reasons.

– Changes in geography, such as municipal boundary changes, may result in the need for domain estimation and these changes accumulate over time. A redesign provides an opportunity to align survey strata with the latest geographical boundaries.
– The needs of users of the survey's outputs change over time, in terms of geography, frequency, and level of detail. These changing needs can be taken into account during a redesign of the survey.
– As the population changes, the sample may no longer be "in the right place" because of uneven growth and migration. A redesign is an opportunity to reallocate the sample.
– Related to the previous point, inclusion probabilities (and therefore weights) become increasingly inaccurate. This is not a concern for the bias of survey estimates but it is for their variance (efficiency).
– A redesign provides an opportunity to introduce improvements (new methods, new technology).
– For surveys with a clustered design, if all sampled clusters are carefully relisted as part of the redesign process, this will reduce undercoverage (missed dwellings) and put all clusters on the same footing (until cluster rotations start occurring).

Because of these benefits, surveys invest in periodic redesigns even though they can be very expensive. Typically, redesigns take place shortly after a population census since data from the census and census geography are key inputs for the design of household surveys.

## 9. Conclusions

In the introduction, we mentioned some major trends in household surveys since the 1940s. We conclude this chapter by taking a nonexhaustive look at current and future challenges. We have already noted that the theory of sample design is well-developed for traditional household surveys. A traditional area where there is scope for further development is the coordination of surveys. The U.S. Census Bureau recently conducted a study comparing four methods based on either systematic sampling or permanent random numbers for their household surveys (see Flanagan and Lewis, 2006). The goal of the study was to find the best method for avoiding selection of a given household in more than one survey over a certain time period. Studies of this type are needed in other contexts as well.

Most new developments are likely to stem from technological changes, particularly the internet. Currently, the internet is a useful medium for data collection, but it is not as useful as a basis for selecting representative samples of people or households. Perhaps this will change in the future: will there come a time when each individual will have a unique and persistent internet address? We have already mentioned the challenges (in developed countries) and opportunities (in developing countries) posed by the increased use of mobile phones. The future of telephone surveys depends on the development of the mobile phone industry and its impact on landline telephone usage.

Like the theory of sample design, estimation theory for sample surveys is mature, especially for relatively simple parameters such as totals, means, ratios, and regression coefficients. However, there is still a great deal to do for analytical problems, especially those associated with longitudinal surveys. Another area of active research is small area estimation, which we mentioned only briefly in this chapter.

Perhaps the biggest estimation-related challenges in household surveys are associated with nonsampling errors: how to measure them and how to fix them or take them into account. Despite their importance, space considerations prevented us from addressing this topic here, and Section 6 of this chapter barely skimmed the surface. We expect that there will continue to be a great deal of research on topics such as nonresponse and imputation, errors and biases due to reporting problems (including work on questionnaire design and cognitive research), and variance estimates that reflect more than simply sampling variability.

A common element underlying the challenges mentioned in the previous two paragraphs is the need for statistical models. Traditionally, national statistical agencies have favored purely design-based methods where possible, minimizing the use of explicit models. To deal with the problems now facing them, survey statisticians in these agencies recognize the need to use models explicitly in many areas, such as imputation and small area estimation.

Finally, we mention the influence of cost considerations on household survey methodology. In most countries, there is constant pressure to reduce survey costs. In countries with a high penetration of landline telephones, this has led to increased use of telephone interviewing, but we have noted that there is a reversal under way and that there is scope to use the internet as a response medium to counteract this reversal. We expect that efforts to improve the survey collection process using paradata and other technology-based tools such as interviewer monitoring will continue (see Groves and Heeringa, 2006). Sharing of experiences in this area among national statistical agencies (e.g., what works, what are the savings) would be beneficial.

## Acknowledgments

The authors are grateful to Ibrahim Yansaneh, Charles Lound, and Stephen Miller, whose comments helped to improve the chapter.

*Sample Surveys: Design, Methods and Applications, Vol. 29A*
ISSN: 0169-7161

DOI: 10.1016/S0169-7161(08)00017-5

# Sampling and Estimation in Business Surveys

*Michael A. Hidiroglou and Pierre Lavallée*

## 1. Introduction

A *business survey* collects data from businesses or parts thereof. These data are collected by organizations for various purposes. For instance, the System of National Accounts within National Statistical Offices of several countries uses them to compile annual (and sometimes quarterly) data on gross product, investment, capital transactions, government expenditure, and foreign trade. Business surveys produce a number of economic statistics such as: *production* (outputs, inputs, transportation, movement of goods, pollution, etc.); *sales* (wholesale and retail services, etc.); *commodities* (inputs, outputs, types of goods moved, shipments, inventories, and orders); *financial statements* (revenues, expenses, assets, liabilities, etc.); *labor* (employment, payroll, hours, benefits, employee characteristics); and *prices* (current price index, industrial price index).

Business surveys differ in a number of ways from social surveys, throughout the survey design. The frame of businesses is highly heterogeneous in terms of size and industrial classification of its units, whereas the one associated with social surveys is more homogenous. Business surveys usually sample from business registers (or equivalent list frames) that contain contact information, such as name, address, contact points, from administrative files. Social surveys, on the other hand, often use area frames to select households, and eventually individuals from within these households.

The literature on the conduct of business surveys is relatively sparse. Deming's (1960) book is the only sampling book that specifically focuses on business surveys. The two recent International Conferences on Establishment Surveys (1993 and 2000) resulted in two books specially dedicated to establishment surveys: Cox et al. (1995) and ICES-II (2001).

This chapter is structured as follows. In Section 2, we discuss sampling frames for business surveys. In Section 3, we will discuss how administrative data form an important component of business surveys. In Section 4, commonly used procedures for stratifying a business register and allocating samples will be introduced. In Section 5, methods for sample selection and rotation will be discussed, highlighting procedures that minimize response burden. The remaining Sections 6 and 7 will include brief coverage of data editing, outlier detection, imputation, and estimation, as they are covered in more depth in other chapters of this book.

## 2. Sampling frames for business surveys

### 2.1. Basic concepts

A *business* is an economic unit (establishment, farm, etc.) engaged in the production of goods or the provision of services that uses resources (labor, capital, raw materials) to produce these goods or services. Businesses operate in economic sectors that include retail trade, wholesale trade, services, manufacturing, energy, construction, transportation, agriculture, and international trade. A *business survey* is one that collects data used for statistical purposes from a sample of businesses or firms.

Businesses are characterized by a set of attributes that include *identification data*, *classification data*, *contact data*, and *activity status*. *Identification data* uniquely identify each unit with name, address, and alphanumeric identifiers. *Classification data* (size, industrial and regional classifications) are required to stratify the population and select a representative sample. *Contact data* are required to locate units in the sample, including the contact person, mailing address, telephone number, and previous survey response history. *Activity status* indicates whether a business is active (live, in-season) or inactive (dead, out-of-season). *Maintenance and linkage data* are needed to monitor and follow businesses through time. They include dates of additions and changes to the businesses and linkages between them. Collectively, the identification, classification, contact, maintenance, and linkage data items are referred to as *frame data*.

A business is also characterized by its legal structure, or its operating structure. Administrative files usually reflect how businesses are structured with respect to their *legal* arrangements, but do not reflect associated *operating structures*. The *legal* structure provides the basis for ownership, entering into contracts, employing labor, and so forth. It is via the legal structure that a business is registered with the government, and subsequently submits tax returns and/or payroll deductions and value-added taxes. The *operating* structure reflects the way the business makes and enacts decisions about its use of resources, production of goods and services, and how its accounting systems keep track of production inventories and personnel (salaries and wages, number and types of employees). These structures are reflected, and maintained, on a business register by representing their linkages with the associated business. The linkages are maintained by regular profiling of the businesses or signals triggered by survey feedback, or from updates from administrative files.

The sampling of businesses takes place by usually transforming operating structures into standardized units known as *statistical units*. The transformation takes into account decision-making autonomy, homogeneity of industrial activity, and the data available from each operating unit. Statistical units are usually represented as a hierarchy, or series of levels that allow subsequent integration of the various data items available at different levels within the organization. The number of levels within the hierarchy differs between statistical agencies. For example, the Canadian Business Register has four such levels: enterprise, company, establishment, and location (see Colledge, 1995 for definitions). In the United Kingdom, the business register has two levels: establishment and local unit (see Smith et al., 2003, for definitions). Statistical units are characterized by size (e.g., number of employees, income), geography, and industry. Statistical units are used for sampling purposes. Such units are called *sampling units*, and the level of the hierarchy that is sampled depends on the data requirements of the specific business survey.

Businesses either have a *simple* or *complex* structure. A simple business engages in a single type of activity at a single location. The vast majority of businesses have a simple structure that consists of a single legal unit that owns and controls a single operating unit. A complex business engages in a range of economic activities taking place in many locations, and can be linked to several legal units that in turn control several operating units.

The *target population* is the set of units about which data are required for a specific business survey. *Target units* within that population can be any of: legal units, operating structures, administrative units linked to businesses, or statistical units. For example, the target population could be the set of all locations that have industrial activity in the industries associated with that survey. The sampling units are at a level equal or higher than the target units.

Data collection arrangements between the statistical agency and a sampled business (defined at the statistical unit level) are established via *collection units*. Three attributes associated with a collection unit are:

- Coverage—defining the relationship between the business from which the data are being acquired and the level within the business (i.e., enterprise, location) for which the data are required;
- Collection mode—the means of obtaining the data (e.g., questionnaire, telephone interview, administrative record, etc.);
- Contact—the respondent name, address, and telephone number within the business operating structure.

Figures 1 and 2 illustrate how statistical units and collection units are related for a simple or complex business, respectively.

Collection units provide one of several means for updating the business frame in terms of frame data: others include administrative data updates and profiling. Collection units

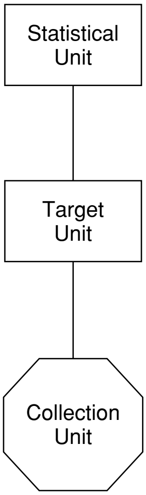


Fig. 1. Simple business.

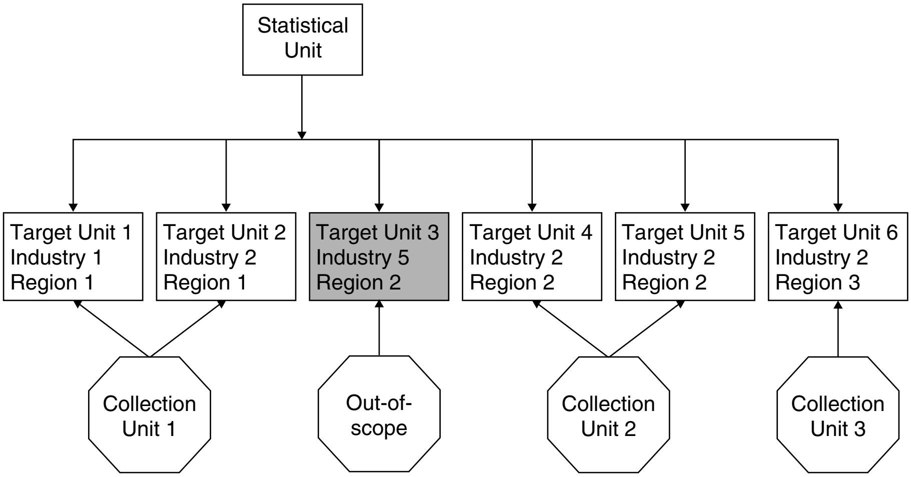

Statistical Unit
Target Unit 1 Industry 1 Region 1
Target Unit 2 Industry 2 Region 1
Target Unit 3 Industry 5 Region 2
Target Unit 4 Industry 2 Region 2
Target Unit 5 Industry 2 Region 2
Target Unit 6 Industry 2 Region 3
Collection Unit 1
Out-of-scope
Collection Unit 2
Collection Unit 3

Fig. 2. Complex business.

also represent the vehicle for monitoring respondent contacts and assessing respondent burden. Collection units are created only for statistical units in a survey sample, and are survey specific. Collection units are automatically generated with rules that depend on statistical-operating links and data availability. They can be modified manually, as need be, to take into account information related to nonstandard reporting arrangements requested by respondents.

### 2.2. *Types of sampling frames*

Sampling frames for business populations, such as retail stores, factories, or farms, are constructed so that a sample of units can be selected from them. There are two main types of sampling frames used for business surveys: *list frames* and *area frames*.

A list frame is list of businesses with their associated frame data (such as administrative identification, name, address, and contact information). This list, also known as a *business register*, should represent as closely as possible the real-life universe of businesses. Business surveys are carried out in most countries by sampling businesses from the business register. For National Statistical Agencies, administrative files are by far the preferred way to maintain the business register, as they are relatively inexpensive to acquire from government tax collecting agencies by the surveying agency. Examples of administrative files provided by tax collecting agencies to National Agencies include the Unemployment Insurance system in the Bureau of Labor Statistics in the United States and the Value Added Tax files in Britain. In Canada, a wide range of administrative files maintained by the Canada Revenue Agency is available to Statistics Canada's Business Register. These include files on corporate tax, individual tax, employee payroll deductions, goods and services tax, importers.

An area frame is a collection of geographic areas (or area segments) used to select samples using stratified multistage designs. All businesses within the selected areas are enumerated. The use of area frames for business surveys presents both advantages and disadvantages. An advantage is that it ensures the completeness of the business frame. However, their use presents a number of disadvantages. It is expensive to list and maintain a list of businesses within an area frame, as they have to be personally enumerated. The sampling design is inefficient on account of the clustering of the selected segments, and the high skewness of the data associated with businesses.

A business survey may be based on more than one of these frames, and all possible combinations have been used by National Statistical Agencies. For instance, in New Zealand, business surveys were at one time based solely on an area frame. Business surveys have always been based on a list frame in the United Kingdom, whereas Canada and the United States used at one time a combination of list and area frames. Area frames are much more costly to develop than list frames. Consequently, their use as a sampling frame for business surveys is warranted if they represent a significant portion of the estimates, or if they represent the only means to obtain a list of businesses. It is for that reason that Canada abandoned the area sample component of its retail and wholesale businesses in the late 1980s: the lack of an area frame was compensated by adjusting the weights to account for undercoverage. The United States followed suite (see Konschnik et al., 1991) in the early 1990s for their monthly retail trade surveys. The joint use of area frames with list frames results in a multiple frame. Kott and Vogel (1995) provide an excellent discussion of problems and solutions encountered in this context. From hereon, the discussion will focus on the building of business sampling frames using administrative files.

### *2.3. Maintenance*

A business universe is very dynamic. There are five main types of changes: (i) *births* due to brand new business formation, mergers or amalgamations; (ii) *deaths* resulting from either splits or physical disappearances of exiting businesses; (iii) *structural* changes in existing businesses; (iv) *classification* changes of existing businesses in terms of industry, size, and/or geography; and (v) *contact information* changes. In the case of mergers or amalgamations, the statistical units are to be linked prior to this change are inactivated and the resulting statistical units are birthed with a new identifier. Also, if a business splits, the parent statistical units are inactivated and the resulting descendents are birthed. Such changes are tracked by a combination of (a) continuously matching the administrative files to the business register; (b) profiling of existing businesses on the business register; and (c) using feedback from surveys that use the business register as their sampling frame.

The ideal system would keep all such changes up-to-date on the business register. The reality is that this is not always possible, and errors in coverage (missing, extraneous, duplicate units), classification (size, industry, geography), and contact information (name, address, telephone number) do occur. Reasons for coverage errors include improper matching of the business register to administrative files, delays in updating new births and structural changes to existing units, and delayed removal of deaths from the register. Haslinger (2004) describes the problems associated with matching administrative files to a business register in further detail. Hedlin et al. (2006) propose a methodology for predicting undercoverage due to delays in reporting new units.

Survey feedback updates the classification status, structures, and contact information of existing businesses. Although survey feedback from surveys is beneficial for updating a register, it can result in biased estimates if changes in classification stratification and/or activity status (live to dead) are used for the same sampled units in future occasions of the same survey. The problem can be avoided using updating the frame with a source independent of the sampling process. If the frame is simultaneously updated through several sources, indicators on the frame that reflect the source of the update will allow the application of independent updates in an unbiased manner. If an independent update source is not available, then domain estimation is required. This is achieved by maintaining a copy of the original status of the stratification information of the in-scope sampling units that allows computation of the survey weights as units were first selected. Domain estimation can then take place reflecting any changes in stratification and activity status of the sampled units. Although domain estimation results in unbiased estimates, their variability will eventually become too large.

We illustrate how an independent source can be used to handle dead units. Dead units in a sample are representative of the total number of dead units on the business frame. Such units are initially retained on the frame and treated as zeroes in the sample. Given that the independent administrative source identifies dead units, how do we use it? We restrict ourselves to two occasions, and for a single stratum, to illustrate how this can be handled during estimation.

On survey occasion 1, a sample $s_1$ of $n_1$ units is selected using simple random sampling without replacement (*srswor*) from a population $U_1$ of size $N_1$. On the second occasion, the universe $U_2$ consists of all the original units in $U_1$ as well as a set $U_b$ of $N_b$ universe births that have occurred between the two occasions. Suppose that a subset of $U_1$ has died, between the creation of $U_1$ and data collection for $s_1$. This subset denoted as $U_d$, consists of $N_d$ unknown dead units. Suppose that the independent administrative source identifies $A_d$ of the $N_d$ unknown dead units, where $A_d < N_d$. During data collection of $s_1$, $n_d$ deaths are also observed in sample $s_1$, and $a_d(a_d < n_d)$ of these deaths are also identified by the administrative source. The sample $s_1$ is enlarged with a representative sample $s_b$ of size $n_b = f_1 N_b$, where $f_1 = n_1/N_1$, selected using *srswor* from $U_b$. If all the deaths are retained in the sample, then the resulting sample consists of $n_2 = n_1 + n_b$ units of which $n_d$ are known to be dead.

Suppose that the parameter of interest is the population total $Y_2 = \sum_{k \in U_2} y_k$. An unbiased estimator of $Y_2$ is given by $\hat{Y}_{2,\mathrm{HT}} = \frac{N_1}{n_1} \sum_{k \in s_2} y_k$ where $s_2 = s_1 \bigcup s_b$. Note that at least $n_d$ units are dead in that sample (because some more deaths have occurred during the collection of data after the second selection). A more efficient estimator of $Y_2$ is given by the poststratified estimator $\hat{Y}_{2,\mathrm{PS}} = \frac{N_2}{\hat{N}_2} \hat{Y}_{2,\mathrm{HT}}$ where $\hat{N}_2 = \frac{N_1}{n_1}(n_1 - a_d + n_b)$ and $N_2 = N_1 - A_d + N_b$. This estimator is of a ratio form and is therefore approximately unbiased.

## 3. Administrative data

Administrative data have been increasingly used by many National Statistical Agencies for a number of years. Data are becoming more readily available in computer readable format and because that their potential to replace direct survey data reduces overall survey costs. Brackstone (1987) classified administrative data records into six types, based on their administrative purpose: the regulation of the flow of goods and people across national borders; legal requirements to register particular events; the administration of benefits or obligations; the administration of public institutions; the administration of government regulation of industry; and the provision of utilities (electricity, phone, and water services).

### *3.1. Uses*

Brackstone (1987) divided the use of administrative data into four categories: direct tabulation, indirect estimation, survey frames, and survey evaluation. These categories have been refined into seven types of use (see Lavallée, 2007a).

#### *3.1.1. Survey frames*

Administrative files have long been used by National Statistical Agencies to build and maintain their business register. The objective is to use the business register to select samples for all business surveys.

#### *3.1.2. Sample design*

Administrative data can be used as auxiliary variables for improving the efficiency of sample designs in terms of sample allocation, for example.

#### *3.1.3. Partial data substitution in place of direct collection*

Some of the variables on an administrative file can be used instead of corresponding variables collected by a direct survey. The practice of partial data substitution has been adapted by Statistics Canada for both annual and subannual surveys: annual tax data on incorporated businesses are used to replace direct collection of financial variables for annual surveys; and Goods and Services Tax (GST) data are used for monthly surveys. Erikson and Nordberg (2001) point to similar practices in Sweden's Structural Business Survey: administrative data replace direct data collection for small enterprises that have less than 50 employees.

#### *3.1.4. Edit and imputation*

Administrative data can be used to assess the validity of collected variables. For example, we should expect expenses on wages and salaries, collected via a direct survey, to be smaller than the total expenses of the business. As total expenses are also available for the corresponding unit on the administrative file, one is in a better position to decide whether a collected value is valid. In the event that collected data have been declared as incorrect, administrative data may be used to replace them, provided that the concepts and definitions are comparable between the survey and administrative data.

#### *3.1.5. Direct estimation*

As administrative data are often available on a census basis, estimates such as totals and means are obtained by summing the corresponding administrative data. Although the resulting estimates are free of sampling errors, they will be subject to all of the nonsampling errors associated with administrative data.

#### *3.1.6. Indirect estimation*

Administrative data can be used as auxiliary data to improve the precision of collected data. Calibration procedures such as those given in Deville and Särndal (1992) are used for that purpose.

### *3.1.7. Survey evaluation*

Once the survey process has been completed, administrative data can be used to evaluate the quality of the resulting process. Validation compares the survey-based estimates to corresponding administrative values to ensure that the results make sense. Such survey evaluations can be done at the microlevel (i.e., the record level), and at the macro level (i.e., the estimate level) as well. For example, in the Current Employment Statistics survey conducted by the Bureau of Labor Statistics, administrative data available on a lagged basis are used for that purpose.

## *3.2. Advantages and disadvantages*

The use of administrative data offers both advantages and disadvantages which we discuss briefly. We begin with the advantages. First, administrative data are often the only data source for essential statistics (e.g., births, customs transactions). Second, because most of the administrative data are available in computer form, considerable savings in terms of capture costs are realized. This does not, however, reduce processing costs to edit, impute, and transform them into a usable format for a specific application. Third, they can also contribute to the reduction of response burden. Fourth, as administrative data are often available on a census basis, there are no sampling errors associated with statistics obtained from them. Another consequence of their availability on a census basis is that it is possible to produce statistics for any domain of interest, including those with a very small number of units. The production of domain estimates is, however, constrained by the availability of that describes the domains on the administrative files.

We next note some of the disadvantages of using administrative data: some of them are similar to those associated with direct surveying. First, there is limited control on data timeliness, content, and quality as the administrative data originator's main objective may not be to use these data for statistical purposes. This will have a negative effect on national agencies' statistical programs. For example, a problem may occur if the frequency for compiling administrative data is changed in mid-stream (e.g., changing from monthly to quarterly). Furthermore, even though there are automated procedures for assigning industrial classification codes to administrative records, the resulting codes may be erroneous because of the limited available information describing industrial activity. Administrative data may as well not be checked as thoroughly as possible at source, and this means that the user needs to build edit checks that verify data consistency. Data in error (missing or failing edit checks) are either corrected using logical checks or imputed. Second, because administrative data have a limited number of variables, they need to be supplemented with data collected by direct surveys. Third, there are coverage problems if the population represented by the administrative data differs from the target population of the survey that uses them.

## *3.3. Calendarization*

Administrative data can cover time periods that differ markedly from the reporting periods required by surveys. For example, the ideal reporting period for an annual survey would be a calendar year, while the one for a monthly survey would be a calendar month. These time periods are also known as *reference periods*. When the reference periods of the administrative data and the survey requirements differ, the administrative data are transformed using a method known as *calendarization*. In Canada, reference periods may differ between different records of an administrative file, and may even change within the same record.

We will assume, without loss of generality, that the calendarization of the administrative data is required at a monthly calendar level. The following is a summary of Quenneville et al. (2003). Formally, the objective of calendarization is to generate monthly estimates for a variable of interest $y$ over a selected range of $T$ months, called the *estimation range*, from the set of $N$ transactions of a given unit. The available reporting periods may either partially or fully cover the months in the estimation range. If the set of transactions does not cover all the estimation range, there are *gaps* between some of the transactions and after the last transaction. The generated monthly estimates $\hat{\theta}_t$ for each month $t$ are called interpolations when they are within the span of the transactions. These interpolations provide monthly estimates for all the months associated with the transactions, as well as the gaps. The generated monthly estimates $\hat{\theta}_t$ are called extrapolations when they are outside the span of the transactions. These extrapolations provide monthly estimates for transactions not yet received. Figure 3 illustrates some of the ideas given in this paragraph.

Calendarization benchmarks a monthly indicator series $x$ to the administrative data $y$. The monthly indicator series is a series obtained from another data source that reflects the seasonal pattern of the series to be calendarized. The indicator series $x$ is in fact used for taking seasonality into account. The benchmarking procedure is based on a regression model with autocorrelated errors. It is a generalization of the method of Denton (1971), which now explicitly recognizes the timing and the duration of the data.

The benchmarking model for calendarization is represented by two linear equations. The first one, given by $y_k = \sum_{t=1}^{T} \gamma_{k,t}\theta_t + \varepsilon_k$, specifies the relationship between the reported value $y_k$ of each of the $N$ transactions and the unknown, but true, interpolations $\theta_t$. This is the key to calendarization. It states that a transaction $y_k$ corresponds to the temporal weighted sums of the true interpolations $\theta_t$ over its reporting period. The quantity $\gamma_{k,t}$, called the *coverage fraction*, is the fraction of month $t$ covered by $y_k$. For example, if $y_k$ covers from July 1 to August 31, the coverage fractions are equal to 31/31 for July, 31/31 for August, and 0 for all the other months. As another example,

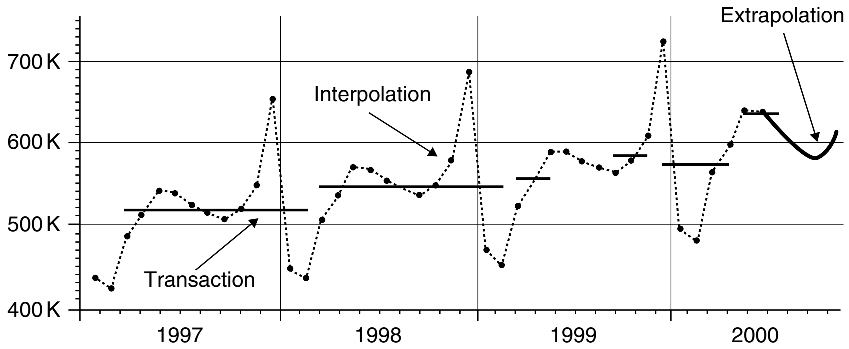


Fig. 3. Calendarization for a monthly series.

if $y_k$ covers from June 16 to August 17, the coverage fractions are equal to 15/30 for June, 31/31 for July, 17/31 for August, and 0 for all the other months. It is assumed that $E(\varepsilon_t) = 0$, $V(\varepsilon_k) = \sigma_k^2$, $\mathrm{Cov}(\varepsilon_k, \varepsilon_{k'}) = 0$ for $k \neq k'$.

The second linear equation, $x_t = \theta_t + c_t e_t$, states that the monthly indicator series $x_t$ is the sum of the true interpolation $\theta_t$ and a measurement error $c_t e_t$, $t = 1, \ldots T$. It is assumed that the indicator series $x_t$ is available for all the months $t = 1, \ldots T$ in the estimation range. It is further assumed that $E(e_t) = 0$, and $E(e_t e_{t'}) = \rho(|t - t'|)$ where $\rho(l)$ is the autocorrelation at lag $l = 0, 1, \ldots, T - 1$ of a stationary and invertible Auto-Regressive Moving Average (ARMA) process (Box and Jenkins, 1976). We also have $E(\varepsilon_k e_t) = 0$. The quantities $c_t$ are known constants proportional to a power of $|x_t|$. Note that the indicator series needs to be rescaled to the level of the data by multiplying it by the factor $\left(\sum_{k=1}^{N} y_k\right) \Big/ \left(\sum_{k=1}^{N} \sum_{t=1}^{T} \gamma_{k,t} x_t\right)$.

These equations can be written in matrix notation as: $\mathbf{y} = \boldsymbol{\gamma}\boldsymbol{\theta} + \boldsymbol{\varepsilon}$, where $E(\boldsymbol{\varepsilon}) = 0$ and $\mathrm{Cov}(\boldsymbol{\varepsilon}) = \mathbf{V}_{\boldsymbol{\varepsilon}}$; and $\mathbf{x} = \boldsymbol{\theta} + \mathbf{Ce}$, $E(\mathbf{e}) = \mathbf{0}$, $\mathrm{Cov}(\mathbf{e}) = \boldsymbol{\rho}_e$, where $\mathbf{y}$ is the column vector containing the reported values $y_k$, and so on for $\mathbf{x}$, $\boldsymbol{\varepsilon}$, and $\mathbf{e}$. We define $\boldsymbol{\gamma} = [\gamma_{k,t}]_{N \times T}$, $\mathbf{V}_\varepsilon = \mathrm{diag}(\sigma_k^2)$, $\mathbf{C} = \mathrm{diag}(c_t)$, and $\boldsymbol{\rho}_e = [\rho(|t - t'|)]_{T \times T}$.

Using a Generalized Least Squares procedure such as the one given in Dagum et al. (1998), the estimated monthly interpolations $\hat{\theta}_t$ are obtained from $\hat{\boldsymbol{\theta}} = \mathbf{x} + \mathbf{C}\boldsymbol{\rho}_e\mathbf{C}\boldsymbol{\gamma}'(\boldsymbol{\gamma}\mathbf{C}\boldsymbol{\rho}_e\mathbf{C}\boldsymbol{\gamma}' + \mathbf{V}_\varepsilon)^{-1}(\mathbf{y} - \boldsymbol{\gamma}\mathbf{x})$. The estimated interpolations can be shown to exactly satisfy the benchmarking constraint by setting $\mathbf{V}_\varepsilon = 0$ and premultiplying both sides of the previous equation in $\hat{\boldsymbol{\theta}}$ by the matrix $\boldsymbol{\gamma}$. This leads to $\boldsymbol{\gamma}\hat{\boldsymbol{\theta}} = \mathbf{y}$, which shows that the estimated interpolations exactly satisfy the benchmarking constraint $\mathbf{y} = \boldsymbol{\gamma}\boldsymbol{\theta}$, because we set $\mathbf{V}_\varepsilon = 0$.

In Canada, calendarization of the Goods and Service Tax data provided by the Canada Revenue Agency has contributed significantly to reducing survey costs and response burden of conducting monthly business surveys in a number of industrial sectors that include wholesale, retail, manufacturing, and services.

## 4. Sample size determination and allocation

### 4.1. Choice of sampling unit

The sampling of a business universe is usually done in two steps. First, the in-scope target universe is defined, and a set of target units are obtained. Second, the sampling unit is defined at some level of the statistical units. The sampling level will be at least at the level of the target units. For example, suppose that locations and establishments are the only two types of statistical units on the business register. Given that the target unit is the location, the sampling unit could either be the location or the establishment.

### 4.2. Stratification of sampling units

Once the population of businesses has been partitioned into sampling units, the sampling units are stratified. The selection of samples is done independently in each of these strata. The strata are usually based on geography (e.g., Canadian provinces and major metropolitan centers), standard industrial classification (e.g., restaurants, agents and brokers, garages, department stores), and some measure of size (e.g., number of employees, gross business income, net sales). Cochran (1977) gives four main reasons for stratification. First, it reduces the variances of survey estimates, if they are correlated with the stratification variables. Second, stratification may be dictated by administrative convenience if, for example, the statistical agency has field offices. Third, sampling problems may differ markedly in different parts of the population such that each part should be considered independently. Finally, reliable estimates for designated subpopulations that have high overlap with the design strata can be obtained as a by-product.

The selection of samples in business surveys frequently uses simple random sampling techniques applied to each stratum. A feature of most business populations is the skewed nature of the distribution of characteristics such as sales, employment output, or revenue. A "certainty" or "take-all" stratum of the very largest sampling units is usually created to reduce the variances of estimates: all sampling units within the certainty strata are selected in the sample. Noncertainty strata are then formed and the remaining sampling units are placed in them according to their size.

The optimality of stratification breaks down over time, resulting in a less efficient sample design. Deterioration of stratification of the frame requires that the whole frame be restratified. A new sample that is optimal with respect to the newer stratification is then selected, in general with as much overlap as possible with the previous sample. This overlap ensures continuity of the estimates in a periodic survey, and is less expensive than a complete redraw from a collection perspective.

Factors that affect realized precision include: population size; overall sample size; stratification of the frame in terms of the number of strata and the stratum allocation scheme; the construction of stratum boundaries for continuous stratification variables; the variability of characteristics in the population; the expected nonresponse; cost, time, and operational constraints; and the targeted precision of summary statistics such as means and totals of the target variables.

### *4.3. Allocation*

#### *4.3.1. Notation*

We introduce notation to deal with allocation for a single $x$-variable (univariate allocation). The finite population $U$ of $N$ units is divided into $L$ nonoverlapping subpopulations or strata $U_h$, $h = 1, 2, \ldots, L$, with $N_h$ units each, and $N = \sum_{h=1}^{L} N_h$. A sample $s_h$ of size $n_h$, $h = 1, \ldots, L$, is selected independently by simple random sampling without replacement (*srswor*) within each $h$-th stratum, yielding an overall sample of size $n = \sum_{h=1}^{L} n_h$. Let $x_{hk}$ (known from a previous census or survey) denote the $k$-th observation within stratum $h$. An unbiased estimator of the population total $X = \sum_{h=1}^{L} X_h = \sum_{h=1}^{L} \sum_{k \in U_h} x_{hk}$ is given by $\hat{X} = \sum_{h=1}^{L} N_h \overline{x}_h$, where $\overline{x}_h = \sum_{k \in s_h} x_{hk} / n_h$, and its associated population variance is

$$\begin{aligned} V(\hat{X}) &= \sum_{h=1}^{L} N_h^2 \left( \frac{1}{n_h} - \frac{1}{N_h} \right) S_h^2 \\ &= \left( \sum_{h=1}^{L} A_h / n_h \right) - D \qquad (1) \end{aligned}$$

where $A_h = N_h^2 S_h^2$, $D = \sum_{h=1}^{L} N_h S_h^2$, $\overline{X}_h = X_h / N_h$, and $S_h^2 = \sum_{k \in U_h} \left(x_{hk} - \overline{X}_h\right)^2 / (N_h - 1)$.

The $x$-values for the current population will not be known. However, an estimate $V'(\hat{X})$ of the population variance $V(\hat{X})$, based on a sample $s_h'$ of size $n_h'$ from a pilot survey, a past survey, or from administrative data, can be used as a substitute. Here, $V'(\hat{X}) = \left(\sum_{h=1}^{L} A_h'/n_h'\right) - D'$ with $A_h' = N_h^2 \hat{S}_h'^2$, $D' = \sum_{h=1}^{L} N_h \hat{S}_h'^2$, $\hat{S}_h'^2 = \sum_{k \in s_h'} (x_{hk} - \overline{x}_h')^2/(n_h' - 1)$, and $\overline{x}_h' = \sum_{k \in s_h'} x_{hk}/n_h'$.

### *4.3.2. Some allocation schemes*

Let $a_h$ denote the proportion allocated to the $h$-th stratum, where $0 \le a_h \le 1$ and $\sum_{h=1}^{L} a_h = 1$. The number of units of units allocated to a given stratum $h$ is given by $n_h = n a_h$ for $h = 1, 2, \ldots, L$.

Assume that the cost of collecting data is the same for all units. The allocation of the sample $s$ of size $n$ to strata $s_h$ can be carried out in two ways:

i Require that the variance of $\hat{X}$ should be minimal given that the overall sample size $n$ is fixed or the overall cost is fixed;

ii Specify a tolerance on the precision of the estimate $\hat{X}$ as a predetermined coefficient of variation $c$, that is $c^2 = V(\hat{X})/X^2$. In that case, the objective is to minimize the sample size $n$ (or total cost), and is computed using the chosen allocation scheme. Substituting $n_h = n\, a_h$ and $V(\hat{X}) = c^2 X^2$ into (Eq. 1) and solving for $n$, we obtain:

$$n = \left(c^2 X^2 + D\right)^{-1} \left(\sum_{h=1}^{L} A_h / a_h\right) \tag{2}$$

A number of allocation schemes for stratified *srswor* are summarized using the above notation.

*4.3.2.1. N-proportional allocation* ($a_h = N_h/N$). This scheme is generally superior to simple random sampling of the whole population if the strata averages $\overline{X}_h$ differ considerably from each other. A slight reduction in variance results only if the strata means are similar. It is often used in business surveys to equalize the sampling weights between strata whose units are known to have a high probability to change classification.

*4.3.2.2. X-Proportional allocation* ($a_h = X_h/X$). $X$-proportional allocation is used in business surveys because distribution of data is quite skewed. The largest units are sampled with near certainty and the remaining units are sampled with probability less than one.

*4.3.2.3. Optimal allocation* $\left(a_h = \left(\sum_{h=1}^{L} (N_h S_h)\right)^{-1} (N_h S_h)\right)$. More sample units are allocated to the larger strata and/or strata that have the highest variances. This type of allocation is also known as Neyman allocation (see Neyman, 1934). Optimal allocation is similar to $X$-proportional allocation if $S_h/\overline{X}_h$ is assumed constant across strata. The difficulty with this allocation is that the population variance $S_h^2$, or its estimate $\hat{S}_h'^2$, may be unstable.

*4.3.2.4. $\sqrt{N}$ or $\sqrt{X}$-proportional allocation* $\left(a_h = \left(\sum_{h=1}^{L} \sqrt{N_h}\right)^{-1} \sqrt{N_h}\right.$ *or* $\left.\left(\sum_{h=1}^{L} \sqrt{X_h}\right)^{-1} \sqrt{X_h}\right)$. This scheme results in good reliability of strata estimates $\hat{X}_h$, but it is not as efficient as Neyman allocation for the overall estimate $\hat{X}$. This type of allocation was first proposed by Carroll (1970), and provides fairly similar coefficients of variation for stratum totals $\hat{X}_h$. Bankier (1988) extended the concept by considering $a_h$ as $\left(\sum_{h=1}^{L} (X_h)^q\right)^{-1} (X_h)^q$ where $0 \le q \le 1$. Note that setting $q$ to 0.5 results in the Carroll allocation.

## 4.4. Some special considerations

Nonresponse, out-of-datedness of the frame, initial over-allocation of units to strata, and minimum sample size within strata are additional factors to account for in the computation of the sample size.

### 4.4.1. Nonresponse

Nonresponse reduces the effective sample size, and hence the reliability of summary statistics. Assume the sample size is $n = \sum_{h=1}^{L} n_h$ units and the nonresponse rates (known from experience) are expected to be $r_h (h = 1, 2, \ldots, L)$, where $0 \le r_h < 1$ within each stratum. The resulting effective sample size would be $n_{\text{eff}} = \sum_{h=1}^{L} n_h(1 - r_h) < n$ after data collection. The sample size can be increased to $n_h' = n_h/(1 - r_h)$ within each stratum $h$ to compensate for the nonresponse. This increase assumes that the nonrespondents and respondents have similar characteristics. If they differ, a representative sample of the nonrespondents needs to be selected to represent the nonresponding part of the sample.

### 4.4.2. Out-of-date frame

The impact of an out-of-date frame should be reflected in the sample size determination and allocation method. The out-of-datedness of a frame occurs because the classification (geography, industry, status: live or dead) of the units is not up to date. In our case, we just focus on estimating the total of the live units for a variable $y$, given that a number of dead units are present on the frame but identified as active. Consequently, a representative portion of them will be included in the sample. The universe of "active" units is labeled as $U$. The corresponding universe of live units (but unknown) is denoted as $U_\ell$, where $U_\ell \subset U$. Let the parameter of interest be the domain total $Y_\ell = \sum_{k \in U} y_{k\ell}$, where $y_{k\ell}$ is equal to $y_k$ if $k \in U_\ell$ and zero otherwise. We need to determine the sample size $n$, such that: (i) the allocation to the design strata is $n_h = n\, a_h (0 < a_h < 1)$ and (ii) the targeted coefficient of variation $c$ is satisfied, that is, $V(\hat{Y}_\ell) = c^2 Y_\ell^2$. Simple random samples $s_h$ of size $n_h$ are selected from $U_h (h = 1, \ldots, L)$, without replacement. The corresponding estimator is $\hat{Y}_\ell = \sum_{h=1}^{L} \hat{Y}_{h\ell}$ with $\hat{Y}_{h\ell} = \sum_{k \in s_h} (N_h/n_h)\, y_{k\ell}$. We obtain the required sample size $n$ using $V\left(\hat{Y}_\ell\right) = c^2 Y_\ell^2 = \sum_{h=1}^{L} N_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h}\right) S_{h\ell}^2$

where $S_{h\ell}^2 = (N_h - 1)^{-1} \sum_{k \in U_h} \left(y_{k\ell} - \overline{Y}_{h\ell}\right)^2$ with $\overline{Y}_{h\ell} = \sum_{k \in U_h} y_{k\ell}/N_h$. The required sample size is

$$n = \frac{\sum\limits_{h=1}^{L} N_h^2 \left( \tilde{S}_{h\ell}^2 + (1 - P_{h\ell}) \tilde{\overline{Y}}_{h\ell}^2 \right) P_{h\ell}/a_h}{c^2 \left( \sum\limits_{h=1}^{L} N_h \overline{Y}_{h\ell} P_{h\ell} \right)^2 + \sum\limits_{h=1}^{L} N_h \left( \tilde{S}_{h\ell}^2 + (1 - P_{h\ell}) \tilde{\overline{Y}}_{h\ell}^2 \right) P_{h\ell}},$$

where $\tilde{S}_{h\ell}^2 = (N_{h\ell} - 1)^{-1} \sum_{k \in U_{h\ell}} \left( y_k - \tilde{\overline{Y}}_{h\ell} \right)^2$, $\tilde{\overline{Y}}_{h\ell} = \sum_{k \in U_h} y_{k\ell}/N_{h\ell}$, $N_{h\ell}$ is the number of units belonging to domain $U_{h\ell} = U_\ell \cap U_h$, and $P_{h\ell} = N_{h\ell}/N_h$ is the expected proportion of units that belong to $U_\ell$ and initially sampled in stratum $h$. Note that the case of $P_{h\ell} = 1$ yields the usual sample size formula. The mean and variance components can be estimated from previous surveys. The required sample sizes at the stratum level are then simply $n_h = n\, a_h$ for $h = 1, \ldots, L$. It is not recommended to use an approximation of the type $n_h^* = n_h/P_h$ to compensate for unknown dead units, where $n_h$ is computed ignoring the existence of unknown dead units in the universe.

### 4.4.3. *Over-allocation*

Optimum allocation (Neyman), $X$-proportional or $\sqrt{X}$-allocation may result in sample sizes $n_h$ that are larger than the corresponding population sizes $N_h$ for some strata. The resulting overall sample size will be smaller than the required sample size $n$. Denote the set of strata where over-allocation has taken place as "OVER." Such strata are sampled with certainty, that is, $n_h = N_h$, with total sample size $n_{\text{OVER}} = \sum_{h \in \text{OVER}} n_h$. The remaining set of strata, denoted as "NORM," is allocated the difference $n - n_{\text{OVER}}$ using the chosen allocation scheme. That is, for $h \in \text{NORM}$, $n'_h = (n - n_{\text{OVER}})\, a'_h$ where $a'_h$ is computed according to the given allocation scheme, with $\sum_{h \in \text{NORM}} a'_h = 1$. The process is repeated until there is no over-allocation. A similar procedure is used in the case where the overall sample size is chosen to satisfy reliability criteria. The only difference is that $n_h = a'_h \dfrac{\sum_{h \in \text{NORM}} A_h/a'_h}{c^2 X^2 + D'}$.

### 4.4.4. *Minimal sample size*

A minimal sample size within each stratum is a requirement to protect against empty strata occurring on account of nonresponse. It also provides some protection against allocations that are poor for characteristics not considered in the sample design. A minimal sample size of three to five units is quite often used in large-scale surveys: at least two units are required to estimate variances unbiasedly. Denote as $m_h (h = 1, 2, \ldots, L)$ the minimal sample size within the $h$-th stratum: $m_h$ will most likely be the same for all strata. The minimum sample size may be applied before or after the allocation of given sample size $n$ has been established. If it is applied before, a sample size $m'$ is initially set aside for minimum size requirements across all strata, where $m' = \sum_{h=1}^{L} m'_h$ and $m'_h = \min\{N_h, m_h\}$. The remaining sample size $n - m'$ is allocated to the population strata of size $N_h - m'_h$ using the chosen allocation method. If the minimum size is applied after allocation, the sample size for the $h$-th stratum is $n'_h = \min\{\max[n_h, m_h], N_h\}$. The sum of the overall sample $\sum_{h=1}^{L} n'_h$ may be greater than $n$.

### *4.4.5. Equalization of the coefficient of variation among strata*

A property of power allocation is that the coefficient of variation for the estimates of the totals will be fairly similar for each stratum. However, it may be required to have exactly equal levels of reliability for estimates of the totals at the stratum level: that is, $V(\hat{X}_h)/X_h^2 = c_1$ for all $h = 1, \ldots, L$, where $c_1$ is not known and is bounded between 0 and 1. If the overall coefficient of variation has been fixed at $c$, it follows that $c_1 = c\sqrt{X/\sum_{h=1}^{L} X_h^2}$, and the sample size within each stratum is $n_h = A_h/\left(c_1^2 X_h^2 + D_h\right)$. If the overall sample size has been fixed at $n$, we solve iteratively $f(c_1) = n - \sum_{h=1}^{L} \dfrac{A_h}{c_1^2 X_h^2 + D_h}$, an increasing function in $c_1$, using the Newton–Raphson procedure.

### *4.4.6. Simultaneous level of reliability for two stratification variables*

Assume that the population has been stratified at the *geography* $(h = 1, \ldots, L)$ and *industry* $(\lambda = 1, \ldots, M)$ levels. Specified coefficients of variation of totals are required at the subnational and industry levels: let these be $c_{h.}$ and $c_{.\lambda}$, respectively. The sampling takes place within a further size stratification of these LM possible cross-classifications. The required sample size for each of these levels can be computed if we can obtain the corresponding coefficient of variation given the marginal (i.e., *geography* and *industrial* reliability constraints). This coefficient can be obtained using a raking procedure (see Deming and Stephan, 1940). Let $X_{h\lambda}$ be the population total of a given variable of interest (say $x$), and let $X_{h.}$ and $X_{.\lambda}$ be the associated marginal totals. The $h\lambda$-th coefficient of variation at the $r$-th iteration is given by

$$c_{h\lambda}^{(r)} = c_{h\lambda}^{(r-1)} \frac{(c_{h.} X_{h.})\,(c_{.\lambda} X_{.\lambda})}{\sqrt{\sum\limits_{h=1}^{L} c_{h\lambda}^{(r-1)} X_{h\lambda}^2}\sqrt{\sum\limits_{\lambda=1}^{M} c_{h\lambda}^{(r-1)} X_{h\lambda}^2}}$$

The starting point for this algorithm is $c_{h\lambda}^{(0)} = (\dot{c}_{h.} + \dot{c}_{.\lambda})/2$, where $\dot{c}_{h.}$ and $\dot{c}_{.\lambda}$ are marginal coefficients of variation given by $\dot{c}_{h.} = c_{h.} X_{h.}\Big/\sqrt{\sum_{\lambda=1}^{M} X_{h\lambda}^2}$ for $h = 1, 2, \ldots, L$ and $\dot{c}_{.\lambda} = c_k X_{.\lambda}\Big/\sqrt{\sum_{h=1}^{L} X_{h\lambda}^2}$ for $\lambda = 1, 2, \ldots, M$. In practice, five iterations are sufficient to stabilize the $c_{h\lambda}^{(r)}$ values. The sample size required to achieve the required marginal coefficients of variation for each $h\lambda$-th cell is then $n_{h\lambda} = \left(\left(c_{h\lambda}^{(R)}\right)^2 X_{h\lambda}^2 + D_{h\lambda}\right)^{-1} A_{h\lambda}$, where $A_{h\lambda} = N_{h\lambda}^2 S_{h\lambda}^2$, $D_{h\lambda} = N_{h\lambda}^2 S_{h\lambda}^2$, and $c_{h\lambda}^{(R)}$ is the coefficient of variation at the final iteration $R$.

## *4.5. Construction of self-representing strata*

### *4.5.1. Using known auxiliary data x*

Stratification of a population into natural strata based on geography and industrial activity usually increases the efficiency of a sample design. Further stratification by size of business (employment, sales) always increases the efficiency of the sample design in business surveys because business populations are typically composed of a few large units (accounting for a good portion of the total for the variable of interest) and many small units. It is therefore desirable to construct stratum boundaries that split the businesses into a take-all stratum containing the largest units (being sampled with certainty) and a number of take-some strata containing the remaining units (sampled with a given probability). The resulting stratification offers two advantages. First, the overall sample size required to satisfy reliability criteria (denoted as $c$) is dramatically reduced (or alternatively, the variance of the estimated total is minimized for a fixed overall sample size). Second, because the largest units are sampled with certainty, the chance of observing large values for the units selected in the take-some strata is reduced.

Consider a population $U$ of size $N$ where the units have the size measures, $x_1, x_2, \ldots, x_N$. Define order statistics $x_{(1)}, x_{(2)}, \ldots, x_{(N)}$ where $x_{(k)} \leq x_{(k+1)}, k = 1, \ldots, N-1$.

We first provide two approximations due to Glasser (1962) for fixed sample size $n$ and Hidiroglou (1986) for fixed coefficient of variation $c$ for splitting the universe into a take-all and a take-some stratum. Glasser's (1962) rule for determining an optimum cut-off point is to declare all units whose $x$ value exceeds $\overline{X}_N + \sqrt{NS_N^2/n}$ as belonging to the take-all stratum, where $\overline{X}_N = \sum_{k \in U} x_k/N$ and $S_N^2 = \sum_{k \in U}(x_k - \overline{X}_N)^2/(N-1)$. Hidiroglou's (1986) algorithm is iterative. The take-all boundary $B_r (r = 1, 2, \ldots)$ at the $r$-th iteration is given by

$$B_r = \overline{X}_{N-T_{r-1}} + \left\{ \frac{(n - T_{r-1} - 1)}{(N - T_{r-1})^2} c^2 X^2 + S^2_{N-T_{r-1}} \right\}^{1/2}$$

where $T_{r-1}$ is the number of take-all units at the $(r-1)$-th iteration, and $\overline{X}_{N-T_{r-1}}$ and $S^2_{N-T_{r-1}}$ are the corresponding take-some stratum population mean and variance. The process is started by setting $T_0$ to zero, and the iterative process continues until $0 < (1 - T_r/T_{r-1}) < 0.10$ has been met. Convergence usually occurs after two to five iterations.

Lavallée and Hidiroglou (1988) provided a procedure for stratifying skewed populations into a take-all stratum and a number of take-some strata, such that the sample size is minimized for a given level of precision. They assumed power allocation of the sample for the take-some strata, as this type of allocation tends to equalize coefficients between the strata. Their algorithm uses Dalenius's (1950) representation of a finite population in terms of a continuous population. That is, given a continuous density function $g$ of the auxiliary variable $x$ in the range $(-\infty, \infty)$, the conditional mean and variance of the $h$-th stratum $U_h$ can be expressed as $\mu_h = \int_{b_{(h-1)}}^{b_{(h)}} yg(y)/W_h$ and $\sigma_h^2 = \int_{b_{(h-1)}}^{b_{(h)}} y^2 g(y)/W_h - \mu_h^2$ where $W_h = \int_{b_{(h-1)}}^{b_{(h)}} g(y)dy$. The overall sample size is given by

$$n = NW_L + \frac{N \sum\limits_{h=1}^{L-1} W_h^2 \sigma_h^2 / a_h}{N\left(c \sum\limits_{h=1}^{L} W_h \mu_h\right)^2 \mu^2 + \sum\limits_{h=1}^{L-1} W_h \sigma_h^2} \tag{3}$$

where $a_h = \dfrac{(W_h \mu_h)^p}{\sum_{h=1}^{L-1}(W_h \mu_h)^p}$ for $h = 1, \ldots, L-1$. Hidiroglou and Srinath (1993) proposed a more general form of $a_h$, given by $a_h = \gamma_h / \sum_{h=1}^{L-1} \gamma_h$ where $\gamma_h = W_h^{2q_1} \mu_h^{2q_2} \sigma_h^{2q_3}$,

$q_i \geq 0$ ($i = 1, 2, 3$). A number of different allocations are obtained with various choices of the $q_i$'s. For example, Neyman allocation is obtained by setting $q_1 = q_3 = 0.5$ and $q_2 = 0$.

The optimum boundaries $b_1, b_2, \ldots, b_{L-1}$, where $x_{(1)} \leq b_1 < \ldots < b_{L-1} \leq x_{(N)}$, are obtained by taking the partial derivatives of (3) with respect to each $b_h$, $h = 1, \ldots L-1$, equating them to zero, and solving the resulting quadratic equations iteratively using a procedure suggested by Sethi (1963). The initial values are set by choosing the boundaries with an equal number of elements in each group. Although the Lavallée–Hidiroglou method is optimal, Slanta and Krenzke (1996) and Rivest (2002) noted that it does not always converge, and that convergence depends on providing the algorithm with reasonable initial boundary values.

Gunning and Horgan (2004) recently used the geometric progression approach to stratify skewed populations. They based their algorithm on the following observation stated in Cochran (1977): when the optimum boundaries of Dalenius (1950) are achieved, the coefficients of variation ($CV_h = S_h/\overline{X}_h$) are often found to be approximately the same in all strata. Assuming that the $x$ variable is approximately uniformly distributed within each stratum, their boundaries are $b'_h = a\tau^h$ for $h = 1, 2, \ldots, L-1$ where $a = x_{(1)}$ and $\tau = \left(x_{(n)}/x_{(1)}\right)^{1/L}$. The advantages of Gunning–Horgan's procedure are that it is simple to implement, and that it does not suffer from convergence problems. However, two weaknesses of the procedure are that it does neither stratify a population according to an arbitrary sample allocation rule (represented by $a_h$), nor does it require the existence of a take-all stratum. The stratification boundaries obtained by the Gunning–Horgan procedure could be used as starting points for the Lavallée–Hidiroglou algorithm to ensure better convergence.

### 4.5.2. *Using models to link auxiliary data x and survey variable y*

A number of authors have developed models between the known auxiliary data $x$ and the survey variable $y$. They include Singh (1971), Sweet and Sigman (1995), and Rivest (2002). The last three authors incorporated the impact of the model in the Lavallée–Hidiroglou algorithm. As Sweet and Sigman (1995) and Rivest (2002) demonstrated, the incorporation of the model could lead to significant improvements in the efficiency of the design.

## 5. Sample selection and rotation

As mentioned earlier, strata are often cross-classifications of industry and geography by size. These strata are either completely enumerated (take-all) or sampled (take-some). We denote the required sampling fraction within a take-some stratum as $f$ (the subscript $h$ is dropped in this section to ease the notation). It is equal to unity for the take-all strata.

The sampling mechanism of in-scope units in business surveys needs to account for a number of factors. First, the units should be selected using a well-defined probability mechanism that yields workable selection probabilities for both estimation ($\pi_k$'s, $k = 1, \ldots, N$) and variance estimation ($\pi_{k,k'}$'s). Note that $\pi_k \approx f$ within the strata. Second, the resulting samples should reflect the changing nature of the universe in terms of births, deaths, splits, mergers, amalgamations, and classification changes. Third, the selection should allow for sample rotation of the units to alleviate response burden across time. Fourth, there should be some control of the overlap of the sampled units between various business surveys occurring concurrently. Fifth, if there are significant changes in the stratification of the universe, it should be possible to redraw a sample that reflects the updated stratification and sampling fractions.

Response burden occurs *within surveys* and *across surveys*. Response burden within surveys is minimized if a selected business remains in sample for as few occasions as possible. Response burden across surveys is minimized if a business is selected in as few surveys as possible at the same time. However, these preferences will not normally agree with what is best for a survey in terms of reliability within and between occasions.

Two types of coordination can be distinguished for selecting several samples from the same frame: they are *negative* and *positive coordination*. *Negative coordination* implies that response burden is reduced, by ensuring that a business is not selected in too many surveys within a short time frame. *Positive coordination* implies that the overlap is maximized as much as possible between samples.

### 5.1. *Selection procedures*

*Poisson sampling* and its variants form the basis for sampling business surveys in most national agencies. This method allows for response burden control within and across surveys. Poisson sampling as defined by Hájek (1964) assigns each unit in the population of size $N$ a probability of inclusion in the sample denoted as $\pi_k = np_k$, $k = 1, \ldots, N$. Here, $n$ is the required sample size and $p_k$ is usually linked to some measure of size of the unit $k$. Ohlsson (1995) provides the following procedure for selecting a Poisson sample of expected sample size $n$. A set of $N$ independent uniform random numbers $u_k$ is generated, where $0 \leq u_k \leq 1$. If these random numbers are fixed and not regenerated for the same units between two survey occasions, they are called permanent random numbers (PRN). A starting point $\alpha$ is chosen in the interval [0, 1]. A population unit $k$ is included in the sample if $\alpha < u_k \leq \alpha + np_k$, provided $\alpha + np_k \leq 1$. If $\alpha + np_k > 1$, it is included in the sample if $(\alpha < u_k \leq 1) \cup (0 < u_k \leq \alpha + np_k - 1)$. The value of $\alpha$ is usually set to zero when a survey sample is first selected. Sampling from a stratified universe occurs by assigning the required $p_k$'s and sample sizes within each stratum. Births to a business universe are easily accommodated with Poisson sampling: a PRN is generated for the birth unit, and it is selected using the previously stated algorithm. Rotation of the sample takes place by incrementing $\alpha$ by a constant $\kappa$ on each survey occasion. The constant $\kappa$ reflects the required rotation rate.

A special case of Poisson sampling is *Bernoulli* sampling: the $p_k$'s are equal to $1/N$ within each stratum. It should be noted that, conditioning on the realized sample size, Bernoulli sampling is equivalent to *srswor*. Poisson and Bernoulli sampling are often not used in practice, because the realized sample sizes may vary too much around the expected sample sizes. A number of procedures have been developed over the years to control this weakness. These include collocated sampling (Brewer et al., 1972), sequential Poisson sampling (Ohlsson, 1995), and Pareto sampling (Rosén, 2001). Statistics Canada uses Bernoulli sampling to sample tax records from Canada Revenue Agency's administrative tax files. PRNs are created by transforming the unique identifying numbers on the administrative files to pseudorandom numbers using a hashing algorithm. This approach, introduced by Sunter (1986), maximizes sample overlap between sampling occasions.

Bernoulli variants have been used in a number of agencies. Statistics Sweden samples from their business register using sequential *srswor*. Sequential *srswor*, described in Ohlsson (1990a,b), involves the selection of the first $n_h$ units within the ordered list of the PRNs within each stratum of size $N_h$. The synchronized sampling methodology used by the Australian Bureau of Statistics, developed by Hinde and Young (1984), is quite similar to the one developed at Statistics Sweden. It differs from the Swedish one with respect to its definition of the start and end points of the sampling intervals. The start and end points are equal to the PRNs associated with the units at the time of selection. In-scope population units are selected if they belong to these sampling intervals. The start point is in sample but the end point is not. A desired sample size $n$ is achieved by including the start point, and the remaining $n - 1$ successive PRNs. The incorporation of births and deaths is done by moving the start or end points to the right to prevent units reentering the sample. The procedure allows for rotation, as well as periodic restratification of the frame. Negative or positive coordination is achieved by allowing different surveys to use well defined intervals on the [0, 1) interval. More details of this methodology are available in McKenzie and Gross (2001). Sampling of the business register at Statistics Canada is a blend of collocated sampling described in Brewer et al. (1972), and the panel sampling procedure given by Hidiroglou et al. (1991). Details of the procedure are given in Srinath and Carpenter (1995).

A variant of Poisson sampling, known as Odds Ratio Sequential Poisson sampling, has been used to sample businesses in the petroleum industry. Saavedra and Weir (2003) provide more details of the methodology, which is really Pareto sampling as described in Rosén (2001). This method provides fixed sample sizes in the Poisson context, and the resulting probabilities of selection closely approximate the desired probabilities to be proportional to size.

### 5.2. *Accounting for response burden*

The methodology used by the "Central Bureau voor de Statistiek" (CBS) in the Netherlands incorporates PRNs to control sample rotation across and within their business surveys, while accounting for response burden. De Ree (1999) briefly described this methodology: at the time of initial sample selection, sampling units on the business register are assigned a PRN, and ranked accordingly. A PRN remains associated with a given unit on the register throughout its life. However, the manner in which samples are selected may vary. Businesses can be selected several times successively for a specific survey (a subannual survey), or by several different surveys. Each time a business is surveyed, its associated response-burden coefficient is increased. After each selection, the ordering changes so that businesses with a lower cumulated response burden are placed before businesses with a higher cumulated response burden. Earlier versions of this methodology are presented in more detail in Van Huis et al. (1994). The CBS sampling system has useful features: (i) it integrates sampling amongst several surveys, and takes into account the response burden; (ii) it allows user specified parameters for defining the sampling and rotation rates. However, there is some limitation in the choice of stratification.

Rivière (2001) discusses a somewhat different approach, whereby the sample selection procedure does not change, but the initially assigned random numbers are systematically permuted between units for different coordination purposes: smoothing out the burden, minimizing the overlap between two surveys, or updating panels. Permutations of the random numbers are carried out within intersections of strata that are referred to as microstrata. The microstrata method was developed in 1998 in the framework of Eurostat's SUPCOM project on sample coordination. The methodology was initially implemented in a program known as SALOMON in 1999. Improvements to SALOMON resulted in a program known as MICROSTRAT in 2001.

In the microstrata method, the initial procedure is to assign a random number to every unit on the business register that is in-scope for sampling. As with the CBS methodology, every unit is also assigned a response-burden coefficient that cumulates every time the unit is selected for a given survey. The random numbers never change but they can be permuted between the units. The permutation of the random numbers is controlled by the cumulated burden similarly to the CBS procedure. The most important difference from the CBS methodology is that the permutations are done within the microstrata. A microstratum is the largest partition that can be defined so as to sort the units by increasing response burden without introducing bias. Using this technique, the random numbers remain independent and identically distributed with a uniform distribution.

The main drawback to microstratification is the possible creation of microstrata so small that the sample coordination becomes ineffective. However, this can be avoided using a different sorting procedure. On the other hand, microstratification has several benefits. The method has good mathematical properties and gives a general approach for sample coordination in which births, deaths, and strata changes are automatically handled. There is no particular constraint on stratification and rotation rates of panels. It is unbiased, as shown by Rubin–Bleuer (2002).

## 6. Data editing and imputation

### 6.1. Data editing

Business survey data are not free of errors and this holds true whether they have been collected by direct surveys or obtained through administrative sources. Errors are detected and corrected by editing data both at the *data capture* and *estimation* stages. A number of data editing procedures are used to detect errors in the data. The associated edits are based on a combination of subject-matter experts' knowledge, as well as data analysis. Edits are either applied to each individual observation or across a number of them. The former is known as *microediting*, whereas the latter as *macroediting*.

Microediting takes place both during data capture and estimation. Microediting can be manual (e.g., a human declaring data in error) or automated (e.g., a computer rejecting data using predetermined editing rules). Edits associated with microediting include validation edits, logical edits, consistency edits, range edits, and variance edits. Validity edits verify the syntax within a questionnaire. For example, characters in numeric fields are checked, or the number of observed digits is ensured to be smaller or equal to the maximum number of positions allowed within the questionnaire. Range edits identify whether a data item value falls within a determined acceptable range. Consistency edits ensure that two or more data items (mainly financial variables) within a record do not have contradictory values. They follow rules of subject-matter experts to verify that relationships between fields are respected. Variance edits isolate cells with suspiciously high variances at the output stage, that is, when the estimates and variances have been produced. Erroneous or questionable data are corrected, identified for follow-up, or flagged as missing to be later imputed. Edits may be differentiated to declare resulting errors as either fatal or as suspicious.

Macroediting is carried out at the estimation stage. Errors in the data set missed by microediting are sought out via the analysis of aggregate data. The objective of the procedure is to detect suspicious data that have a large impact on survey estimates (Granquist, 1997). If this impact is quite large, suspicious data can be considered as outliers. Macroediting offers a number of advantages. First, significant cost savings can be obtained without loss of data quality. Second, data quality can be improved by redirecting resources and concentrating on editing of high impact records. Third, timeliness improvements are achieved by cutting down survey processing time and subject-matter experts' data analysis time. Finally, follow-ups are reduced, thereby relieving respondent burden.

A drawback of microediting is that too many records can be flagged for follow-up without accounting for the relative importance of individual records and the high cost in editing all records. This is remedied by *selective* editing, which cuts down on checking all records declared in error by focusing on a subset. Selective editing (also known as significance editing) selects records in error for follow-up if it is expected that the corrected data will have a large impact on the estimates. Such methods have been developed at Statistics Canada (Latouche and Berthelot, 1992), the Australian Bureau of Statistics (Lawrence and McDavitt, 1994; Lawrence and McKenzie, 2000), and the Office for National Statistics (Hedlin, 2003). Records in error that are not followed-up are imputed.

Latouche and Berthelot (1992) defined a score function to determine which records to follow up. Their score function was based on the magnitude of historical change for that record, the number of variables in error for that record, and the importance of each variable in error. One of the score functions that they suggested is given by

$$\text{Score}_k(t) = \sum_{q=1}^{Q} \frac{w_k(t)E_{k,q}(t)I_q\left(x_{k,q}(t) - x_{k,q}(t-1)\right)}{\sum_{s} w_k(t)x_{k,q}(t-1)}$$

where $E_{k,q}(t)$ equals 1 if there is an edit failure or partial nonresponse, and 0 otherwise; $w_k(t)$ is the weight for unit $k$ at time $t$, and $I_q$ reflects the relative importance for variable $q$. For example, if variable $x_q$ is considered more crucial or important than variable $x_{q'}$, then this is reflected in the score function by assigning a larger value to $I_q$ than $I_{q'}$. Suspicious records are ranked by their associated score. Records with scores above a given threshold are followed up.

Hedlin's (2003) procedure differs from Latouche and Berthelot's (1992) procedure in that his score function minimizes the bias incurred by accepting records in error. For a sample $s$, let the clean data be denoted as $y_1, y_2, \ldots y_n$ and the raw data as $z_1, z_2, \ldots z_n$. The score for $z_k$ is computed as $\text{Score}(z_k) = w_k \times |z_k - E(z_k)| / \hat{Y}$, where $\hat{Y} = \sum_{k \in s_d} w_k y_k$, and $s_d$ is part of the current sample for a specified domain $d$ of interest (say the $d$-th industrial sector). The $E(z_k)$ term is usually the previous "clean" value of that record $y_k$, or the median of the corresponding domain. A record $k$ will be rejected if $\text{Score}(z_k)$ exceeds a prespecified threshold based on historical data. The threshold is set so that the coverage probability of the estimate is nearly unchanged.

### 6.2. *Detection of outliers*

Outliers are a common feature of almost all surveys. This is especially true for business surveys due to the highly skewed nature of their data. Outliers may result in unrealistically high or low estimates of population parameters, such as totals.

Outliers can come from two sources. First, they can be erroneous values, due to data entry or measurement problems for example. Second, they can be improbable or rarely occurring, but valid values. Erroneous values that are detected as outliers should be corrected, or removed from the dataset. On the other hand, improbable values identified as outliers should be left in the dataset, but special treatments should be applied to them to reduce their effects on estimates.

Chambers (1986) classifies outliers in sample surveys into two groups: *representative* and *nonrepresentative*. Outliers are representative if they have been correctly recorded and represent other population units similar in value to the observed outliers. Nonrepresentative outliers are those that are either incorrectly recorded or unique in the sense that there is no other unit like them. Errors that lead to outliers should be detected and corrected at the editing stage. In what follows, we focus on outliers that are free of error, and such outliers may either be representative or nonrepresentative.

Suppose that we have observed a sample $s$ of size $n$ with values $y_k$, and associated weights $w_k$, $k = 1, \ldots, n$. Outliers will be *influential* if the joint effect of the data and associated weight is significant. This is so whether they are representative or nonrepresentative. A typical example is a frame that is out-of-date in terms of size classification of its units. Suppose that a unit classified as small or medium size should have been classified as a large unit. The joint effect of the sampling weight $w_k$ and large observed value $y_k$ may result in declaring unit $k$ as an influential observation.

In this section, we focus on a number of procedures to detect outliers. We present a number of those used in practice for business surveys. The treatment of outliers is discussed by Beaumont and Rivest (Chapter 11 of this book).

#### 6.2.1. *Top-down method*

This simple procedure sorts the largest entered values (top 10 or 20) and starts the manual review from the top or the bottom of the list. Units that have an abnormally large contribution to an estimator of interest such as the sample total are flagged and followed up.

Let $y_{(1)} \leq \ldots \leq y_{(n)}$ denote the ordered values of the observed $y$-values in the sample $s$. The cumulative percent distribution of the top $j$ units to the $y$ total of all the sampled units is computed. Unweighted or weighted versions of the cumulative percent distribution are computed. The unweighted version identifies units that may be in error. Once the unweighted top-down method is performed, the weighted version provides us with an idea of units that will be influential on account of their very large $w_k y_k$ product.

We can illustrate how the unweighted cumulative percent distribution is computed. The computations for the weighed version are identical with the exception of incorporating the weights $w_k$ into the computations. The cumulative percent contribution $P_{(j)}$ to the total for each of the $j$ top units is given by $P_{(j)} = 100 \times \sum_{k=j}^{n} y_{(k)}/Y_s$, where. For $j = n$, we have $P_{(n)} = 100 \times y_{(n)}/Y_s$. For $j = n - 1$, we have $P_{(j)} = 100 \times (y_{(n-1)} + y_{(n)})/Y_s$, and so on. More details on the top-down method are available in Granquist (1987).

### 6.2.2. *Standardized distance*

Let $z_k = w_k\, y_k$ be the product of the sampling weight $w_k$ and observed value $y_k$. Let $m_z$ and $\sigma_z$ be the estimates of the location and scale of $z_k$. A typical measure used to detect outliers is the standardized distance $\delta_{z,k} = (z_k - m_z)/\sigma_z$. A unit $k$ is identified as an outlier if the absolute value of $\delta_{z,k}$ is larger than a predetermined threshold. Location and scale estimates could be the sample mean and the standard deviation of the $z_k$ values. Such estimates are nonrobust because they include some of the potential outlier values. Including all units in the computations reduces their probability of being declared as outliers. This “masking” effect is avoided by computing robust estimates of $m_z$ and $\sigma_z$. Robust outlier-resistant estimates of $m_z$ and $\sigma_z$ are the median $Q_{2,z}$ and interquartile distance $(Q_{3,z} - Q_{1,z})$ of the $z_k$ values respectively, where the $Q_{j,z}$ values are the $j$-th $(j = 1, 2, 3)$ quartiles of the population (or the sample). Note that we could have used the nonweighted variable $\delta_{y,k} = (y_k - m_y)/\sigma_y$ as well. Units are declared as outliers if their $z_k$ values fall outside the interval $\left(m_z - \delta_{\text{Low}}\sigma_z,\ m_z + \delta_{\text{High}}\sigma_z\right)$, where $\delta_{\text{Low}}$ and $\delta_{\text{High}}$ are predetermined values. These bounds can be chosen by examining past data or using past experience.

### 6.2.3. *Hidiroglou–Berthelot method*

The standardized distance can be used to detect whether the ratio of two variables $y$ and $x$ for a given sampled unit differs markedly from the ratios of the remaining units. Such comparisons do not account for size differences between units. Incorporating a measure of size (importance) with each unit places more emphasis on small ratios associated with those larger values. Hidiroglou and Berthelot (1986) extended the standardized distance procedure by incorporating a size component, and transforming the ratios to ensure symmetry. The extended method has been adapted by several national agencies to detect suspicious units. The procedure consists of six steps: (i) Ratios $r_k = y_k/x_k$ are computed for each unit $k$ within the sample $s$. (ii) Data are transformed to ensure outliers can be detected at both tails of the distribution. The transformed data are given by $s_k = 1 - (\text{med}\, r_k)/r_k$ if $0 < r_k < \text{med}\, r_k$, and $r_k/(\text{med}\, r_k) - 1$ otherwise. (iii) The data’s magnitude is incorporated by defining $E_k = s_k \max(x_k, y_k)^{\phi}$ where $0 < \phi < 1$. These $E_k$ values are called effects. The parameter $\phi$ provides a control of the importance associated with the magnitude of the data. It controls the shape of the curve defining upper and lower boundaries. (iv) The first $(E_{Q1})$, second $(E_{Q2})$, and third $(E_{Q3})$ quartiles of the effects $E_k$ are computed. (v) The interquartile ranges $d_{Q1} = \max(E_{Q2} - E_{Q1}, |a\, E_{A2}|)$ and $d_{Q3} = \max(E_{Q3} - E_{Q2}, \left|a\, E_{Q2}\right|)$ are computed. The quantity $\left|a\, E_{Q2}\right|$ reduces the tendency of declaring false outliers, and “$a$” is usually set to 0.5. This problem may arise when the $E$ values are clustered around a single value and are one or two deviations from it. (vi) Units are declared to be outliers if their associated $E_k$ value is outside $(E_{Q2} - cd_{Q1}, E_{Q2} + cd_{Q3})$. The parameter $c$ controls the width of the acceptance region. Belcher (2003) suggested a procedure to determine the values of the different parameters entering the Hidiroglou–Berthelot method. Figure 4 illustrates how these steps lead to identifying outliers.

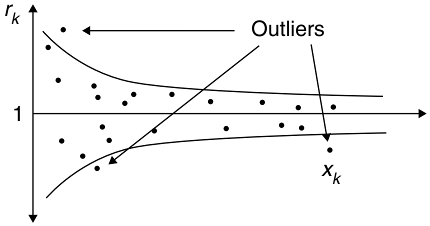


Fig. 4. Hidiroglou–Berthelot method.

### 6.3. Imputation

Edited records have either passed or failed the edits. A subset of these records may have been declared as outliers as well, regardless of their edit status. Records also considered as having failed edits are those that have not responded to the survey (or unit nonresponse), or that have provided incomplete data (partial response). Furthermore, some data items may have been manually deleted if they have been considered in error as a result of the editing process. The overall impact of edit failure is that it results in *missing data*.

There are several options available for dealing with missing data. The simplest one is to do nothing. That is, missing values are flagged on the output data file, leaving it up to the data user or analyst to deal with them. This "solution" to missing data is usually adopted when its reveal to be too difficult to impute values with sufficient accuracy. For example, this occurs for variables that have no direct relationship with any other collected variable. In farm surveys, for example, livestock and crops cannot be used to impute each other, and it is then preferable to leave the missing values not imputed.

Another option is to adjust the survey weights for nonresponse. Although this procedure is mainly meant for unit nonresponse, it can be used for partial nonresponse. However, the drawback is that there will be as many weight adjustments as there are missing fields across the records. Methods such as calibration, or mass imputation, are used to insure consistency between the resulting tabulations. These approaches have been considered by Statistics Netherlands for the construction of a consistent set of estimates based on data from different sources (see Kroese and Renssen, 2001).

The preferred option for survey users is to impute missing data within individual records. The imputation procedures should be based on the Fellegi–Holt principles (Fellegi and Holt, 1976) which are as follows: (i) data within individual records must satisfy all specified edits. (ii) The data in each record should be made to satisfy all edits by changing the fewest possible variables (fields). (iii) Imputation rules should be derived automatically from edit rules. (iv) Imputation should maintain the joint distribution of variables.

In business surveys, because there are usually strong accounting relationships between collected variables, manual imputation is often considered, especially for small surveys where the resources to be devoted to imputation systems are minimal. This approach is not reasonable as it can lead to different tabulations, thereby yielding inconsistent results. The resulting manual imputation may not be the best method to use if the imputation is based on incomplete knowledge.

Imputation methods can be classified as *deterministic* or *stochastic*. A deterministic imputation results in unique imputed data. Examples of deterministic imputation methods often used for business surveys include: logical, mean, historical, sequential (ordered) hot-deck, ratio and regression, and nearest neighbor imputation. A stochastic imputation results in data that are not unique, as some random noise has been added to each imputed value. Stochastic imputation can be viewed as being composed of a deterministic component with random error added to it. Stochastic imputation, unlike deterministic imputation, attempts to preserve the distribution of the data. Examples of stochastic imputation include random hot deck, regression with random residuals, and any deterministic method with random residuals added.

Imputation of plausible values in place of missing values results in internally consistent records. Imputation is also the most feasible option for partial nonresponse. Good imputation techniques can preserve known relationships between variables, which is an important issue in business surveys. Imputation also addresses systematic biases, and reduces nonresponse bias. However, imputation may introduce false relationships between the reported data by creating "consistent" records that fit preconceived models. For example, suppose it is assumed that $x < y$ for two response variables $x$ and $y$. If this constraint is false, imputing the missing variable $x$ (or $y$) will result in incorrectly imputed data.

Imputation will normally use reported data grouped within subsets of the sample or population. Such subsets are known as *imputation groups* or *imputation classes*.

### *6.3.1. Deterministic imputation methods*

*6.3.1.1. Logical (or deductive) imputation.* Missing values are obtained by deduction, using logical constraints and reported values within a record. Typical examples include deriving a missing subcomponent of a total. This type of imputation is often used in business surveys if there is a strong relationship between variables (especially financial ones).

*6.3.1.2. Mean value imputation.* Missing data are assigned the mean of the reported values for that imputation class. Mean value imputation should only be used for quantitative variables. Respondent means are preserved, but distributions and multivariate relationships are distorted by creating an artificial spike at the class mean value. The method also performs poorly when nonresponse is not random, even within imputation classes. It is often used only as a last resort.

Note that the effect on the estimates of mean value imputation corresponds exactly to a weight adjustment for nonresponse within imputation groups. Weighting is useful for unit nonresponse as relationships between variables are preserved.

*6.3.1.3. Historical imputation.* This is the most useful method in repeated economic surveys. It is effective when variables are stable over time, and when there is a good correlation between occasions for given variables within a record. The procedure imputes current missing values on the basis of the reported values for the same unit on a previous occasion. Historical trend imputation is a variant of the procedure: previous values are adjusted by a measure of trend, based on other variables on the record. Historical imputation is heavily used for imputing tax data for incorporated businesses at Statistics Canada (Hamel and Martineau, 2007). Historical imputation can be seen as a special case of regression imputation (see later).

*6.3.1.4. Sequential hot-deck method.* This method assumes that the order of the data items is fixed, even though they might have been sorted according to some criterion (measure of size or geography). Missing data are replaced by the corresponding value from the preceding responding unit in the data file. The data file is processed sequentially, storing the values of clean records for later use, or using previously stored values to impute missing variables. Care is needed to ensure that no systematic bias is introduced by forcing *donors* (reported data items) to always be smaller or larger than *recipients* (missing data items). Sequential hot-deck is used for the United States Current Population Survey.

Sequential hot-deck uses actual observed data for imputation. The distribution between variables tends to be preserved, and no invalid values are imputed, unlike with mean, ratio, or regression imputation. This is ensured if the imputed variables are not correlated with other variables within the record. If the variables are correlated (as it is the case with financial variables, for example), then imputing the missing variables by those from the preceding unit will not preserve the distribution. This problem is resolved, in practice, by imputing complete blocks of variables at a time: a block being a set of variables with no relationship with variables outside the block (see Hamel and Martineau, 2007).

*6.3.1.5. Nearest neighbor imputation.* Nearest neighbor imputation uses data from clean records to impute missing values of recipients. It uses actual observed data from recipients. Donors are chosen such that some measure of distance between the donor and recipient is minimized. This distance is calculated as a multivariate measure based on reported data. Nearest neighbor imputation may use donors repeatedly when the nonresponse rate is high within the class. This method is the second most heavily used one, after historical imputation, for imputing tax data for incorporated businesses at Statistics Canada (Hamel and Martineau, 2007).

*6.3.1.6. Ratio and regression imputation methods.* These imputation methods use auxiliary variables to replace missing values with a predicted value that is based on a ratio or regression. This method is good for business surveys when auxiliary information is well correlated with the imputed variable. However, accounting relationships between variables need to be respected, and this leads to the need for some adjustment of the predicted values. For example, suppose that the relationship $x + y = z$ holds for three variables $x$, $y$, and $z$ . If $x$ and $y$ are imputed, we might need to adjust the imputed values (e.g., by prorating) to insure that this relationship still holds.

The response variable needs to be continuous for these methods to be effective. For regression, the independent regression variables may be continuous or dummy variables if they are discrete. A disadvantage of this method is that the distributions of the overall data set may have spikes.

### *6.3.2. Stochastic methods*

Deterministic methods tend to reduce the variability of the data, and to distort the data distributions. Stochastic imputation techniques counter these negative effects by adding a residual error to deterministic imputations that include regression imputation. Another approach for stochastic methods is to use some form of sampling in the imputation process, as it is the case of hot-deck imputation. Whether or not stochastic methods are used, it is possible to compute variances that take into account the effect of imputation (see Beaumont and Rivest in Chapter 11 of this book).

## 7. Estimation

Business survey data are collected for reference periods that are monthly, quarterly, or annual. The resulting data are usually summarized as *level* and *change*. Level is measured as a total for a given variable of interest $y$. Change is defined as the difference between, or the ratio of, two estimated totals at two different time periods.

Factors that need to be taken into account for estimation include the sample design, the parameters to be estimated, domains of interest, and auxiliary data. The sample design is usually straightforward for business surveys. These surveys mostly use one-phase or two-phase stratified simple random sampling or Bernoulli sampling without replacement at each phase. Domain estimation is used in three ways for business surveys. First, the classification (i.e., geography, industry, or size) of the sampled units may differ from the original one. Second, the classification associated with domains for tabulating the collected data may differ from one used for the stratification purposes. Third, a unit originally sampled in-scope for a given survey may become out-of-scope either by ceasing its business activities or changing its classification to one that is not within the target population (e.g., a unit sampled within the retail sector becomes a wholesaler, which is out-of-scope to the survey).

The increasing use of auxiliary data for business surveys is associated with the wider availability of sources outside the survey, such as regularly updated administrative sources or annual totals from a larger independent survey. Auxiliary data yield several benefits. They improve the efficiency of the estimates when the auxiliary data (say $\boldsymbol{x}$) are correlated with the variable(s) of interest $y$. Given that there is some nonresponse, the potential nonresponse bias is reduced if the variables of interest are well correlated with the auxiliary data. A by-product of using auxiliary data is that their weighted totals add up to known population totals. We limit estimation to one-phase stratified Bernoulli sampling in what follows.

### *7.1. Estimation for level*

Let $U = \{1, \ldots, k, \ldots, N\}$ denote the in-scope population of businesses. The population of businesses is stratified by geography, industry, and size as $U_h$, $h = 1, \ldots, L$. The population size of $U_h$ is $N_h$, and a probability sample $s$ is selected from $U$ with inclusion probability $\pi_k$ for $k \in s$. If Bernoulli sampling has been used, the inclusion probabilities associated with stratum $U_h$ are given by $\pi_k = f_h = n_h/N_h$. The expected sample size within the sample stratum $s_h$ is $n_h$, and the realized sample size will be $n_h^*$. Each sampled unit $k \in s_h$ will have sample design weights given by $w_k = 1/\pi_k = N_h/n_h$.

Suppose that we wish to estimate the total of $y$ for a given domain of $U$, say $U_{(d)}$, $d = 1, \ldots, D$. This population total is given by $Y_{(d)} = \sum_{h=1}^{L} \sum_{k \in U_h} y_{(d)k}$ where $y_{(d)k} = y_k\, \delta_{(d)k}$ with $\delta_{(d)k}$ is one if $k \in U_{(d)}$ and zero otherwise.

If no auxiliary data are used, the population total $Y_{(d)}$ is estimated by the expansion estimator $\hat{Y}_{(d)}^{(\mathrm{EXP})} = \sum_{h=1}^{L} \hat{Y}_{(d),h}^{(\mathrm{EXP})}$ where $\hat{Y}_{(d),h}^{(\mathrm{EXP})} = \sum_{k \in s_h} w_k y_{(d)k}$. Although this estimator is unconditionally unbiased, it is conditionally biased given the realized samples size $n_h^*$ (see Rao, 1985). Given that Bernoulli sampling has been used, the estimated variance of $\hat{Y}^{(\mathrm{EXP})}_{(d)}$ will be $\mathrm{v}(\hat{Y}^{(\mathrm{EXP})}_{(d)}) = \sum_{h=1}^{L} \sum_{k \in s_h} (1 - \pi_k) y^2_{(d)k} / \pi_k^2$, which does not compare favorably with the corresponding expression for stratified *srswor* given the same (expected) sample sizes at the stratum level.

Consider the estimator

$$\hat{Y}^{(\mathrm{HAJ})}_{(d)} = \sum_{h=1}^{L} \frac{N_h}{\hat{N}_h} \sum_{k \in s_h} w_k y_{(d)k} \tag{4}$$

where $N_h$ are known population strata counts, and $\hat{N}_h = \sum_{k \in s_h} w_k$ is the estimated population strata counts (Brewer et al., 1972). This estimator, also known as the Hájek estimator, "adjusts" for the discrepancy between expected and realized sample sizes, assuming that $P\left(n_h^* = 0\right)$ is negligible for $h = 1, \ldots, L$. This estimator has the following two desirable properties. First, it is conditionally nearly unbiased given $n_h^*$. Second, its variance estimator is approximately equal to the one that we would get with *srswor* with realized sample sizes $n_h^*$ selected from populations of size $N_h$, $h = 1, \ldots, L$. That is,

$$V\left(\hat{Y}^{(\mathrm{HAJ})}_{(d)}\right) \doteq \sum_{h=1}^{L} \frac{N_h^2}{n_h} \left(1 - \frac{n_h}{N_h}\right) S^2_{h(d)} \tag{5}$$

where $S^2_{h(d)} = (N_h - 1)^{-1} \sum_{k \in U_h} \left(y_{(d)k} - \overline{Y}_{h(d)}\right)^2$ and $\overline{Y}_{h(d)} = \sum_{k \in U_h} y_{(d)k} / N_h$. This variance can be estimated using

$$v\left(\hat{Y}^{(\mathrm{HAJ})}_{(d)}\right) \doteq \sum_{h=1}^{L} \frac{N_h^2}{n_h^*} \left(1 - \frac{n_h}{N_h}\right) \hat{S}^2_{h(d)} \tag{6}$$

where $\hat{S}^2_{h(d)} = (n_h^* - 1)^{-1} \sum_{k \in s_h^*} \left(y_{(d)k} - \hat{\overline{Y}}_{h(d)}\right)^2$ and $\hat{\overline{Y}}_{h(d)} = \sum_{k \in s_h^*} y_{(d)k} / n_h^*$.

Estimator (7.1) is reasonable if the realized sample size is sufficiently large within each stratum $U_h$; if it is not, strata with insufficient realized sample sizes need to be combined with others to reduce the relative bias. The estimator $\hat{Y}^{(\mathrm{HAJ})}_{(d)}$ can alternatively be written as $\hat{Y}^{(\mathrm{HAJ})}_{(d)} = \sum_{h=1}^{L} \sum_{k \in s_h} w_k g_k y_{(d)k}$, where $g_k = N_h / \hat{N}_h$ for $k \in U_h$ is known as the *g-weight* (see Särndal et al., 1992).

The separate count ratio estimator is the simplest example of an estimator that uses auxiliary data. Multivariate auxiliary data can be incorporated into the estimation process via the well known regression estimator, $\hat{Y}^{(\mathrm{REG})}_{(d)} = \hat{Y}^{(\mathrm{EXP})}_{(d)} + \left(\mathbf{X}_d - \hat{\mathbf{X}}_d\right)' \hat{\mathbf{B}}_{(d)}$, where $\hat{\mathbf{B}}_{(d)}$ is obtained by minimizing the variance of $\hat{Y}^{(\mathrm{REG})}_{(d)}$. The regression estimator can also be written as $\hat{Y}^{(\mathrm{REG})}_{(d)} = \sum_{k \in s} \widetilde{w}_k y_{(d)k}$ where $\widetilde{w}_k = w_k g_k$ is known as regression weights. The regression weights are the products of the original design weights $w_k$ with the $g_k$-weights given by $g_k = 1 + \left(\sum_{k \in U} \mathbf{x}_k - \sum_{k \in s} w_k \mathbf{x}_k\right)' \left(\sum_{l \in S} \frac{x_l x_l^t}{\lambda_l \pi_l}\right)^{-1} \frac{\mathbf{x}'_k}{\lambda_k}$: the $\lambda_k$ term incorporates the optimality of the estimator. In the case of the ratio estimator, we have that $\lambda_k = c_k x_k$. The Huang and Fuller's (1978) iterative procedure is implemented in Bascula (Nieuwenbroek and Boonstra, 2001).

The calibration procedure of Deville and Särndal (1992) minimizes distance measures between the original weights and final weights $\tilde{w}_k$ subject to $\mathbf{X} =$

$\sum_{k\in s}\widetilde{w}_k\mathbf{x}_k$. They propose several such distance measures, and the one defined by $\sum_{k\in s}(\widetilde{w}_k - w_k)^2/w_k\lambda_k$ corresponds exactly to the regression weighs given by $g_k = 1 + \left(\sum_{k\in U}\mathbf{x}_k - \sum_{k\in s} w_k\mathbf{x}_k\right)' \left(\sum_{l\in s}\dfrac{x_l x_l^t}{\lambda_l \pi_l}\right)^{-1} \dfrac{\mathbf{x}_k'}{\lambda_k}$. It should also be noted that Deville and Särndal's procedure allows for bounding the $g$-weights. A good comparison of these two approaches in given is Singh and Mohl (1996).

The $\sum_{k\in s}\widetilde{w}_k\mathbf{x}_k = \sum_{k\in U}\mathbf{x}_k$ constraint can be applied to subpopulations $U_p \subseteq U(p = 1,\ldots,P)$ of the population $U$, where $U_p \cap U_{p'} = \emptyset$ for $p \neq p'$ and $U = \bigcup_{p=1}^{P} U_p$. These subpopulations are also referred to as poststrata. The previous constraint translates into $\sum_{k\in s_p}\widetilde{w}_k\mathbf{x}_k = \sum_{k\in U_p}\mathbf{x}_k$ where $s_p = s \cap U_p$. For this case, the $g$-weights are of the form:

$$g_k = 1 + \left(\sum\nolimits_{k\in U_p}\mathbf{x}_k - \sum\nolimits_{k\in s_p} w_k\mathbf{x}_k\right)' \left(\sum_{l\in s_p}\frac{w_l\mathbf{x}_l\mathbf{x}_l'}{c_l}\right)^{-1} \frac{\mathbf{x}_k}{c_k} \quad \text{for } k \in U_p \qquad (7)$$

Hidiroglou and Patak (2004) show that domain estimation efficiency (in terms of variance) is improved when auxiliary data are available for poststrata that are close to the domains of interest. This holds when there is a constant term in the auxiliary data vector, and when the variance structure associated with the model linking the variable of interest $y$ to the auxiliary data $\mathbf{x}$ is constant. It does not necessarily hold, however, that the incorporation of unidimensional auxiliary data at the poststratum level into $\hat{Y}_{(d)}^{(\mathrm{HAJ})}$ will increase the efficiency of the resulting estimator.

If the poststrata sizes are too small, raking on margin variables (counts or continuous quantities) will also improve the efficiency of the expansion estimator $\hat{Y}_{(d)}^{(\mathrm{EXP})}$. Hidiroglou and Patak (2006) displayed how gross business income incorporated into raking ratio estimation could improve the efficiency of the estimates of total sales for the Monthly Canadian Retail Trade Survey by raking on margins based on industry and geography.

Two-phase sampling is also used in business surveys to obtain (or inherit) relatively inexpensive first-phase information that is related to the characteristic of interest. This information may be in the form of discrete variables that yield estimated counts used in poststratified estimation, or continuous $x$ variables for improving the efficiency of the estimators of interest. A number of surveys use a two-phase sample design at Statistics Canada. An example is the Quarterly Canadian Retail Commodity Survey. This sample design was chosen to reduce collection costs by using as the first-phase sample the Canadian Monthly Retail Trade Survey. Auxiliary information (annualized sales) from the first-phase sample is used in all of the Canadian Retail Commodity Survey design steps to maximize the efficiency. More details of sample design are given in Binder et al. (2000).

### 7.2. *Estimation for change*

Let $U(t) = \{1,\ldots,k,\ldots,N(t)\}$ denote the in-scope population of businesses at the $t$-th survey occasion. Note that because of births and deaths of units, we expect $U(t)$ to be different from $U(t')$ for $t \neq t'$. As in the previous section, we consider the case where

Bernoulli sampling has been used, with inclusion probabilities associated with stratum $U_h(t)$ given by $\pi_k(t) = f_h(t) = n_h(t)/N_h(t)$.

At survey occasion 1, a sample $s(1)$ of size $n^*(1)$ has been selected, using the selection intervals defined by the starting point $\alpha(1) \in [0, 1]$. Recall that a population unit $k$ falling in stratum $h$ is included in the sample if $\alpha(1) < u_k \leq \alpha(1) + f_h(1)$, provided $\alpha(1) + f_h(1) \leq 1$ (see Section 5.1). For survey occasion 2, a rotation of $(100 \times r)\%$ of the sample has been performed by shifting the parameter $\alpha(1)$ by $r \times f_h(2)$ in all strata of survey occasion 2. The starting point $\alpha(2)$ is then given by $\alpha(1) + r \times f_h(2)$ (assuming, for simplicity, that $\alpha(1) + rf_h(2) + f_h(2) \leq 1$). This rotation results in a selected sample $s(2)$ of size $n^*(2)$, with $n^*(c)$ units in common with $s(1)$. Note that the inclusion probability $\pi_k(1, 2)$ of unit $k$ being selected in both survey occasions is given by $\pi_k(1, 2) = \min\left[\max\left[0, f_h(1) - rf_h(2)\right], f_h(2)\right]$.

Change between the survey occasions can be defined in a number of ways. A simple definition is to consider the difference $\Delta_{(d)} = Y_{(d)2} - Y_{(d)1}$, where $Y_{(d)t}$ is the population total of domain $d$ at survey occasion $t$. This difference can be estimated using $\hat{\Delta}^{(\mathrm{HAJ})}_{(d)} = \hat{Y}^{(\mathrm{HAJ})}_{(d)2} - \hat{Y}^{(\mathrm{HAJ})}_{(d)1}$, where $\hat{Y}^{(\mathrm{HAJ})}_{(d)t}$ is defined by (4) for $t = 1, 2$. The variance of $\hat{\Delta}^{(\mathrm{HAJ})}_{(d)}$ is given by

$$V\left(\hat{\Delta}^{(\mathrm{HAJ})}_{(d)}\right) = V\left(\hat{Y}^{(\mathrm{HAJ})}_{(d)2}\right) + V\left(\hat{Y}^{(\mathrm{HAJ})}_{(d)1}\right) - 2\mathrm{Cov}\left(\hat{Y}^{(\mathrm{HAJ})}_{(d)2}, \hat{Y}^{(\mathrm{HAJ})}_{(d)1}\right) \tag{8}$$

The variance $V\left(\hat{Y}^{(\mathrm{HAJ})}_{(d)t}\right)$ is given by (5), and it can be estimated using (6). The covariance $\mathrm{Cov}\left(\hat{Y}^{(\mathrm{HAJ})}_{(d)2}, \hat{Y}^{(\mathrm{HAJ})}_{(d)1}\right)$ is approximately given by

$$\mathrm{Cov}\left(\hat{Y}^{(\mathrm{HAJ})}_{(d)2}, \hat{Y}^{(\mathrm{HAJ})}_{(d)1}\right) \doteq \sum_{h=1}^{H}\sum_{g=1}^{G} N_{gh}(c)\frac{N_h(1)N_g(2)}{n_h(1)n_g(2)}\left(f_{gh} - \frac{n_h(1)n_g(2)}{N_h(1)N_g(2)}\right)S_{gh(d)} \tag{9}$$

where, to simplify the notation, we denoted the strata of survey occasion 1 by $h(h = 1, \ldots, H)$, and those of survey occasion 2 by $g(g = 1, \ldots, G)$. The quantity $N_{gh}(c)$ is the size of the common population $U_{gh}(c)$ crossing stratum $g$ of survey occasion 2 and stratum $h$ of survey occasion 1. The quantity $f_{gh}$ is given by $f_{gh} = \min\left[\max\left[0, f_h(1) - rf_h(2)\right], f_h(2)\right]$, and $S_{gh(d)} = \left(N_{gh}(c) - 1\right)^{-1} \sum_{k\in U_{gh}(c)} \left(y_{(d)k}(2) - \overline{Y}_{gh(d)2}\right)\left(y_{(d)k}(1) - \overline{Y}_{gh(d)1}\right)$ with $\overline{Y}_{gh(d)t} = \sum_{k\in U_{gh}(c)} y_{(d)k}(t)/N_{gh}(c)$ for $t = 1, 2$. The covariance (10) can be estimated using

$$\mathrm{Cov}\left(\hat{Y}^{(\mathrm{HAJ})}_{(d)2}, \hat{Y}^{(\mathrm{HAJ})}_{(d)1}\right) = \sum_{h=1}^{H}\sum_{g=1}^{G} \frac{n^*_{cgh}}{f_{gh}}\frac{N_h(1)N_g(2)}{n^*_h(1)n^*_g(2)}\left(f_{gh} - \frac{n_h(1)n_g(2)}{N_h(1)N_g(2)}\right)\hat{S}_{gh(d)} \tag{10}$$

where $\hat{S}_{gh(d)} = \left(n^*_{gh}(c) - 1\right)^{-1} \sum_{k\in s^*_{gh}(c)} \left(y_{(d)k}(2) - \hat{\overline{Y}}_{h(d)2}\right)\left(y_{(d)k}(1) - \hat{\overline{Y}}_{h(d)1}\right)$ and $\hat{\overline{Y}}_{h(d)t} = \sum_{k\in s^*_h(t)} y_{(d)k}(t)/n^*_h(t)$ for $t = 1, 2$. If the realized sample size $n^*_{cgh}$ is not sufficiently large in some cross-strata $gh$; then some collapsing must be done within the strata $g$ of survey occasion 2, or the strata $h$ of survey occasion 1, to reduce the relative bias.

*Sample Surveys: Design, Methods and Applications, Vol. 29A*
ISSN: 0169-7161

DOI: 10.1016/S0169-7161(08)00018-7

18

# Sampling, Data Collection, and Estimation in Agricultural Surveys

*Sarah M. Nusser and Carol C. House*

## 1. Introduction

Surveys that provide information about land use, land stewardship, agricultural production, and the economics of managing both the agricultural industry and our natural resources are of necessity complex, but vitally important. Agriculture is a $240 billion industry in the United States alone. It forms the foundation of an even larger food and fiber industry that contributes 12.3% of the U.S. gross national product and over 17% of the total employment. This industry provides $71 billion in U.S. exports, whereas the U.S. imports more than $65 billion in food and fiber products from other countries. Many developed countries in the world have similar economic dependence on their agricultural industry, and the economies of developing countries are tied even more tightly to their agrarian infrastructure.

Information obtained from surveys of agricultural production and the economics of that production allows the market economy to function efficiently. “With accurate information, individuals can make sound decisions that allow them to adjust their actions to the situation at hand. The value of publicly provided information [to these markets] is often underestimated.” (Roberts and Schimmelphennig, 2006). Apart from producers and buyers, this information is used by such entities as businesses supplying inputs to production, to those deciding where to build processing facilities, and those directing train car distributions. These surveys further provide the information needed by policy makers, local governments, academic researchers, and other stakeholders to extract knowledge needed for making informed decisions.

Farming and ranching are closely tied to the use and management of land. There are 1.9 billion acres of land in the United States, with nearly 1.4 billion acres in rural (non-Federal) areas. Non-Federal rural areas include about 400 million acres in active cropland and cropland set aside via conservation programs, more than 500 million acres of pastureland and rangeland used to raise livestock, and 400 million acres of forestland (Natural Resources Conservation Service, 2007a). The quality of this land is important to both agriculture and natural resources. For example, programs to reduce soil loss on cropland and increase wetlands on agricultural land have helped to generate reductions of 43% in soil erosion since 1982 (Natural Resources Conservation Service, 2007b) and net increases in agricultural wetlands of 98,000 acres during the last decade (Natural Resources Conservation Service, 2007c). Thus, surveys to obtain information about land utilization and stewardship are linked closely to those of agricultural production, and this chapter will include these in the discussions. These surveys provide information on such topics as the changing use of land, the spread of urbanization, the conservation practices employed on cropland, and the effectiveness of environmental policies.

From the perspective of applying statistical methods, conducting agricultural surveys requires the use of a distinct mix of methodology. Farms are businesses and much of the data that are collected on agricultural surveys are facts related to the operation of those businesses. During the past three decades, much of production agriculture in developed countries has become consolidated into larger operations. Now in the United States, less than 10% of the farms produce almost 75% of the total production. Many farming enterprises are part of vertically integrated, multinational corporations. Thus, survey methodology developed for business surveys is relevant to agricultural surveys in developed countries (see Chapter 17).

Although many farms have become corporate in nature, many more remain small and family operated. Within the United States, 90% of all farms are still small, family owned operations. In developing countries, small family run farms are the predominant type of production agriculture. Often, there is an interest in the farm household as well as the farm enterprise. Surveys of small farms, and particularly those collecting information about the farm household, may be similar to household surveys of other population groups. Thus, the survey methodology created for household surveys also applies to agricultural surveys (see Chapter 16).

Finally, some surveys of agriculture involve assessments of conditions on the land. These surveys consist of sampling land units and then collecting information and making direct measurements on those land units. This may be done on-site or remotely via imagery and without any input from a human respondent. Thus, methodological strategies that are designed to minimize error in measurements made directly from the field or other resources are also important. In addition, statistical approaches used in environmental surveys are relevant (see Chapter 19).

Thus, agricultural surveys present many challenges, most of which are related to the complex and diverse nature of the target population and the types of information acquired through these surveys. These challenges include:

- Using the appropriate mix of methodology for businesses, households, and direct measurements;
- Choosing a sampling frame that has both reasonable completeness and reasonable efficiency;
- Handling changes in the population, such as the consolidation of production agriculture into fewer population units from which more and more information is demanded, or changes in boundaries of land types;
- Coping with decreasing response rates;
- Creating estimates that are consistent with multiple independent administrative data sources and previously released estimates;
- Addressing the emerging need for all information to be geo-referenced;
- Serving the increasing need for small population estimates related to small geographic areas and specialized agricultural commodities that have been developed and marketed;

- Providing for the growing demand of analysts and researchers for more data, more access, greater accuracy, and to be able to import data into sophisticated tools; and
- Ensuring confidentiality of respondent data.

This chapter provides an overview of methods used for surveys of agricultural producers and land areas that focus on agriculture and natural resource concerns. In doing so, we emphasize emerging methods and future challenges in sampling, data collection, and estimation.

## 2. Sampling

In surveys of agriculture and the land, a careful review and understanding of the population inferences that are needed from the survey are critically important. This is certainly the case for all surveys, but these issues seem to be easily obscured in agricultural and land surveys. This is illustrated by the following example. There is a current need in the United States and Europe to understand how quickly farmland may be disappearing to development, and to make policy decisions to slow down that development. The first question to ask is "what is farmland." Is it land that is currently used for growing a crop or raising livestock? Is it more broadly any land owned or managed by a farming operation, whether or not it is currently used for producing crops or livestock? Is it any undeveloped land? Are woods and forests included? What if cattle are kept in those woods and forests? Do you include subdivisions of 10 acre homesites (perhaps with horses), or is this land already considered "developed"? The answers to these types of questions are essential to defining the target population, sampling units, reporting units, and observations for a survey.

Sampling methods for agricultural surveys are tied to land, either directly or indirectly. Not surprisingly, area frames are utilized for many agricultural surveys. An area sample frame allows for the direct sampling of units of land, usually through a multistage sampling process to improve efficiency in sampling and/or data collection. A major advantage of the use of an area frame is that it provides complete coverage of the targeted land area. It can be an efficient frame for drawing samples to collect data and make measurements that are highly correlated to that land area, such as land use, production of major crops, general conservation practices, etc. An area frame, although a complete frame, will generally provide very large sampling errors when used for collecting data on uncommon items. For example, the acres of corn in Iowa can be estimated very effectively with an area frame. The estimation of acres of apples in that same state would be problematic.

Area frames may be constructed in different ways. Several important area frames are in use for agricultural statistics, and we will describe in some detail three that have distinctly different designs. One has been built and maintained by the U.S. Department of Agriculture for the estimation of agricultural production. Land area in the United States is divided into primary sampling units (PSUs), which are then stratified by general land use classifications, and sampled at different rates. Strata definitions differ by state, but usually include the following breakouts: >75% cultivated, 51–75% cultivated, 15–50% cultivated, <15% cultivated, agri-urban, urban, nonagriculture, water. Sampled PSUs are further broken down into secondary sampling units of approximately one square mile in size, and sampled for data collection. Usually one secondary unit is selected within each PSU. Breakdown of all units are made along visible and natural boundaries such as roads, streams, ditches, etc. This approach to the development and use of an area sampling frame for agricultural statistics has been used in a number of developing countries.

Another important area frame is used by the U.S. Department of Agriculture (USDA, 2005) to monitor conditions on the land via the National Resources Inventory (NRI). The NRI is a large natural resource monitoring survey conducted by the USDA Natural Resources Conservation Service in cooperation with the Iowa State University Center for Survey Statistics and Methodology. This inventory captures data on land cover and use, soil erosion, prime farmland, soils, wetlands, habitat diversity, selected conservation practices, and related resource attributes. It does so primarily via remote sensing but is periodically augmented by a subsample of field visits (Nusser and Goebel, 1997). The NRI area frame encompasses all land in the United States and its territories. The sample for the first 1982 inventory formed the basis of subsequent inventories. The first-stage area sample was stratified within counties (or equivalents) using political divisions or geographic coordinate systems without regard to visible features or natural resource boundaries (Figure 1). The use of political or geographic stratum boundaries avoids problems associated with changing natural resource boundaries. Small strata were used in each county to ensure geographic spread and control over sampling rates for diverse geographic domains. Area segments within strata are generally a square of land one-half mile on each side, although segments may be smaller or larger as heterogeneity of land conditions warrant (Figure 1). Sampling rates within strata generally range from two to four percent. For nearly all sampled area segments, three points are selected within the segment using restricted randomization to encourage geographic spread. Sample points are used to obtain detailed trending information for most survey variables.

A third area frame is one developed for the USDA Forest Service's Forest Inventory and Analysis (FIA) program, designed to estimate conditions over time for private forests in the United States (Bechtold and Patterson, 2005). The frame design consists of a system of grid points that are the center coordinates of a hexagonal tessellation of the land surface of the United States. The FIA uses a three-phase sample design. Remote sensing methods are used in the first phase to classify these grid points into forest and non-forest strata. FIA's second and third phases are used to obtain field measurements. The second phase sample is used to obtain traditional inventory field measurements at a systematic subsample of approximately 125,000 first phase forest points. The third phase is a subsample of approximately 8,000 second phase points (about 1 in 16 phase two points) to collect more intensive forest health field measurements (USDA, 2007b). One-fifth of the Phase 2 and Phase 3 samples are observed each year. A complex plot design has been developed to facilitate collection of a wide variety of biological and environmental properties at different scales (Figure 2; USDA, 2007a). The layout at a Phase 2 point consists of four circular subplots with a 24 foot radius, one centered at the sample point and the other three surrounding the center point at regular intervals 120 feet from the center plot. A circular microplot with a 6.8 foot radius is also established within each circular subplot.

The integration of field and remote sensing surveys via two-phase sampling to extend the depth of information available for individual sampling units is also used with the NRI. Subsamples of NRI photo-interpretation survey samples are selected to investigate

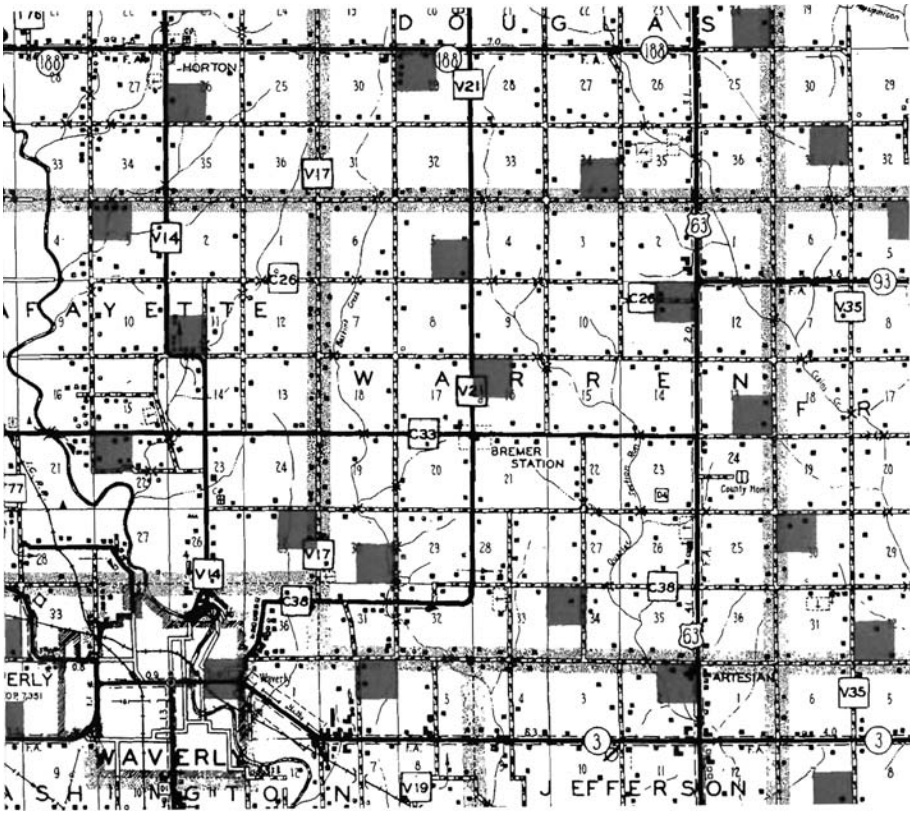


Fig. 1a. Area sample examples for the National Resources Inventory: (a) standard Public Land Survey sample in the central U.S. with a 4% area sample using third townships as strata and quarter sections as area sampling units (each square on the map is a 1 mi × 1 mi section, each solid sample unit is a 0.5 × 0.5 mi quarter section, each stratum is 2 mi × 6 mi).

field conditions for specific kinds of land, such as the health of rangeland or the quality of cropland soils. Over the last few years, the Conservation Effects Assessment Program has involved a collaboration among USDA agencies in which interviews were conducted with producers whose land was associated with a subsample of NRI points. This approach has led to more detailed practice information at a point than would be available via remote sensing and administrative databases, and has been used to develop simulation models that predict outcomes under alternative practices for policy evaluation.

In Europe, designs that involve remote sensing and field components are also being used to generate timely information on agricultural production for implementation of its Common Agricultural Policy. The MARS (Monitoring Agriculture by Remote Sensing) project is sponsored by the Directorate General Joint Research Center of the European Commission. This project provides crop and yield monitoring by utilizing agro-meteorological models, low resolution remote sensing methods, and area estimates using high resolution data combined with ground surveys. Echoing features of the FIA design, Italy's agricultural survey (AGRIT) also relies on satellite imagery to stratify points for

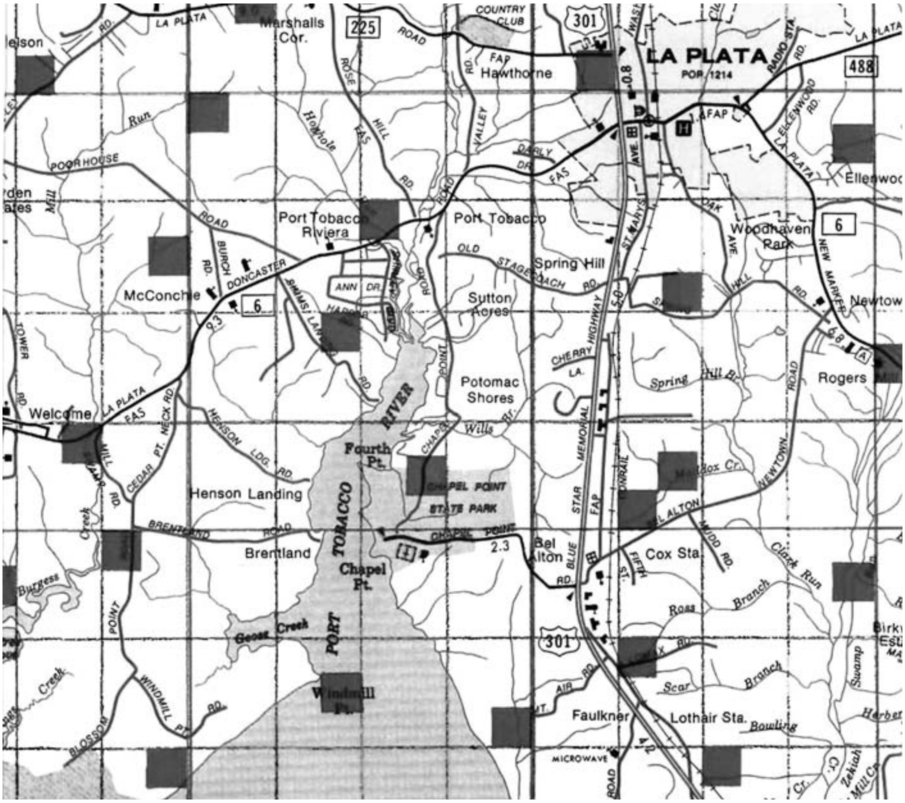


Fig. 1b. Area sample examples for the National Resources Inventory: (b) similar design in the eastern U.S. with latitude and longitude boundaries for strata and sampling units.

agricultural field surveys (Carfagna and Gallego, 2005). This is especially helpful in fragmented landscapes such as those in Europe.

Technologies used in area frame sampling have evolved considerably over the last 50 years. Initial frames relied on paper products—county road maps and aerial photography to define strata and area segments. Today, geographic information systems (GIS) are used to assist in stratification and in selecting and managing samples. Utilizing satellite imagery and digital aerial photography, the steps in the process involve computer-assisted delineation of sampling structures against a digital image. Delineation is improved by the capacity to view different scales. Further, digital storage of strata and segments enables more efficient production of sampling lists and field materials for the survey. For land-based surveys, the ability to quickly generate area sampling geometries on the earth's surface via GIS greatly increases flexibility in choice of design features. For example, the FIA program redesigned their survey based on a tessellation of hexagons for their first phase of sampling, an approach that would be difficult without GIS. Regardless of the type of sample units, the use of GPS (global positioning system) receivers enables field staff to find sample points created in GIS frames when they are not associated with visible features.

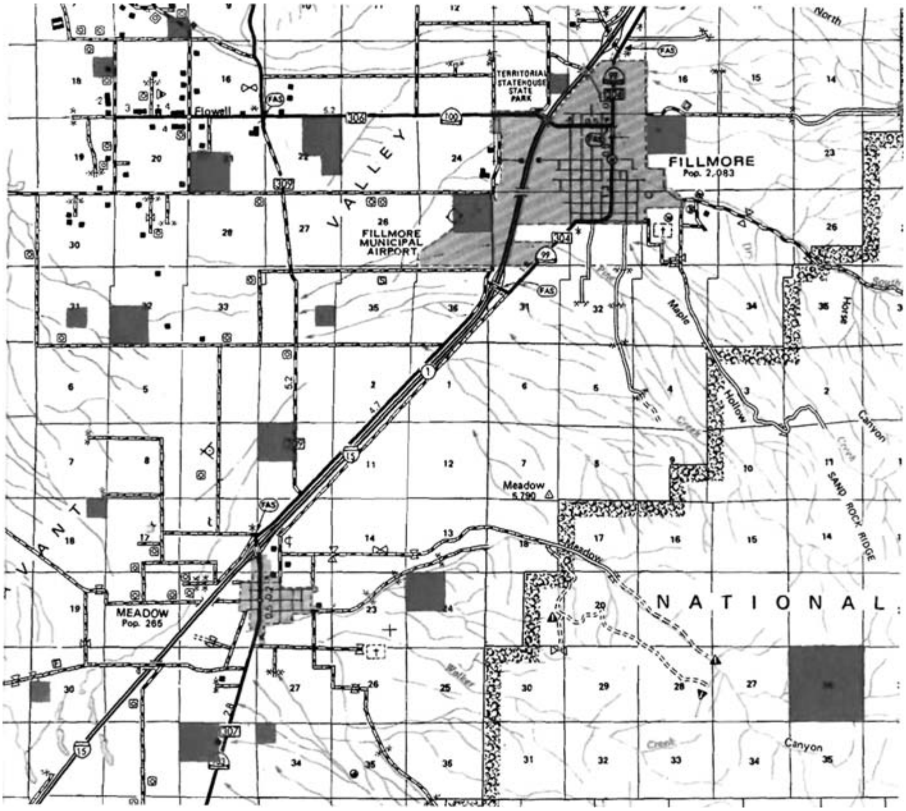


Fig. 1c. Area sample examples for the National Resources Inventory: (c) area sample in variable land use area of the western U.S. with smaller area segments for heterogeneous areas and larger segments in homogeneous areas.

List sampling frames are also very important for agricultural surveys. Generally, these lists are of land owners or of the operators/managers of the agricultural enterprises. These lists can be efficient sampling frames for agricultural statistics particularly if they contain useful information about the agricultural enterprises. Thus, a list of agricultural enterprises that have produced or sold hogs in the past could be an effective sampling frame for a survey to measure hog inventories. If that list also had information about how many hogs each enterprise produced in the past, it would allow stratification of the list for increased precision in the estimate.

Lists of agricultural enterprises may be available through producers associations, producer oriented magazine publishers, or certain publicly available administrative record systems. Most such list sources are likely to have considerable "out-of-scope" records, and will also suffer from serious under-coverage of the target population. Local land and tax offices may have information on land owners that could be used to build a list frame of such owners.

Many governments provide financial and technical support for agricultural production and land stewardship through various programs, and create administrative record

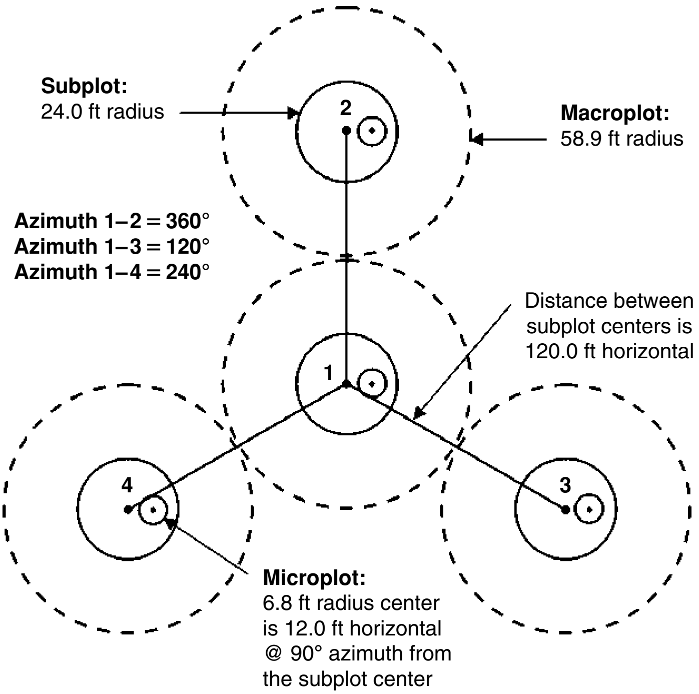

Subplot:
24.0 ft radius
Macroplot:
58.9 ft radius
2
Azimuth 1–2 = 360°
Azimuth 1–3 = 120°
Azimuth 1–4 = 240°
Distance between subplot centers is 120.0 ft horizontal
1
4
3
Microplot:
6.8 ft radius center is 12.0 ft horizontal @ 90° azimuth from the subplot center

Fig. 2. The Forest Inventory and Analysis plot design (USDA, 2007a).

systems to run these programs. If accessible, these records can be very valuable in sampling for agricultural surveys. (They can also be extremely valuable in direct estimation, as will be discussed later in this chapter.) Administrative records are created when farmers sign up for production support payments, when they receive government assistance for land or waste management enhancements, or when they apply for a license to build new facilities or to spray certain pesticides.

List sampling from administrative records is perhaps one of the most widely used methodologies for agricultural surveys worldwide. Canada conducts its Census of Agriculture every five years in conjunction with its Census of Population. The database generated from this activity creates a list sampling frame for agricultural surveys. This sampling frame may be updated for births and deaths, and augmented by specialty lists. Countries with centrally controlled economies (including those which have had such economies in the past) usually have extensive administrative systems of records. The Russian Federation, for example, bases its frame for surveys of rural agricultural households on its “Land Taxpayer Register.” The only ancillary variable for stratification is total land area. South Africa historically used administrative data from its marketing board to generate information on agricultural production. As that country has moved to a free market economy, that administrative source has become much less complete as a sampling frame. Many of these countries are experimenting with area frames, remote sensing, and other methodological innovations. For example, South Africa is experimenting with a point sample area frame survey in which land use and crops are identified from an ultra-light aircraft.

Sampling from a list frame for general purpose agricultural production surveys can be very complex, even when a good sampling frame with ancillary data with size of operation and commodities produced is available. Often the intent is to measure agricultural production for numerous crop and livestock commodities within a single survey. For efficiency, larger producers (who can report for a larger percentage of the total commodity) should be sampled more heavily than smaller producers. Because farmers do not produce every commodity, care must be taken to allocate the sample so that sufficient responses are received to estimate all targeted commodities. Stratified sampling can be used for this purpose. Strata would be defined first by commodity, and then by the size of operation for those producing the commodity. Because many farming operations will generally produce more than one commodity, each could be eligible for inclusion in more than one stratum. A priority system may be used to assign each population unit to one and only one stratum. Such a system should give higher priority to the less common commodities.

The stratified sampling design described earlier can be effective when only a few commodities are targeted. When there are many commodities, a large number of strata are needed and the prioritization becomes more complex. An alternative sampling procedure for general purpose agricultural production surveys is discussed by Kott and Bailey (Bailey and Kott, 1997; Kott and Bailey, 2000). Their approach selects a Poisson sample (Ghosh et al., 2002) from multiple list frames and then uses calibration estimation. Kott and Bailey propose independently assigning a Poisson Permanent Random Number (PRN) to each population unit. They then create a separate list frame of farming operations for each commodity, creating multiple partially overlapping frames. A specific population unit would be in a commodity frame if ancillary information indicated that farming operation produced the commodity. Many population units would be in more than one commodity list frame. A targeted minimum sample size is determined for each commodity, and a probability proportional to size sample simultaneously drawn from each frame. Use of the permanent random numbers forces overlaps between the commodity samples, thus minimizing the total sample size. Kott and Bailey describe the process of how this overlap is forced, as well as how to calculate the probability of inclusion in the overall sample. In practical application, instead of having a separate sampling frame for each commodity, it may be more appropriate to group similar commodities (such as major row crops) within the same commodity frame.

Using a list frame and an area frame in a multiple frame design can be very effective for agricultural surveys. A well-developed list of enterprises with appropriate ancillary variables can provide a very efficient sample, but may also have considerable incompleteness. An area frame sample may be less efficient for collecting certain types of data, but offers complete coverage. Using standard multiple frame methodology (see Chapter 4), estimates from the overlap domain are weighted by the inverse standard errors for each frame. In practice, the standard errors using the area frame are often considerably larger than those of list frame and thus the overlap domain is estimated almost exclusively from the list frame.

Repeated observations over time in surveys that target units of land are possible, and often desirable, but will increase the complexity of sample designs (Fuller, 1999). Survey objectives that involve estimating both status at a particular point of time and trends over time are in conflict and need to be balanced. For surveys that target agricultural enterprises directly, longitudinal designs are problematic because individual farming enterprises change their structure over time. They go in and out of business, take on or dissolve partnerships, change their land holdings, and change names. Any longitudinal survey that requires human response needs to take into account respondent burden and the effect of frequent contact on both response rates and the quality of response.

Land surveys with area sampling frames and longitudinal data sets generate a different set of concerns, which are well expressed through the current design for the NRI. Pressures for more frequent NRI data led to a redesign in 2000 involving a shift to an annual survey. This replaced a design that involved observations on 300,000 area segments once every five years from 1982 to 1997. Research evaluating the trade-offs between status and trend estimates resulted in a supplemented panel design in which 40,000 segments are observed every year along with a rotating panel of approximately 30,000 segments that augments the continuous panel (Breidt and Fuller, 1999). The rotation period varies across types of segments. Segments with points that have land/cover uses that are more likely to change or are of primary interest are given shorter rotation periods. Longer rotation periods are defined for other segments. The 1997 NRI is viewed as a first phase sample for subsequent annual samples for the purposes of estimation (Legg et al., 2006).

Land surveys may seek to produce a time series of information about land-based sample units for constructing tables of gross change (Fuller, 1999). Gross change tables provide estimates of change into and out of various categories of land, and offer a detailed understanding of the dynamics of change. Area data can be used for this purpose, but it is quite difficult to track polygons that represent changes in the extent of area features over time. A more practical method is to use repeated observations at a specific point on the land, which generates direct time series information to support gross change estimates.

## 3. Data collection

Data collection methods vary with the type of observation unit, the type and complexity of information to be collected, the time frame for data collection, and budget considerations. Many surveys require producer inputs and thus involve collecting data from farm operators. Responses may be collected either in person, by telephone, through the mail or via the web, to collect data on crop and livestock production, conservation practices, and/or pesticide use. As with the other household or business surveys, mail collection can be very economical, but may not yield the required response rates. Multimode surveys are common, using less-expensive mail or web surveys coupled with nonresponse follow-up on the telephone or with in-person visits. Web surveys have an important future for surveying agricultural producers and agri-businesses, especially as internet access becomes more consistent in rural areas. In 2005, 51% of all farmers in the United States, and 72% of commercial farmers, reported having internet access. However, most reported still using dial-up rather than high-speed service, and as yet most appear not to be inclined to respond to surveys over the web.

Most in-person agricultural surveys in the United States are completed using paper survey instruments, in part because of the difficulty of conducting interviews in outdoor farm environments. However, recent research indicates that as technologies for tablet computers improve, it will be feasible to conduct computer-assisted interviews for production surveys (Nusser, 2004). In-person interviews may focus on one or more parcels of land operated by the producer, and thus often rely on an aerial photograph to identify and document parcel boundaries. With the use of computer-assisted interviews, it will be possible to use GIS-based capture of parcel boundaries on digital photographic backdrops (Nusser, 2004). Such an approach also provides the opportunity for computer-based measurement of areas associated with fields and buffers, which may be more accurate than operator-reported areas and less burdensome to obtain.

Land-use surveys often involve field observations of attributes at points, along transects or within areas, including possibly the location and extent of geographic features. However, remote sensing and image analysis have become very effective and efficient means of collecting information on land usage and management. Data collected via remote sensing generally involve variables that can be directly observed using geographic information sources (e.g., aerial photographs, satellite images, topographic maps, soils maps). These may be the extent or surface area of geographic features on the land, or conditions at a specific line segment or point on the land. To supplement directly observable data, some elements, such as conservation program participation, may be obtained from administrative data, again not requiring contact with a human respondent. These approaches save respondent burden and data collection resources. However, even though the costs of remote sensing materials are far less than in the past, these costs still may be substantial, particularly if high-resolution materials are desired for detailed observations.

Computer-assisted data collection methods for land features are considerably more complex than standard questionnaires, but surveys are beginning to take advantage of advances in geographic information technologies. In recent years, the NRI has moved to directly capturing boundary delineations of features such as water bodies and building structures using custom-developed geographic data tools (Nusser, 2005). The geographic interface is integrated with more traditional computer-assisted forms to record land classifications. Prior to the availability of geographic tools in the survey instrument, data collectors manually measured abstracted summaries of features, such as areas and lengths. The geographic survey instrument now enables collection of raw information on the feature, and algorithms process the geographic data.

Although land use surveys rely more heavily on remote sensing imagery and analysis, they may still require on-site field measurements that are tied to remotely sensed delineation and classification of land areas. In 1996, the NRI began using handheld computers with formal computer-assisted survey instruments in the field with coordinates from the photo-interpretation survey displayed on a GPS receiver (Nusser and Thompson, 1997). Today, GPS can be provided as a resource, integrating easily with a mobile computer and digital imagery.

Repeated observations in longitudinal surveys introduce special problems for remotely sensed or field observations at a sample point. The key concept of repeated measurement is to return to the same physical location as data were collected in the past. It is tempting to consider the stored GPS coordinate as the gold standard for a sample or reporting unit's location, but positional error exists in GPS receivers and the coordinate systems of imagery and other geospatial data layers. These errors may exist even after the coordinates have been orthorectified (a process that aligns a geospatial data source to a standard coordinate system that adjusts for the presence of three dimensional terrain features). Material from the previous observation is required to ensure that the new observation is taken in the same location as the prior observation.

## 4. Statistical estimation

Statistical estimation encompasses data processing, imputation, weighting, and estimation of parameters and variances of parameter estimates. As with any survey, estimation approaches depend greatly upon the objectives of the survey. The choice of methods for surveys that release one-time results are generally less constrained than those that can be used in longitudinal panel surveys in which a series of related estimates are released over time. If inference for small areas is an objective, more complex estimation techniques are needed. Similarly, methods for creating public release data sets are more involved than those used for surveys where only estimates are released.

Even somewhat static estimation of agricultural production and related activities is complex because it generally involves the careful integration of information from various surveys (often collected at different times over the marketing year) and from administrative sources of varying quality. The goal of these endeavors is to estimate current year supply of an agricultural commodity, which can be used with information on demand for that commodity to provide appropriate transparency for proper functioning of the marketplace.

Vogel and Bange (1999) discuss the complexities of crop production estimates and forecasts produced by the U.S. Department of Agriculture and the need for a careful integration of survey and administrative data, both from domestic and foreign sources. The components of production are measured separately, and span the production year. First, planted acreage is estimated based on a multiple frame survey conducted in June. This initial estimate of planted area is subject to revision later in the season when administrative data is available from an USDA production support program. This administrative data provides the area planted to major crops by farming operations signed up for the support program. Because most field crop producers enroll in these programs, these estimates from administrative data must be integrated with survey estimates of planted area. As the season progresses, additional surveys are conducted to estimate area that will be harvested. This amount may vary greatly from year to year, and change throughout the production year, based on both weather and market conditions. Yield is forecasted monthly during the growing season using a combination of estimates from surveys of producers and from direct measurements of the crop in the field. Data from both the producer surveys and the direct measurement surveys are used in models that forecast yield per acre. At harvest, additional surveys (both surveys of producers and those involving direct measurements only) are conducted to estimate harvested area, yield and production.

Following harvest, additional administrative or survey data become available showing utilization of the crop production. For example, all cotton will be ginned. There is no other utilization for this crop. A monthly census of cotton gins around the country provide a critically important estimate of total cotton production that must be integrated with estimates from earlier production surveys to improve the overall estimate of cotton production. Corn is even more complex because some corn enters the market (for food, feed, or ethanol production) and some is fed on-farm to livestock. Corn may be stored for some time without further processing. Administrative and survey data providing estimates of corn exports, corn imports, corn storage, and corn fed to animals are all utilized in an estimation "balance sheet" (see Figure 3) with the initial corn production

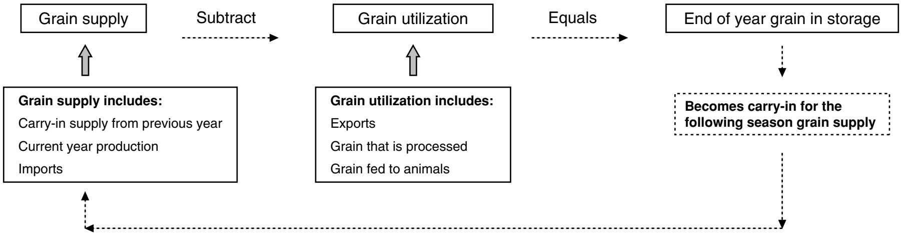


Fig. 3. Balance sheet for grain production and use.

estimates to ensure that the total estimates of corn production, marketing and utilization are consistent for the marketing year.

An even more complex assemblage of methods is required when surveys release related estimates over time and/or of microdata that can be further explored by policy analysts and researchers. The NRI survey program periodically releases estimates of conditions and trends for natural resources as well as public use data sets that have a complete time series for all points in the data set. Numerous issues are addressed in the estimation process. These include ensuring sensible temporal patterns for observed and imputed point data, integrating known trends from external data sources such as program participation data and land areas, appropriately reflecting these trends in small area estimates, and promoting consistency with estimates generated from previous releases. These problems require sophisticated estimation methods and if data are to be released, estimation methods must also create a data set that is simple for analysts to generate routine statistics and their estimated variances.

To address these issues, the NRI uses several strategies. In data review, historical time trends are given special attention to evaluate whether rare or impossible time series have been created in the classification sequences for a point. However, it can be difficult to identify all unusual or unexpected time series, or even inconsistent values for suites of related variables. A recent advance was made by Wang and Opsomer (2006), who describe the use of cluster algorithms to identify unusual combinations of values or unusual trends over time. This approach may reduce the need to create specific computer checks or manually review data for unusual trends. Small changes can also occur in individual time series for continuous measures that are entirely due to measurement error and do not represent meaningful changes. Smoothing procedures have been developed to minimize the impact of such fluctuations.

To improve small area estimates, an imputation approach is used to ensure area data known to correspond to segments or watersheds within counties are geographically allocated to these same locations for subpopulation estimates (Breidt et al., 1996; Nusser and Goebel, 1997). NRI sample area segments can be viewed as representing the first phase in a two-phase sample with points as the second phase. However, instead of using segment data in a standard two-phase weighting approach (see Chapter 3), segment area data are used to impute points that reflect observed segment changes when such patterns are not observed in the segment's points. Weights that reflect the size of segment areas corresponding to specific time trends are assigned to imputed and observed points that match conditions observed in the segment data. Local attribution of these areas improves the small area properties of estimates relative to using a standard two-phase estimator (Breidt et al., 1996).

As mentioned earlier, it is important that NRI–generated estimates are consistent over time. It is equally important that they are consistent with information from certain administrative record systems (such as that tabulating area enrolled in the Conservation Reserve Program). In NRI, consistency with prior estimates or administrative totals is handled via raking and ratio estimation methods.

Recent research in the NRI has considered an estimated generalized least squares estimator (EGLSE) to incorporate correlations across sample panels and the full 1997 NRI sample (Legg et al., 2005, 2006). These estimators are consistent and asymptotically as efficient as generalized least squares estimators, which have minimum variance among the class of linear unbiased estimators. EGLSEs are used in a ratio step near the end of the weighting process to improve the properties of estimates for key variables.

Because of the complexity of the NRI estimation process, replication methods are used for variance estimation. Replicates are created for delete-a-group variance estimation (Kott, 2001; Lu et al., 2006) that attempt to reflect both sampling variance and variance due to estimation methods such as point imputation. Another approach being investigated, but not yet implemented, is fractional imputation (Kim and Fuller, 2004). In fractional imputation, multiple imputation outcomes are generated for a point and the weight associated with the point prior to imputation is divided equally among the imputation realizations for the point. This approach is expected to provide a better estimate of variance due to imputation than the current method.

The National Agricultural Statistics Service also uses satellite data to improve the precision of estimates of crop acreage, especially at the county level (Allen et al., 2002). Studies have shown that relative efficiencies of 3.0 or more have been achieved over the area frame estimate of planted acreage. However, Allen et al. (2002) point to complexities in this process: "There is a common misunderstanding that crop type signatures are so unique that they could be determined once and for all. Then later classifications would be a matter of running a new satellite data file against known parameters. This is called signature extension. Satellite-based crop classification is based on the measurement of energy emitted or reflected by plants. Those readings do differ somewhat from one crop type to others in different wavelengths. However, that pattern differs throughout the growing season of a particular crop. There can also be considerable differences between healthy plants and plants of the same crop but under serious stress. The density of crop planting and the presence or absence of weeds and recent precipitation also can affect crop response. On top of all other factors, the atmosphere through which the crop response is being measured is not the same from one day to the next."

Because most agricultural surveys are inherently linked to the land, there is great pressure to use the geospatial locations of sample points, even if they are not publicly available. Research agreements can be used as one vehicle to enable point-specific modeling. Agencies are also looking towards the possibility of using image classification in creating map-based products. Supervised image classification involves using ground truth (or reference) data in combination with satellite imagery to produce a data layer depicting a specific theme, such as land cover or crop cover. For example, in selected states, NASS has created a popular data product called the Cropland Data Layer, which is a public use GIS data file with crop specific categorization (at 30 m resolution) for each crop season. This data layer is used in conjunction with other GIS layers to aid in watershed monitoring, soils utilization analysis, agribusiness planning, crop rotation practices analysis, animal habitat monitoring, and prairie water pothole monitoring (Allen et al., 2002).

## 5. Confidentiality

Maintaining the confidentiality of individual survey responses is a critical part of the survey estimation process for agricultural and land based surveys, as it is for surveys of other populations. Disclosure avoidance issues for farms mirrors many of the issues with surveys of businesses (see Chapter 15). Namely, there are large or specialized producers that are easily identifiable even when their responses are combined with other responses. Thus disclosure avoidance programs and processes must ensure an appropriate number of responses in an estimation cell, but also must ensure that one or two responses do not individually account for a predominant portion of the cell total.

Confidentiality methods for surveys that involve observations of the land are simpler than for surveys that involve respondents. Coordinates of sample points are not publicly released in order to protect land owners and to preserve the integrity of plots that are revisited over time. Thus, the primary concern in these surveys is to evaluate disclosure risks within geographic polygons created with classification variables such as county, watershed or eco-region. The NRI combines polygons with small sample sizes with adjacent polygons to reduce disclosure risk. As the time series at a point becomes more extensive, additional disclosure limitation methods may be needed.

## 6. Concluding remarks

Agriculture and land utilization issues sit at the vertex of many global concerns and conflicts. The disagreement over governmental subsidy programs for agricultural production is a major barrier to negotiating international free trade agreements. There are ongoing concerns about shortages of food in drought stricken parts of the world. Global concerns about human health extend to food-borne diseases such as bovine spongiform encephalopathy (BSE) in cattle, avian influenza, and *E. coli*. There is global concern about the clearing of forestland for agriculture and about the conversion of agricultural land to development. One common thread through all of these concerns is the need for high quality information with which to make policy and trade decisions. Agricultural and land-based surveys will continue to be an important tool for information gathering in the foreseeable future.

There are a number of methodological opportunities and challenges ahead for these surveys. From the data user viewpoint, there are several critical needs. Because a single survey will never encompass the breadth of interacting components of the agricultural sector, there is an increasing need for linking data from surveys and other administrative sources together in appropriate ways to support more global analyses. There is an increasing need to link data of all types back to a land base, and then to be able to layer that data within a GIS to draw inferences based on increasingly smaller geographic areas. There is an increasing need for improved methods of small area estimation in general. This extends not only to facilitate analysis of small geographic areas, but also because agricultural commodities are becoming so specialized that finer and finer breakouts of production are required. There is an increasing need to look at change over time, and to create data and techniques that are appropriate for longitudinal analysis. Finally, data users want more access to disaggregate data, they want online access to that data, and they want online tools available with the data sets that will allow them to perform a wide range of analysis.

From the data providers' viewpoint, there are equally critical needs ahead. In developed countries there continues to be fewer enterprises engaged in commercial agriculture. Thus, the population base from which to sample for agricultural surveys is shrinking while the need for data from that population is growing. Methodological challenges before us are to minimize the reporting burden by improving electronic reporting capabilities, increasing the utility of other sources of information, and developing and improving various modeling techniques that require less input data. Another critical need for data providers is to maintain the confidentiality of their individual responses and of remotely sensed data associated with a producer's land. There are significant methodological changes ahead as we try to meet data user needs of more online access to disaggregate data from a shrinking population that is becoming more and more concerned about privacy. From a statistical perspective, these requirements ultimately lead to more complex sampling and estimation methods, particularly for surveys conducted over time, which must be balanced with the need to provide easily accessible and usable data to researchers and the public.

## Acknowledgments

Some of the research discussed in this chapter was supported in part by Cooperative Agreement No. 68-3A75-4-122 with the USDA Natural Resources Conservation Service, U.S. Department of Agriculture. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do neither necessarily reflect the views of the Natural Resources Conservation Service, the National Agricultural Statistics Service, nor the U.S. Department of Agriculture.

*Sample Surveys: Design, Methods and Applications, Vol. 29A*
ISSN: 0169-7161
© 2009 Elsevier B.V. All rights reserved
DOI: 10.1016/S0169-7161(08)00019-9

19

# Sampling and Inference in Environmental Surveys

*David A. Marker and Don L. Stevens Jr.*

## 1. Introduction

In this chapter, we focus on surveys of environmental resources, which we loosely define as the air, water, soil, and associated biota that sustain our environment. The objective of the surveys we consider will generally be an assessment of status, condition, or extent of a resource. The target population of the survey may be discrete and finite, for example, small lakes or wetlands, with well-defined population units; or may be a one-, two-, or three-dimensional continuum, for example, a stream network, a forest, or the volume of water in a large lake. Each of these calls for different types of frames and sampling techniques. The survey may be a one-time assessment or may include a long-term monitoring objective to assess change or trend. Addressing both objectives requires a balance of revisiting sites to assess trend and adding new sites to assess status.

Traditionally, the focus of sampling in the environmental sciences has been on relatively small and well-delimited systems, e.g., at the scale of a lake or watershed or forest stand. However, some current environmental issues, such as global warming, contamination of surface and ground water by pesticides and other pollutants, and extensive landscape alteration are not localized. Quantifying the extent of symptoms of widespread concerns requires large-scale study efforts, which in turn needs environmental sampling techniques and methodology that are formulated to address regional, continental, and global environmental issues.

Survey design is a well-developed and established area in the statistical literature. There are many textbooks that provide excellent accounts of the essential attributes of good survey design, such as the necessity of clear definitions of the population of interest, the sample units, the sample frame, and how the sample is to be drawn (Cassel et al., 1993b; Cochran, 1977; Kish, 1967; Lohr, 1999; Sarndal et al., 1992; Thompson, 2002; Yates, 1981). However, designs for environmental sampling often present additional challenges which we identify below. These include the need for broad population description; spatial context of the population; availability of ancillary information; inadequate frames; difficult access; multiple objectives, including status and trend; evolving objectives; and the need to satisfy multiple stakeholders.

The focus of most survey methodology is estimating the mean value or total of a population. In contrast, an environmental survey often has a more general object, such as estimating the distribution function or the proportion of the population in various classes, for example, the proportion of lakes that meet designated use criteria. There may be many environmental responses of interest that are interdependent. A common objective of environmental surveys is to characterize the status of some resource as well as the change or trend in that status. These two objectives have somewhat conflicting design criteria: status is generally best assessed by sampling as much of the resource as possible, whereas trends are generally best detected by observing the same resource locations over time. Frequently, a secondary objective is the evaluation of relationships between attributes, both measured at the site and available on the frame.

A characteristic of overwhelming importance for environmental populations is that they exist in a spatial context. The response will have spatial pattern and structure. Sites near to one another will tend to have similar physical substrate and be subjected to similar stressors, both natural and anthropogenic. Response can be influenced by topography, hydrology, and metrology. All these influences will tend to induce spatial patterns in the response.

Another important characteristic of environmental populations is that some ancillary information (in addition to location) is almost always available. Currently, there is a wealth of remotely sensed information available from satellites or aerial photography that may be used to structure the sampling design or used in analysis.

Environmental resources are often expensive and time-consuming to sample. Logistics can be difficult; often, the population of interest includes sites in remote locations, for example, lakes in wilderness areas. There can be considerable time and money expended in traveling between sites. Laboratory costs for analyzing individual samples may be nontrivial. Some environmental metrics can be time-consuming to evaluate, for example, quantifying the species richness, and abundance of a macroinvertebrate sample requires the services of a skilled benthic taxonomist. For a large program, it may be a year or more after data are collected in the field before laboratory analyses are available. These may also be subject to substantially more measurement error than routinely found in other types of surveys. Nonresponse in environmental sampling can be substantial for reasons such as ease of physical access, safety, or permission.

A practical complication frequently encountered in environmental sampling is the difficulty in obtaining an accurate sampling frame. In many instances, available sampling frames include a substantial portion of nontarget elements or fail to cover the entire population. The frame problem is aggravated by the sheer difficulty of collecting and analyzing samples.

Environmental sampling almost always occurs with a backdrop of political, economical, and societal considerations so that statistical considerations represent only one aspect of a sampling design. Furthermore, because environmental issues can impact human populations, there are often multiple groups, agencies, and organizations that have an interest in the products of the survey. The interests of the multiple stakeholders are not always perfectly aligned. Meeting the interests of multiple stakeholders, while maintaining a scientifically and statistically rigorous design, can be a challenge.

In many instances, the need for an environmental sample will be driven by the need to assess the condition of an environmental resource because of concern over potential degradation. The design needs to address current environmental issues but that is not sufficient. The current issues will eventually be resolved, but new, presently unrecognized issues will emerge. These issues will manifest themselves in unforeseeable ways, and they will affect resources that cannot now be identified. An environmental assessment program with the dual objectives of status and trend must be able to accommodate regrouping, recombining, expansion and contraction of the sample to permit such emerging issues, and evolving objectives to be addressed. The issues of inadequate frames, nonresponse and missing data, and evolving objectives drive a need for sampling designs with the flexibility to add, remove, or reallocate samples.

Below, we review some of the sampling methodology that has been developed to meet the challenges that sampling environmental populations present: focus on broad population description; spatial context; ancillary information; inadequate frames; difficult access; evolving objectives; and the need to satisfy multiple objectives and stakeholders.

## 2. Sampling populations in space

Historically, many environmental samples have been chosen for convenience or subjectively to be representative. Both of these selection methods have severe shortcomings (Paulsen et al., 1998; Peterson et al., 1999). There are two widely accepted, statistically and scientifically rigorous approaches to selecting an environmental sample: probability-based and model-based. These two approaches begin from different theoretical bases but can both address the common working objective. In probability-based sampling, the response is viewed as fixed but unknown. A model-based approach views the response as one realization of a random process. In environmental sampling, we are usually interested in an attribute of an environmental resource, and a probability-based design objective is to estimate that attribute. A parallel model-based design objective is to predict the outcome of the process and to calculate the attribute from the predicted outcome. The focus in this chapter is on probability-based design and inference. However, there are some insights from model-based optimal design that are relevant to probability-based sampling, and these are discussed in Section 19.7.

Because space has a central role in environmental sampling, much of the relevant theory and practice has dealt with spatial sampling. A number of authors have investigated designs for sampling in space. The papers that have a sampling theoretic orientation tend to consider only finite populations; some of those with a spatial statistics model-based orientation consider continuous populations. Overton and Stehman (1993) compared three designs (systematic (SYS), simple random sampling (SRS), and random tessellation stratified (RTS)), then contrasted them in terms of their precision and variance-estimation properties. They concluded that the designs ranked RTS<SYS<SRS in order of increasing variance. Many investigations of two-dimensional sampling have taken a superpopulation approach (Cochran, 1946; Das, 1950; Quenouille, 1949). Matérn (1986) investigated sampling in continuous two-dimensional space and derived some comparisons of stratified and systematic sampling, using several systematic arrangements and spatial covariance functions. He concluded that a systematic sample on a triangular grid was optimum for a wide class of nonincreasing, isotropic covariance functions. Olea (1984) compared several variations on systematic designs that give good sample dispersion yet avoid the potential problems with periodicity that strict systematic designs have. Iachan (1985) derived some asymptotic comparisons of two-dimensional sampling designs and extended Cochran's (1946) result that under some restrictions on

the covariance function, systematic sampling is more efficient than stratified sampling, which in turn is more efficient than simple random sampling. Dalenius et al. (1961) showed that with some restrictions on the spatial covariance, sample designs using a triangular grid are optimal, a result supported by later work by McBratney et al. (1981) and Yfantis et al. (1987).

## 3. Defining sample frames for environmental populations

Probability-based sample designs were defined in Chapter 1. The first characteristic of a probability sample is that each unit in the population must have an explicit definition. That definition is used to develop a frame for the population, that is, a construct from which population elements can be selected via a random process. The construction of a frame is a first step for any probability sample, but environmental frames can take some different forms than that usually found in survey methodology.

In the case of discrete environmental resources with distinct sampling units (such as lakes), it would be possible in concept to develop an exhaustive list of all elements of the resource. The Eastern Lake Survey (ELS) (Linthurst et al., 1986) and the Western Lake Survey (WLS) (Landers et al., 1987), which were conducted by the EPA as a part of the National Acid Precipitation Assessment Program (NAPAP), took this approach. The frame for these surveys was developed by listing each lake in the target region on U.S. Geological Survey (USGS) 1:250,000-scale topographic maps.

A serious drawback to a list frame is the amount of time required to construct the list. Even in those fortunate circumstances when a nearly ready-made frame exists, considerable effort must be expended to verify that the frame completely covers the population of interest without excessive inclusion of nontarget populations. In the ELS, for example, investigators discovered that many bodies of water represented as lakes on 1:250,000-scale maps were in fact bogs, intermittently flooded areas, or wide spots in a stream. These nontarget units can be eliminated in the sampling, but they do complicate population estimation procedures. There is no easy way to compensate for units in the target population that were omitted from the list.

Another strategy for sampling environmental resources is to develop an area sampling design based upon a single-area sampling frame. In such a frame, the entire region to be monitored (e.g., the conterminous United States) is partitioned into a set of mutually exclusive and exhaustive areas. These areas are frequently designated primary sampling units (PSUs). The partition can be based on arbitrary geometric figures or on some characteristic of the landscape, such as the USGS hydrologic cataloging units. Commonly, PSUs are chosen with boundaries that are easily discernible in the field, such as permanent roads, railroads, or rivers. A sample is selected from these PSUs according to a probability-based protocol, such as selecting a PSU with a probability proportional to its size. Usually, some restriction is imposed on the sample selection to ensure spatial dispersion of the sample. The resources occurring in each sample PSU are identified, characterized, and measured.

Land use/land cover databases are available that cover the entire United States. These databases can provide ancillary information, such as land use, land form, soils classification, vegetation cover, hydrology, and human-induced modifications that can be used to structure a sampling design. Additional information is available from existing meteorological databases.

The National Hydrography Database (NHD; USGS, 2000) is arguably the best information available for water bodies in the United States, but it is not an ideal frame. Although attributes within NHD can be used to identify a subset of NHD that more closely matches the target, the subset may still include many nontarget entries. For example, suppose the original goal was to describe the population of lakes in California that support fish populations; the NHD has over 9000 entries identified as lakes in California. Many of these are very small (over 1000 are less than 1 ha, and over 5000 are less than 5 ha). Many of those smaller entries identified as lakes are in fact not lakes; they are farm ponds, sewage-treatment holding ponds, or intermittent seepage basins that are dry most of the year.

Conversely, the best available frame may miss some of the target resource. A geographic information systems (GIS) coverage was most likely based on aerial photos, and the utility of the coverage as a population frame is dependent on the skill of the photo interpreter and cartographer. The best source of frame information for wetlands in the United States is the National Wetland Inventory (NWI; USFWS, 2002; http://www.fws.gov/nwi/). However, it can be very difficult to identify wetlands in forested areas because of canopy cover. As many as of 50% of wetlands can be missed in the NWI (Brooks et al., 1999).

A grid can be used to frame a population that is distributed over some spatial extent by superimposing a grid over a representation of the spatial extent and then sampling at or around each grid point. Randomized systematic grids have a long history of application in environmental sampling. At the national level, they have been used extensively by the National Forest System (NFS) and Forest Inventory and Analysis (FIA) (Bickford et al., 1963; Gillespie, 1999; Hazard and Law, 1989) to sample forest growth and production. The National Stream Survey (Kaufmann et al., 1988; Messer et al., 1986) also used a grid frame to locate stream segments for sampling.

With the availability of GIS, electronic representations of maps are becoming more common for environmental populations. The NHD is the most complete information to be had on the extent and location of aquatic resources in the United States. The U.S. EPA's Environmental Monitoring and Assessment Program (EMAP; Messer et al., 1991; http://www.epa.gov/emap/) uses the NHD as a preferred frame (Angradi, 2006). With a GIS representation of a frame, the population can be viewed as finite or as a continuum of points. For example, a list of coordinates of lakes can be obtained from the NHD. Conversely, a forest, a large estuary, or a stream network may be treated as continua. Sample sites can be identified by choosing points from the GIS representation.

The use of a GIS can facilitate the preservation of the spatial context of the sample points. At a minimum, spatial context is the information required to locate a sample point on the landscape, for example, latitude and longitude. However, there is a richer connotation in all the available landscape information that is also attached to geographic coordinates: ecoregion, land use, soil topography, and so on. Knowing the spatial context of a sample from a resource, that is, knowing where the samples are located, and knowing their spatial relationship to one another provides the link of proximity to admit the joint evaluation of multiple resources and to evaluate the effects of stresses with known spatial properties.

A cautionary note on the combining of GIS data from multiple sources is that different levels of accuracy can result in inconsistencies in the derived sampling frame. It is important when combining multiple data sources to check for unacceptable combinations and clean these before proceeding.

## 4. Designs for probability-based environmental samples

The simplest and easiest to implement sampling method is simple random sampling (SRS) (Cochran, 1977). Simple formulae apply to population attribute and variance estimation. However, SRS is rarely appropriate because auxiliary information and prior knowledge about the population are not used in the selection. SRS samples will be inefficient compared to methods that do utilize knowledge about population characteristics or structure.

### *4.1. Multistage designs*

In a multistage cluster design, an area selected at a given stage is further split into subareas, and a sample is selected from the subareas. Complete characterization and measurement take place only at the lowest order set of areas. This is essentially the design used by the National Agricultural Statistics Service (NASS http://www.nass.usda.gov/research/AFS.htm; Cotter and Nealon, 1987; Mazur and Cotter, 1991) in their June Enumerative Survey of national agricultural production, where each sampled PSU is split into secondary sampling units called segments. Field visits are made to a sample of segments.

The National Resource Conservation Service (NRCS) (formerly the Soil Conservation Service (SCS)) also uses a two-stage design in several national resource surveys (Goebel, 1998; Goebel and Schmude, 1982; Nusser and Goebel, 1997). The 1958 Conservation Needs Inventory (CNI) used a frame based on 100-acre squares of land in the northeastern states and partitions of public land survey sections (approximately 640 acres) in the rest of the country. The 1967 CNI treated the 1958 sample areas as PSUs and subsampled within them at specific points. The 1977 National Resource Inventory (NRI) also used the 1958 CNI area frame and a two-stage sample. The 1982 NRI also used a two-stage area frame based on public land survey sections in most cases. A similar design was used in 1987 and 1992 (Goebel, 1998).

#### *4.1.1. Spatially constrained designs*

Some prior knowledge about environmental populations is always available. We may have reason to believe that the response is influenced by or is related to a variable for which we have complete information, for example, from remote sensing techniques. One important item of information that is always available for environmental populations is location. As noted in the introduction, environmental populations invariably exhibit spatial structure and pattern.

The advantage of spatial control accrues from the tendency of elements of an environmental population that are near one another to be more similar than elements that are far apart. Observations of elements that are near one another contain redundant information. Thus, samples that are well dispersed over the population domain tend to lead to more precise estimates of population attributes than samples without spatial control. The advantage of a spatially dispersed sample has long been recognized; accordingly, there are many techniques for achieving that dispersion, including area sampling, spatial stratification, systematic and grid-based sampling, spatially structured list frames, and spatially balanced designs.

### *4.1.2. Spatially balanced designs*

The notion of a balanced sample was introduced by Yates (1946). A sample of $Z$ is *balanced* over an auxiliary variable $X$ if the $x$-values (which are known beforehand) are chosen so that the sample mean of the $x$-values is exactly equal to the true population mean of $X$. A stricter version of balance was suggested by Royall and Herson (1973), who required that the first several sample moments of the $x$-sample match the population moments. The intuition behind balancing is that the auxiliary variable is correlated with the unknown response to be assessed. By balancing over the auxiliary variable, we hope to get approximate balance over the unknown response and hence to get a more precise estimate than SRS would give. Kott (1986) noted that an option intermediate between random sampling and strict balancing can be obtained by splitting the range of $X$ into $n$ quantiles and picking one sample element in each quantile. Although this option does not achieve balance in the strict sense of having sample moments match population moments, it does guarantee that the sample distribution function of $X$ will be close to the true distribution function for every sample draw. This is the idea behind all stratified sample designs. Because of the correlation between $X$ and $Z$, the hope is that the sample of $Z$ will be more precise.

If the ancillary variable is location, then we define a sample to be *spatially balanced* if the spatial moments of the sample locations match the spatial moments of the population. The first two spatial moments are the center of gravity and the inertia. The center of gravity for a region $R$ is given by the ordered pair $(\mu_x, \mu_y)$, where $\mu_x$, the moment about the $y$-axis, is given by $\mu_x = \int_{-\infty}^{\infty} x\nu_y(x)\mathrm{d}x$. The function $\nu_y(x)$ is the extent of the cross section of $R$ at the point $x$ and is given by $\nu_y(x) = \int_{-\infty}^{\infty} I_{\{w|(x,w)\in R\}}(y)\mathrm{d}y$. Similar definitions hold for $\mu_y$ and $\nu_x$. The second spatial moment is analogous to the covariance matrix and measures the regularity of the shape of $R$ or of the point pattern formed by the sample points.

In general, a probability sample will not achieve exact spatial balance, but approximate spatial balance is a worthwhile goal. Also, the discrepancy between the sample moments and the population moments can be used as a measure of spatial balance of a sample. The techniques described below for achieving spatial control all do better at achieving spatial balance than does SRS.

## *4.2. Stratified designs*

Sample designs can almost always be improved by introducing stratification. As discussed in Chapter 1, frames can be stratified to assure representation in the sample for units with particular characteristics or to improve precision of estimates. Stratification by analytic domains of interest can assure representation of each domain, thereby improving small area estimation (see Chapters 31 and 32 and Marker, 2001). If units within strata are more homogeneous than the population as a whole, then stratified designs (with corresponding estimators) can improve survey precision. While many environmental surveys use stratified list sample designs similar to those discussed in earlier chapters, the remainder of this section describes other types of stratified designs more common in environmental settings.

### *4.2.1. Systematic sampling*

For some kinds of environmental resources, systematic sampling is an attractive means of achieving spatial dispersion. For a two-dimensional, extensive resource, for example, a forest, a systematic sample can be obtained by placing a grid over a map of the resource and selecting the center points of grid cells or intersections of grid lines as sample points. Olea (1984) discusses several alternate ways of picking points in grid cells so that strict alignment is avoided. Randomness can be achieved by random placement of the grid. For a stream network, sample points could be picked at regular intervals along the network, starting at the outflow and working upstream. A rule for how to proceed at confluences would be needed.

A potential drawback of systematic samples is that they can align with natural or anthropogenic features of the landscape. If those features also influence the response, then high variability of estimators can result. This phenomenon is usually cited in relation to periodic or near-periodic responses but can also occur in responses with a mosaic structure. Another shortcoming with a systematic sample is its inflexibility. Frame errors or inaccessible sites are not easily accommodated, nor is variable probability. A sample can be locally intensified, say by halving the grid spacing, but there are a limited number of intensification factors available (Dacey, 1964; Hudson, 1967). Finally, systematic samples do not yield unbiased variance estimation formulas.

### *4.2.2. Spatially stratified designs*

One of the most popular means of achieving a spatially dispersed sample is through the use of spatial stratification. Strata are defined to be disjoint polygons that tile or tessellate the target domain. Strata can be regular geometric figures such as grid cells; arbitrary polygons such as ecoregions; political boundaries such as state or county borders; or natural boundaries such as drainage basins. Maximal spatial balance will generally be achieved by maximal dispersion over the domain, which in turn will be obtained by choosing strata with few samples per stratum. The aim of defining the strata should be to have an equal amount of the resource (number, length, area) and equal number of samples in each stratum, resulting in an equiprobable design. Commonly, samples are selected within strata using SRS, but other techniques could be used. If the design is not equiprobable, then the aim should still be to have a constant number of samples per stratum, but the amount of the resource per stratum will vary.

Maximal stratification achieves good spatial control, but having only a few samples per stratum limits flexibility. Given the difficulties of environmental sampling, it would be quite possible to lose all the samples from a stratum because of inaccessibility. If lost samples were replaced, as could be done if SRS is used for within stratum selection, then the inclusion probability could be substantially different for that stratum.

Forming strata with equal size and equal number of samples is usually straightforward for equiprobable designs and two-dimensional target resources but can be problematic for finite resources of unequal size. For example, lakes are often treated as finite populations for sampling purposes. The size distribution of lakes is heavily skewed toward small lakes (Larsen et al., 1994; Stevens, 1994). An equiprobable sample would result in mostly very small lakes, which are not likely to be the lakes that are of most interest (e.g., the ones that are accessible, support recreational or commercial fisheries, support other recreational use, have developed shorelines, or are subject to development impacts). Also, the spatial pattern of lakes tends to be clumped rather than uniform. To reap a benefit from spatial stratification, the strata should encompass a more or less homogeneous area of spatial influence, for example, land use/land cover, terrain, ecoregions, and anthropogenic impacts. This suggests strata that are spatially compact, with small perimeter to area ratio. Forming such strata (spatially compact with equal number of samples) can be difficult. Stevens (1994) describes an algorithm used by EMAP to form spatial strata of lakes.

### *4.2.3. Random tessellation stratified designs*

A compromise between SRS and SYS designs is a RTS design. An RTS design is implemented by randomly placing a grid over the population domain and selecting one point at random in each grid cell. See Olea (1984) or Overton and Stehman (1996) for a discussion of RTS designs. For a linear resource, the design is implemented by systematically dividing the resource into units with equal length and then picking a point at random in each unit. It can also be applied to a finite resource by picking one unit at random from the units covered by a grid cell. Although the RTS design does give good spatial dispersion, it also suffers from the same lack of flexibility and unbiased variance estimator that a systematic design does.

### *4.2.4. Spatial address techniques*

One method that has been used to disperse points in space is to induce a linear order on points in two-dimensional space, apply that order to the population elements, and then use systematic sampling along the ordered population. For example, NASS (Cotter and Nealon, 1987; Mazur and Cotter, 1991) has used a serpentine order to arrange the PSUs. Saalfeld (1991) discussed a method for sampling a connected tree structure, such as a stream network, by starting at the base of the network, tracing up one side (following all tributaries to their source) and then down the other side to the point of beginning. The resulting path traces each stream segment on both sides and is thus twice as long as the total length of the network so that every point on the network is mapped to two points on the path. A systematic sample along the path will have good spatial dispersion.

Some methods for creating spatial addresses are related to the concept of space-filling curves, such as first constructed by Peano (1890). Wolter and Harter (1990) have used a construction similar to Peano's to construct a "Peano key" to maintain the spatial dispersion of a sample as the underlying population experiences births or deaths.

The Peano key is an example of a spatial address created via a quadrant recursive function (Mark, 1990). Without loss of generality, we can assume that the two-dimensional population domain has been scaled and translated into the unit square. A quadrant recursive (q-r) function maps the unit square onto the unit interval and has the property that subquadrants of any order are mapped onto subintervals. This property preserves some two-dimensional proximity relationships in the one-dimensional image (Mark, 1990). As the name implies, a q-r function is defined recursively. We illustrate the construction of a q-r function with the Peano key. First, divide the unit square into four quadrants, which are labeled 0 through 3, beginning in the lower left, proceeding up, diagonally down, and then up to end in the upper right (see Fig. 1). The second step then divides each quadrant into four subquadrants labeled in the same order. Successive steps continue the subdivision process to smaller and smaller subquadrants. Figure 1 illustrates the process, with the subdivision carried out only in the first subquadrants. A spatial address is constructed by joining the labels attached to the subquadrants, beginning with the first division and proceeding down the chain, and treating the resulting number as a base 4 fraction. Thus, every point in the crosshatched subquadrant in Fig. 1 will get an address beginning with $0.001_4$. If this process was carried out indefinitely, then the limit is a measurable, 1-1, onto function from the unit square to the unit interval (Stevens and Olsen, 2004).

The basic quadrant recursive function is made into a random map by randomly and independently permuting the order in which labels are attached to the quadrants, at every possible opportunity. This randomization, termed hierarchical randomization (Stevens and Olsen, 1999), preserves the quadrant recursive nature of the map.

Stevens and Olsen (2004) use recursive partitioning to develop a very general technique, the Generalized Random Tessellation Stratified (GRTS) method, for selecting approximately spatially balanced designs. The concept underlying GRTS is to apply recursive partitioning to create a spatial address. At each step in the recursion, the total inclusion probability for each cell is computed as the sum or integral of the inclusion probability of all population elements within the cell. The inclusion probability need not be constant and very general variable probability designs can be accommodated.

The recursion is continued until every cell has total inclusion probability less than one and then hierarchical randomization is applied. The process is illustrated in Fig. 2. Part (A) of the figure shows the q-r address of the first 16 cells, with a line connecting

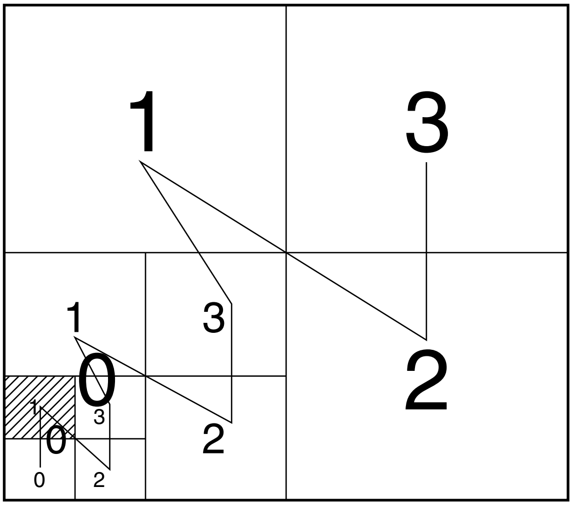


Fig. 1. Illustration of the recursive partitioning steps in construction of the Peano key.

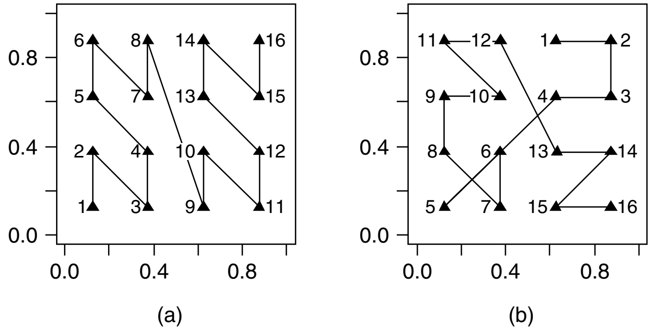


Fig. 2. Example of GRTS q-r addressing and sample location. (a) Nonrandomized q-r address for the first two levels. (b) Hierarchically randomized q-r address.

the cells following the q-r order. In Fig. 2(B), the path connecting the cells follows a hierarchically randomized order. Each cell is assigned a length equal to its inclusion probability, and then the lengths are strung together, forming a line with length equal to the total sample size. A systematic sample is selected along the line. Because of the 1-1 nature of a quadrant-recursive map, every point on the line corresponds to some population element, so the selected points on the line can be mapped back to specific population elements. Details are given in Stevens and Olsen (2004).

One additional step gives tremendous flexibility to the GRTS technique. The samples, as selected, will appear in the proximity-preserving order inherited from the randomized q-r address. Stevens and Olsen (2004) show how to order the sample points so that any consecutive subsequence of the sequence has good spatial balance. This property allows adjustment of sample size based on field experience or changing priorities, adjustment for nontarget sites, and formation of interpenetrating temporal panels.

## 4.3. *Other sample designs*

### 4.3.1. *Adaptive sampling*

Some environmental populations have spatial structure that makes them difficult to sample efficiently, even when using stratagems for spatial balance. For example, natural populations frequently exhibit clustering: individuals of the same type or species tend to group together. One potential technique for improving sample efficiency for clustered populations is adaptive sampling (Thompson, 1990, 1991b). Adaptive sampling allows one to modify the sample based on information as it is collected. The basic idea is best illustrated with an example. Suppose that a regular square grid has been placed over the domain of some clustered population. Further, suppose that the clusters tend to be of a size that covers several grid cells so that the grid cell area is substantially smaller than the average cluster size. An initial sample of grid cells is selected, the cells in the initial sample are visited, and the response (e.g., the number of individual members of the target species in the grid cell) is recorded for each cell. If the response meets some criteria (e.g., number of observed individuals is positive, or greater than some number), then adjacent cells are added to the sample. This sequence of observation/augmentation is continued until no newly observed cell meets the criteria of triggering augmentation.

The resulting sample presents some analysis difficulties because the inclusion probability of a cell is impossible to calculate without complete knowledge of the population structure, which is not available. Thompson (1990) shows how to obtain some modified weights that permits unbiased estimates of the total using an estimator similar to the Horvitz–Thompson estimator. Christman (1997) compares the efficiency of several designs for sampling clustered populations and concludes that adaptive sampling is an efficient sampling scheme for rare, tightly clustered populations (Chao and Thompson, 2001).

There are also some difficulties in applying adaptive sampling in the field. The rule for adding to the sample must be formulated prior to beginning sampling and must be followed in the field. In particular, new neighboring sites must be added so long as the site just observed meets the criteria. Some investigators have reported that has lead to unmanageable sample sizes (Hanselman et al., 2003; Kimura and Somerton, 2006). Thompson (2006) has recently extended the allowable stopping criteria to permit more control over the evolution of the sample. In particular, the new methodology allows the investigator to ensure a fixed sample size. However, the procedure can be computationally intensive.

### *4.3.2. Mark/recapture studies*

Estimation of the size of a wildlife population is a frequent need in environmental studies. Many fish and wildlife populations in the United States are listed under the Endangered Species Act as being either threatened or endangered, and there is a consequent legal requirement to track the abundance of those species. Additionally, most state fish and wildlife agencies use population size information to manage harvest levels and set fishing and hunting seasons.

One of the most popular methods of estimating the size of a wildlife or fish population is known as mark/recapture. In a basic mark/recapture study, an initial sample of individuals is collected, tagged with some permanent mark, and then released. A subsequent sample records the number of marked individuals recaptured from the first sample as well as the total number of individuals. The simplest mark/recapture models for estimating population size assume a closed population (there are no additions to or removals from the population during the observation period), and that each individual has a constant and equal probability of capture at each trapping occasion (Otis et al., 1978; White et al., 1982). For a single recapture event, with data consisting of the number marked (M), the total number in the second sample (C) (including recaptured), and the number recaptured in the second sample (R), Chapman's (1951) estimator of the population size is

$$\hat{N}_C = \frac{(M+1)(C+1)}{R+1} - 1$$

However, in practice, basic mark/recapture studies are rarely used because the required assumptions are not likely to hold. Open population models for mark/recapture studies, known as Jolly-Seber models, were introduced by Jolly (1965) and Seber (1965). The closed-population assumptions have been relaxed (Link and Barker, 2005; Pradel, 1996; Schwarz, 2001; Schwarz and Arnason, 1996, 2000) in a variety of ways to permit estimation of apparent survival and recapture probabilities, the population size at each trapping occasion, and the number of individuals entering the population at each occasion. Seber and Schwarz (2002) provide a recent review of the state of capture–recapture studies.

### 4.3.3. *Designs for assessing trend or status and trend*

Powerful and sophisticated statistical techniques are available to identify and test changes in population parameters, for example, tests for change in the mean. We also have techniques for identifying and quantifying trend at a single site or in a single parameter, for example, we can quantify trend in some chemical concentration at a particular sampling station or trend in the average value of several sampling stations, by fitting a regression model that includes a time-dependent term. However, the notions of regional change and especially regional trend are much less well understood. Duncan and Kalton (1987) identified several types of change that might be addressed by sampling a population over time. Although their discussion was oriented toward human populations, the following kinds of change that they identified are relevant for environmental populations:

*Gross change* is the change at the *site* between two time periods.

*Average change over several time periods* can refer to the rate of change or trend at a *site* (as opposed to regional change or trend).

*Individual instability* is a measure of the variance at a *site*, possibly corrected for trend.

The traditional statistical concept is of *change in a population parameter*, where the usual population parameter of interest is the mean. Change is described by sampling the population and estimating parameters at distinct points in time. The resulting estimates are then analyzed for change/trend, for example, with time series or regression methods, or tested for significant difference.

Duncan and Kalton also described *net change* at the aggregate level. They use the example of change in unemployment rate between two months; however, a more general concept is implicit in net change. One can view the unemployment rate as the mean value of a dichotomous population, coded 1 for unemployed and 0 for employed. From this viewpoint, net change is merely a change in a population parameter. However, defining net change as a change in the population distribution, for example, population cumulative distribution function, captures a more general concept of allowing elements of individual change to counterbalance one another. Thus, it is quite possible for individual elements of a population to change, yet for there to be no net change in the population. A related concept occurs in forestry, where change is sometimes broken down into components consisting of growth of existing trees, mortality, and in-growth of new trees. Each of these components of change could be positive, yet the age and size population distribution could be invariant.

Generally, the most precise information of change (trend) comes from sites that are revisited, whereas the most precise information of status comes from visiting more sites. A critical point in designing a survey with the dual objective of status and trend is the allocation of visits to new sites versus revisits, attempting to describe current status and to detect trends in a set of ecological indicators. Observing the same sites over time eliminates the between-site component of variation. If the sites maintain their identity through time, this can greatly increase the power of trend detection methods. For some environmental resources, this is clearly not an issue, for example, for forested sites. For others, for example, lakes or estuaries, there may be little advantage in returning to the same set of site coordinates. Moreover, even if a site retains its identity, there is potential impact of previous visits on the site stemming from both perturbation due to sampling activity at the site and differential management of the site. That impact, sometimes referred to as "time-in-sample bias" (Bailar, 1989), can be substantial. The gain in precision may be more than offset by loss of representativeness.

Skalski (1990) recommended the use of rotating panel designs for the dual objective of status and trends. These designs partition the total sample into several subsets or panels. Each panel is then revisited on a different schedule. Fuller (1999), McDonald (2003), and Chapter 5 of this volume provide details and nomenclature for a wide variety of panel designs. Stevens and Olsen (1999) show how to use GRTS to form interpenetrating temporal panels so that each panel is spatially balanced as well as the composite. The Oregon Department of Fish and Wildlife (ODFW) uses a panel design to monitor the size of Coho salmon populations on the Oregon Coast (Stevens, 2002). Coho spawn in fresh water, migrate to salt water to spend their adult lives, and then return to spawn in about a three-year cycle. The design ODFW uses is tied to the Coho life cycle. One panel is visited every year. There are three panels visited on a three-year cycle and nine panels visited on a nine-year cycle. Four panels are visited every year: the annual panel, one of the three-year panels, one of the nine-year panels, and one panel of new sites.

The power of panel designs to detect change or trend is addressed in Fuller (1999), Urquhart et al. (1993, 1998), and Urquhart and Kincaid (1999). Their insight is that some frequently visited sites are important (e.g., a small annual panel), and the revisit schedule should be tied to the level of change relative to background noise. Thus, to detect a small but persistent trend, a design with a long revisit cycle will be more powerful for the same level of effort than a design with frequent revisits.

## 5. Using ancillary information in design

Ancillary data can be used in both sampling and inference to improve the accuracy of estimates. In sampling, it can be used to stratify the sampling frame to assure representation of all types of units. It can also be used as a basis for oversampling certain types of units to improve estimates for subdomains of interest. This is true for environmental and other sampling applications. We focus on two applications of ancillary information, which are frequently discussed in environmental applications: additional dimensions to the sampling frame and the appropriate use of ranked set sampling (RSS).

Spatial strata in environmental sampling typically are defined in four dimensions: three geographical and one time dimensions. Sampling bays and streams requires defining the geographical sampling frame in three dimensions. This is also true of sampling land, whether hazardous waste sites or downstream from potential pollution sources. Air pollution monitoring also has three geographical dimensions. Sampling wildlife, on the other hand, is more likely to only have two geographic dimensions. But in all these examples (with the possible exception of nonmoving pollution in the ground), the population of interest, whether wildlife or pollution, is moving, introducing a temporal dimension to the stratification.

The sampling units comprising the sampling frame are usually not difficult to delineate, but ancillary data on these units are often sparse. This is particularly true across time and height/depth. Ancillary data are generally known for a few time periods, so its consistency across time is subject to doubt. Similarly, many historical measurements represent a vertical slice (whether water, land, or air) that has been composited before analyzing, providing limited information on variation across this dimension.

In some environmental applications, it can also be very difficult (or costly) to move data collection locations. For example, only a few monitoring wells are likely to be drilled downstream of a potential pollution source. Once these sampled locations are selected, they can monitor across time, but all the measurements will be at the same location with respect to the other three dimensions of the sampling frame. This physical clustering of the measurements can dramatically reduce the effective sample size and resulting precision of estimates.

As with other traditional sample allocations (see Chapters 1–3), one frequently oversamples rare domains of interest and parts of the sampling frame, where the outcome measure is thought to be more variable. The four-dimensional nature of environmental sampling can make this more difficult. Variability has to be considered across space and time. The basic principle that one should not oversample too heavily if it is not clear which units are to be oversampled, applies to environmental samples. The resulting design can be very inefficient if the units that are oversampled do not increase the frequency of the rare subdomains or have consistent measurements of the domain of interest.

Ranked set sampling is a particular type of two-phase (double) sampling (see Chapter 3). In RSS, a small number of units, $m$, are not measured but are simply ranked, and then the measurement is taken on one unit based on its rank. This is repeated for $m$ sets, each time selecting a different order statistic to be measured. To select a sample of $nm$ measurements, it is necessary to rank $nm^2$ units by taking $n$ cycles of $m$ sets. This method was introduced by McIntyre (1952) to estimate pasture yields but has received renewed interest in recent years (Patil et al., 1994; Takahasi and Wakimoto, 1968). The advantage of RSS is that it does not require you to know how to stratify the sampling frame in advance, nor do you have to take the initial less-costly ranking information on all first-phase units before beginning second-phase measurement. By being able to rank multiple units and measure one immediately, RSS is attractive to the field operations of environmental measurements.

As an example of the possible use of RSS, consider wanting to select a sample of stream riffles (where the water moves roughly across a series of rocks) to measure fish stocks. Stocks are quite possibly correlated with riffle size. Rather than just to take a random sample of one-third of riffles, it is preferable to walk a stream and rank each set of $m = 3$ riffles, selecting the largest of the first set to measure, the middle of the second set, and the smallest of the third set, then repeating this process. (It is possible to modify this balanced approach to oversample units of particular interest. (Patil, 2002b)) With remote sampling locations and a lack of stratifying ancillary information in advance, this process can provide increased precision.

Unfortunately, much of the research on RSS has compared it with SRS. As demonstrated by the earlier chapters of this volume, SRS is rarely appropriate and the correct comparison is against other complex sample designs that might be used. Mode et al. (2002) compared RSS with three other sampling designs: (1) SRS; (2) weighted double sampling with cut points; and (3) double sampling using ratio estimation. They showed that RSS is appropriate when inexpensive (and possibly qualitative) auxiliary data are available for ranking, for which little distributional knowledge exists. If the general distribution of the auxiliary data is known in advance, then determining which to sample by comparing the auxiliary information to the cut points can achieve improved precision. If the auxiliary data are known to be highly correlated and linearly related to the variable of interest, then ratio estimation is preferable.

In situations where the available covariates make RSS a reasonable data collection method, it is important to consider the cost implications (Mode et al., 1999). Ranked set sampling requires ranking $nm^2$ units in addition to sampling $nm$ of them. If ranking has minimal costs relative to measurement, RSS can be used. The relative cost of measuring a single unit compared to ranking can vary depending upon the application. Mode et al. provide examples of cost ratios of 5.3 for crude oil in contaminated sediment, 20 for estimating fish abundance, and 50 for detecting radiation. They found that depending on the shape of the distribution and the accuracy of the ranking, cost ratios exceeding 6–11 were sufficient for RSS to yield improvements for a fixed total cost.

## 6. Inference for probability-based design

The analysis of a probability survey is often called *design-based* because the validity of the population inference rests on the design rather than on an assumed statistical model. The randomness is explicitly included in the sample-selection process and forms the basis for estimating population characteristics. The key quantity in the estimation is the inclusion probability for a population unit, which is the probability that that unit is included in the sample. It must be positive for every unit. In the case of a continuum, the inclusion probability is defined by an inclusion density, usually denoted by $\pi(s)$. In contrast to a probability density, the inclusion density has units. For example, an inclusion density for a point sample from a map might have units of $(number\ of\ sample\ points)/km^2$. In the case of a finite population, the inclusion probability sums to the sample size; in the continuous case, the integral of the inclusion density over the target domain gives the sample size. The importance of the inclusion probability for a sample element is that its reciprocal is a measure of the portion of the population represented by that element. Thus, for example, in a SRS of size $n$ from a finite population with $N$ total elements, the inclusion probability for each sample element is $n/N$, and each sample element represents $N/n$ population elements. If a SRS of $n$ sites were selected in a wetland with area $A$ km$^2$, then the inclusion density would be $\pi(s) = n/A$ and each site would represent $A/n$ km$^2$ of wetland.

The basic analysis tool is the Horvitz–Thompson or $\pi$-weighted estimator (Horvitz and Thompson, 1952; Thompson, 2002). The continuous version of this estimator is given in Cordy (1993) or Stevens (1997). The concept of the $\pi$-weighted estimator is that estimates of totals are obtained by weighting individual observations with a weight inversely proportional to their inclusion probability.

Let $n$ be the number of sample plots, $z_i$ the response for the $i$th sample plot, and $\pi_i$ be the inclusion probability (or density) evaluated at $i$th sample point. Note that $z_i$ could be a numeric score (e.g., per cent forested land cover) or a binary classification, for example, $z_i = \begin{cases} 1, & \text{if } i\text{th plot in degraded condition} \\ 0, & \text{otherwise} \end{cases}$. The Horvitz–Thompson estimate of the total of $z$ is given by $\hat{z}_{\mathrm{T}} = \sum_i^n \frac{z_i}{\pi_i}$ and the estimate of the mean value by $\bar{z} = \frac{\hat{z}_{\mathrm{T}}}{A}$, where $A$, the population size, is the total area of the target population. These formulas are the same for both finite and infinite populations. Note that in the case of $z_i$ being a binary classification, $\bar{z}$ estimates the proportion of the resource in the condition class, for example, the proportion of the watershed in degraded condition.

An alternative estimator of the mean value uses the estimated population size $\hat{A} = \sum_1^n \frac{1}{\pi_i}$ as a divisor in place of $A$. In some circumstances, use of the estimated population size in place of a known population size can lead to a more precise estimate of the mean because of positive covariance between $\bar{z}$ and $\hat{A}$. If the size of the target population is not known, for example, the imperfect frame case described below, then the alternative estimator must be used. Also, if some plots were not accessible, say because access permission was not obtained, then an estimate of the average condition of the *accessible* wetlands is $\bar{z} = \frac{\hat{z}_{\mathrm{T}}}{\hat{A}}$, where both $\bar{z}$ and $\hat{A}$ are computed using only those sites for which a response was obtained. An alternative is to use a nonresponse adjustment to compensate for the nonaccessible locations.

A spatially balanced sample will normally be more precise than a SRS of the same size because its spatial balance capitalizes on the spatial structure of the response. However, because of the restricted randomization inherent in the spatial balance, variance estimation can be an issue. Technically, the variance depends on pairwise or joint inclusion probabilities (the probability that a pair of points are both included in the sample). The restricted randomization implicit in spatial balance makes some of those joint probabilities very small or zero. The joint probabilities appear in the denominator of the usual variance estimators, so the estimators are undefined if joint probabilities are zero and unstable if small. A commonly used approach is to ignore the spatial constraint in the design and apply the SRS variance estimator. The resulting estimator will almost always be biased high. Horvitz and Thompson (1952) derived an unbiased variance estimator to accompany their estimator of the total, but the joint inclusion probability appears as a divisor in the estimator, so it is unsuitable for spatially balanced designs.

Wolter (1985) identified eight one-dimensional variance estimators for one-dimensional systematic sampling. D'Orazio (2003) extended three of these to two-dimensional systematic sampling. A general purpose technique that provides reasonably good results is to apply a postselection spatial stratification with at least two points per stratum. The strata can be selected arbitrarily but the points in a stratum should be close together. The usual stratified sample variance estimator is then applied. Stevens and Olsen (2003) developed a variance estimator specifically for spatially constrained designs that is based on a similar concept. Instead of explicitly forming strata, a local variance is computed at each sample point. The local neighborhood of a point is defined as a region containing the point's four nearest neighbors and then expanded to satisfy a symmetry constraint (if $a$ is in the neighborhood of $b$, then $b$ must be in the neighborhood of $a$). The overall variance estimate is a weighted average of the local estimates.

## 7. Model-based optimal spatial designs

The development of statistical theory or methodology is often driven by the search for optimality, that is, to find a new procedure that is "best." Design optimality involves two choices: which estimator or predictor to use, and which population elements to select, or, in a spatial context, where to place design points. In a statistical context, the standard is usually some measure of closeness of the estimator or predictor to the population attribute. Thus, our working objective needs to combine an estimator and a criterion that can be optimized. This requires the specification of an optimality criterion, which in statistics is usually minimum variance. This is not the only possibility; minimax criteria are also used, where one tries to minimize the maximum unfavorable outcome, for example, minimize the maximum loss. In situations where bias is a major concern, minimum mean square error is often the criterion. Unless otherwise stated, optimal should be interpreted to mean minimum variance.

Statistical models may be used to describe the underlying environmental process that generates the response. The statistical models usually applied in this setting are models of a mean process, possibly depending on ancillary variables, plus models of a spatial random process. The mean process $\mu(s|X, \beta)$ may be a constant, a function of location $s$, or a function of location and ancillary variables $X$ with parameters $\beta$. The spatial random process $Z(s|\theta)$ with parameters $\theta$ is frequently taken to be *intrinsically stationary* so that $E[Z(s+h|\theta) - Z(s|\theta)] = 0$. The spatial covariance of $Z(s|\theta)$ is usually described by the *variogram* $2\gamma(h)$, where $2\gamma(h|\theta) = \text{Var}[Z(s+h) - Z(s)|\theta]$. The quantity $\gamma(h)$ is then called the semivariogram. In this discussion, we will consider a spatial random field given by $Y(s) = \mu(s|X, \beta) + Z(s|\theta)$, $s \in R$ for location $s$ and domain $R$. Frequently, the semivariogram is also assumed to be isotropic so that it depends only on distance and not direction, so that $\gamma(h|\theta) = \gamma(|h||\theta)$.

Some of the early insights on optimal design (Dalenius, 1961; Iachan, 1985; Matérn, 1986) were derived by assuming a known covariance, using the sample mean as an estimator, and by optimizing a variance rate, that is, a variance per unit area. This approach sidesteps the influence of a domain boundary. In practice, the presence of a boundary, especially an irregular boundary, influences the optimal site locations. The results were consistent in suggesting that a systematic sample was better that a stratified sample, which was in turn better than a SRS. Moreover, the compactness property of a triangular grid was also shown to lead to favorable designs.

For the random field model, the sample mean is not the optimal estimator of our working objective. For the case when $\mu(s)$ is an unknown constant or a linear combination of explanatory variables, the optimal (in the sense of minimum squared error loss) predicted value for a new location $s_0$ is given by the kriging or best linear unbiased prediction estimator $\hat{Y}(x_0) = \sum \lambda_i Y(x_i)$, where $\lambda_i$ are the kriging weights and are described in many textbooks on geostatistics such as Cressie (1993) or Schabenberger and Gotway (2005). The variance of the prediction at location $s_0$ is given by $\sigma^2(s_0|S, \gamma) = 2\sum_{i=1}^{n} \lambda_i \gamma(s_i - s_0) - \sum_{i}^{n} \sum_{j}^{n} \lambda_i \lambda_j \gamma(s_i - s_j)$. Note that the prediction variance depends on the location of the sample points and the semivariogram. There is no dependence on the actual values at those points.

To get an optimal design for our working objective, it makes sense to use the optimal estimator and to choose the sample $S$ to minimize the total prediction variance $V_{\text{T}}(S, \gamma) = \int_D \sigma^2(s|S, \gamma)\text{d}s$. In most cases, this integral is very difficult to work with. It is intractable analytically and must be dealt with numerically.

As an alternative, Yfantis et al. (1987) evaluated square, triangular, and hexagonal grids, assuming a known covariance. Their optimality criterion was to minimize the maximum mean square prediction error. Their conclusion was that a triangular grid was optimal. McBratney et al. (1981) reached a similar conclusion using the average prediction variance.

The concept that the optimum location of sampling points for prediction will be some sort of regular arrangement is well established. One approach to optimizing design is to maximize some measure of regularity of the point pattern of the sample locations. The underlying assumption is that a highly regular design will also be a low variance design. An algorithm for locating sample sites that has been used with known domain boundaries or the presence of existing points is spatial simulated annealing (SSA) (Di Zio et al., 2004; Lark, 2002; Stevens, 2006; Van Groenigen, 2000; Van Groenigen and Stein, 1998). Sample points are selected to optimize some criterion that reflects the study objective, for example kriging variance, or a measure of regularity of the resulting spatial point process. The SSA begins with a set of arbitrary locations, and cycles through the points, perturbing each one in turn. At each step, the optimality criterion is calculated. If the new configuration resulting from the perturbation is better than the prior optimum, it is retained as the new optimum configuration. If it is worse, it is retained with a probability that decreases with the number of cycles. The concept behind retaining the suboptimal configuration is to bump the iteration away from a local optimum. Letting the probability of (temporarily) accepting a suboptimal configuration decrease helps to ensure eventual convergence to the global optimum.

Another approach is to modify the criterion somewhat. For example, instead of attempting to optimize over all possible designs, limit the space of potential designs. One way of limiting the design space restricts attention to sequentially optimal designs. In this method, an initial design with $m$ points is chosen, arbitrarily or at random. Then $s_{\mathrm{m}+1}$ is chosen at an optimal location conditional on the locations of the previous points (Cressie et al., 1990). The process is then repeated until all $n$ points have been chosen. Another way to do this is to discretize by replacing the two-dimensional continuous domain with a finite point set, say with a regular grid that covers the domain. In principle, then, one can evaluate all possible designs and pick the optimal one. This has been tried by Di Zio et al. (2004) and Wiens (2005). Even then, the computational burden can be overwhelming unless the design space is severely limited. Other authors have used SSA in conjunction with discretization (Wiens, 2005; Zhu and Stein, 2006).

In most applications, the covariance structure will not be known and must be estimated. Some papers have considered optimal designs solely for estimating the covariance function without regard to prediction. Warwick and Myers (1987) develop a search algorithm for achieving particular distributions of point pair distances, by which they take sums of squares of discrepancies in the realized and desired distributions and select a point pattern with a minimum sum of squares. Müller and Zimmerman (1999) consider generalized least squares fit to the empirical variogram to estimate variogram parameters. They use the determinant of the information matrix as design criteria. They compare several techniques, including the Warwick and Meyers method (1987). Their results show that a more irregular design with some points placed close to each other is better for variogram estimation. Zhu and Stein (2005) use maximum likelihood to estimate covariance parameters. They use minimax and Bayesian criteria to select an optimal designs. The design space is restricted to a fine grid, and SSA is used to locate optimal designs.

The more realistic case where the objective is prediction and the covariance structure is unknown and must be estimated has been considered by several authors, who attempt to consider the impact of covariance parameter estimation on the prediction variance.

Zimmerman (2005) notes that the design objectives for efficient prediction assuming known dependence and efficient estimation of spatial dependence parameters are largely antithetical and often lead to very different optimal designs. Zimmerman introduces a hybrid design that emphasizes prediction but accounts for the uncertainty in the covariance parameters. His approach is to choose the design to minimize an approximation to the variance of the empirical kriging (empirical-BLUP) prediction error. Note that the empirical kriging/BLUP predictor involves evaluating the covariance matrix at the estimated $\hat{\theta}$ rather than the assumed known $\theta$. He makes some empirical comparisons between designs to optimize parameter estimation, prediction variance, and a hybrid design.

Zhu and Stein (2006) compare the designs for (1) prediction using covariance parameters estimated from an existing data set, (2) estimating covariance parameters, and (3) prediction with estimated parameters. They use SSA to locate optimal design configurations. Consistent with previous work, the optimal designs in case (1) are highly regular and approximately triangular grid structure, subject to perturbation because of irregular boundaries. For case (2), the optimal designs consisted of multiple clusters of points. Their case (3) gives a pattern that is mostly regular, with several clusters of closely spaced points.

Diggle and Lophaven (2006) described a Bayesian approach to spatial design that balances the design for parameter estimation with spatial prediction. The designs are efficient for spatial prediction and make an appropriate allowance for parameter uncertainty. They also compare the efficiency of designs based on a regular grid plus extra close pairs to a regular grid with in-filling. Ritter and Leecaster (2007) also evaluate several designs that combine regularly spaced points with clusters of points. They conclude that the clusters are valuable for estimating the semivariogram and offer several recommendations for a design.

## 8. Plot design issues

Environmental measurements are frequently taken as an average over a three- (or four-) dimensional space. Water, land, or air samples are collected from a small physical area rather than a point. This area is referred to as the physical support. Complications arise in inference from environmental samples when the analytic units do not match with the physical support of the samples or when units of different size (or composition) support are combined. These are referred to as change of support problems (Gotway Crawford and Young, 2006).

Combining units of different size does not effect mean estimation, but it can cause significant problems in estimating precision and correlation. This in turn effects estimation of significant differences and distributional percentiles. Cressie (1996) points out that if the physical units are positively autocorrelated, the collapsing of the units into larger physical support will have less effect on the variability of the mean than when this correlation is absent.

In general, the larger the physical support, the lesser the variability in the measurements. This averaging of smaller units into larger ones shrinks the variation among units. This can be vital in many environmental situations. Polluted areas are often defined as those exceeding a set level. The determination of whether or not a site is polluted can be completely determined by the size of the physical support used for collecting the sample, not the underlying amount of contaminant.

Although not an environmental application, Openshaw and Taylor (1979) provide an excellent example of how the size and shape of the physical support can determine the estimate. They examined the correlation between the percentage of Republican voters and elderly voters in Iowa counties. Depending on which groupings of counties were used as the physical support, the correlation varied from −0.99 to +0.99.

Another plot design issue that can have important implications for analysis is when the physical support overlaps analytic domain boundaries. This situation can arise, for example, when plots are based on watersheds or river reaches, but analyses are planned by political boundaries such as states or counties. When the sampled plots cross the analytic boundaries, it makes analyses very difficult, with the accuracy of the estimates a function of the model assumptions that have to be made to allocate the support across domains (see Chapter 31). To minimize this problem, it is important to try and identify key analytic domains before the sampling frame is determined. It is then possible to define sampling units as the intersection of logical geographic units and these planned domains.

Composite sampling (Patil, 2002a) is a tempting methodology for measurements that are much more expensive (or time consuming) to analyze than they are to collect. Common examples are sampling for pesticides in soil, air monitoring, and contaminants in fish. Composite sampling is a logical method if when the analyte is present, it is likely to be in large quantities. For example, when conducting exploratory measurements around a suspected hazardous waste dump site, it is reasonable to take multiple samples from around the site, composite them, and then do the chemical analyses. If the analyte was really dumped on this location, it is assumed that the diluting resulting from combining the different samples, some of which have high levels of the analyte and others having none, will still result in detectable levels. Once the presence has been identified, more careful, noncomposited sampling can be conducted. (Alternatively for fish, the initial field sample collects 10 fish: 5 are composited and 5 are archived. If the composite analysis raises a flag, then the individual fish is analyzed. This avoids the expense of multiple field visits. Sample storage is cheap compared with travel to a remote site.)

There are two dangers associated with composite sampling. First, if the detection limit for the analyte is high relative to the expected levels in relative hot spots, then a composited sample might be below the detection limit, even if a hot spot has been included. For example, if the detection limit is only one-quarter the concentration found in the hot spot and five or more physical locations are composited, it is possible for the composite analysis to be nondetectable, even when it included the high value.

Second, composite samples are very good at producing estimated mean values for the area being composited. Thus, it can be used to produce a daily average air pollution level or an average exposure from digging up soil on a site. Unfortunately, composite samples underestimate the variability about that average. That is, the variation in samples with physical support of size equal to that of individual samples will be much greater than the variation observed from the composites. The difference is proportional to the square root of the number of composited samples, so if sets of four individual samples are composited, the variability will be 50% that of the individual samples. This is particularly important if one is interested in measuring percentiles of a distribution far away from the mean. For example, if one is interested in estimating the 90th percentile of individual soil samples, this number will be quite a bit higher than the 90th percentile of composited samples. This is similar to the fact that the 90th percentile of hourly airborne (or waste water) emissions is much larger than the 90th percentile of daily emissions.

Also note that composite samples may not estimate the average that one is really interested in. Going back to the fish example, the result from compositing is a weighted mean of the fish that went into the composite, with the weight being the actual weight of the individuals. This is not necessarily the same as the average body burden of the five fishes. Unless the weight distribution in the sample matches the weight distribution of the target population, there could be enough discrepancy to be of concern.

## 9. Sources of error in environmental studies

There are a number of sources of error that, if not unique to environmental studies, are more commonly observed in environmental situations than with other types of data. One particularly problematic error source is that the only sources of data are frequently not located at the site of interest. Frequently, measurement requires installation of expensive equipment, which already exists in preset locations. Water pollution measurements are frequently taken near outfalls from industrial facilities, but the concern is the effect on drinking water in peoples' homes. Air pollution monitoring stations are often located near the manufacturing plants producing the pollution, but the concern is with pollution levels in the air breathed by people where they live, often far away from the pollution source. Unlike most other data collection situations, the physical sampling locations are presets based on decisions having nothing to do with optimal statistical sampling. It is left to analysts to model how the data observed in one set of locations migrates to the locations of interest. Madsen et al. (2007) develop a regression model utilizing spatial correlation where the predictor variables are observed at different locations than the response. Zhu et al. (2003) apply a Bayesian hierarchical model to relate incidence of asthma to traffic density data, where the response and the stressor are misaligned in both space and time. Mugglin and Carlin (1998) also use a hierarchical model to interpolate disease incidence counts using spatially misaligned covariates. This migration may be subject to air and water currents, seasonality, and a host of other complicating factors.

Even when the environmental study designer gets to identify their sample locations, they may not be able to gain access to the sites. Often sampled locations are identified from large-scale maps based on GIS or other methods. Although in theory it is possible to go to all such locations, it is not necessarily true in practice. The location might be in middle of river rapids, on a steep slope, or on private property where the landowner refuses to provide permission. While the latter is analogous to refusals in household surveys (Groves and Couper, 1998, also Chapter 9 in this volume), environmental data collection introduces additional situations in which it will be impossible to collect the data from the sampled location.

Seasonality affects many types of data collection and analysis. But in most environmental surveys, seasons affect the location being sampled; while in household surveys the person being measured is associated with a specific location. (Migrant populations are an exception to this generalization and in this situation are more like environmental samples than other surveys of people (Kalton, 2003).) Environmental surveys of living species have the extra source of variability due to the fact that many move locations across seasons. For example, the location of fish varies greatly by time of year. Surveys of salmon in rivers of the Northwestern United States can only be conducted in those locations during spawning season. At all other times of year, there will be no fish to measure. If a single species is being studied, it can be timed appropriately for the migration patterns of the species; but if a general survey is being conducted, the results for specific species may be largely dependent on the time of year at which data are collected.

The two just-discussed sources of error can interact to create additional difficulties. During certain seasons, it may be impossible to collect data from locations that are available other times of year. For example, in a survey of domestic well drinking water quality, it may be necessary to sample from the well pipes before treatments are added. This requires accessing the pipes outside of the home. It is impossible to collect this data in rural Alaska during the winter. The data must be collected during the rest of the year, even though the water is consumed all year long.

Measurement error can come from a great variety of sources, including data collection instruments, laboratories, staff collecting and/or measuring the data, inconsistent physical materials, and detection limits. In addition, measurement errors that are unbiased can result in bias for statistics of interest.

Data collection instruments may not measure accurately. This can result in both extra variability and bias. Frequent recalibration of instruments can reduce bias but with an increase in data measurement costs. For example, X-ray fluorescence (XRF) machines are used to measure lead content of painted surfaces. The XRF machines might regularly underestimate the amount of lead or it might inconsistently measure depending on the underlying substrate. If multiple machines are used at the same time by different data collectors, inconsistencies across machines can introduce more error. Machines in laboratories can be similar sources of measurement error. In addition, there may be practices at laboratories (e.g., cleanliness or data tracking) that can affect data quality as well. Again, if multiple laboratories are used to measure the same analyte, then measurement errors can be compounded.

As with other types of surveys, data collectors can be sources of both variability and bias. In environmental surveys requiring data collectors to use machines, the varying skills of the people using the machines can cause increased levels of error. This makes it very important to develop protocols for data collection that will minimize error, are easy to follow, and are easy to monitor for quality.

The need to collect physical samples from inconsistent physical material is a source of error unique to environmental surveys. For example, collecting samples from municipal dump sites to measure the presence of toxic materials requires developing procedures to assure a representative sample of materials from a combination of computer parts, lawn mowers, furniture, and miscellaneous waste.

Detection limits are another difficulty unique to environmental surveys. Detection limits are "defined as the lowest level of the measurand where the probability of a positive result is at least 95 percent" (Van der Voets, 2002, p. 504). Frequently, samples with values of the analyte that are not detected are assumed to be 0 or possibly one-half of the detection limit. A more sophisticated approach is to model the measured data and then distribute the nondetected values between 0 and the detection limit according to the model. Lambert et al. (1991) describe how it can be improper to assume that all values that are nondetected are really below the detection limit.

The fact that measurement error can cause bias in estimated regression coefficients is well known across all analytic situations (see Fuller, 1987; Stefanski and Buzas, 1995, and Chapter 12 of this volume.) In environmental analysis, the effect of measurement error is increased because the parameters of interest are frequently not means and totals but rather extreme percentiles or the percent of a distribution that falls above a preset limit. The estimated distribution that results from naive analysis of observations with measurement error is an estimate of the convolution of the true distribution of the parameter and the measurement error distribution. Central location may remain unchanged, but the tails of the convoluted distribution will be spread out relative to the true parameter distribution. In these situations, unbiased measurement error can result in biased estimates of these parameters of interest. Cook and Stefanski (1994) introduced the SIMEX estimator (for SIMulation_EXtrapolation) as a way to remove bias from estimates from estimators that may be complicated, nonlinear functions of the data. The simulation step of SIMEX consists of generating multiple sets of pseudodata by adding known levels of error to the observed data and by calculating the estimator for each set of pseudodata. For the extrapolation step, the set of estimates is then regressed against the known level of error contamination. The regression is then extrapolated back to a zero level of error contamination. Stefanski and Buzas (1996) show how to apply the SIMEX estimator to deconvolute the estimated distribution function for a finite population, and Stefanski and Cook (1995) explore the relationship between the SIMEX estimator and the jackknife method (Quenouille, 1956) for reducing bias in nonlinear estimators.

Many environmental measures are right skewed, such as the log-normal distribution shown in Fig. 3. If the goal is to estimate the percent of the distribution above a preset limit (L in Fig. 3), it is clear that a greater percentage of the data are just below L than just above it. Thus, unbiased random measurement error will cause more values that are

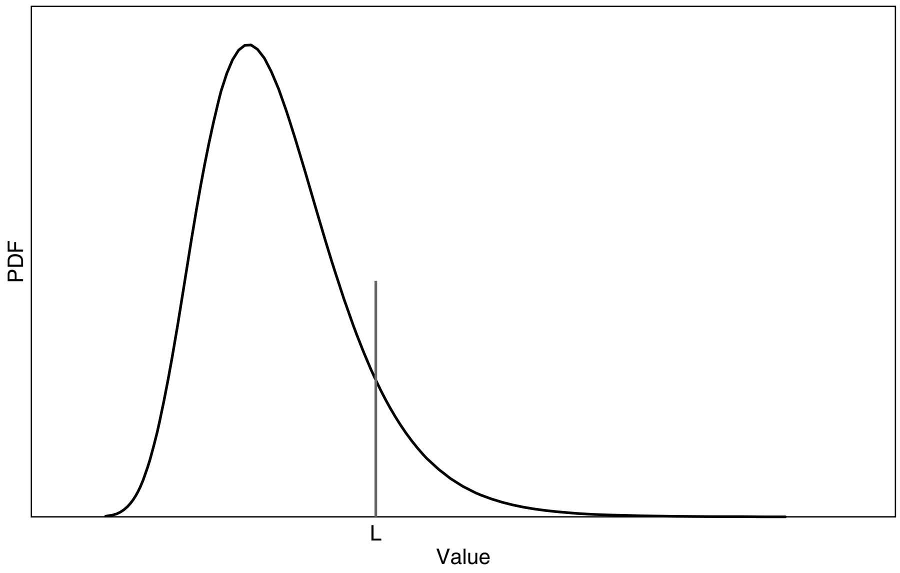


Fig. 3. Log-normal distribution with preset limit L.

truly below L to be measured above L than the reverse. Measurement error will therefore bias upwards the estimated percent above the limit L.

A second source of bias that interacts with this measurement error is that environmental measurements often are concerned not with the percent of measurements above the cutoff L, but with the percent of physical units that have any values above L. Examples include hazardous waste sites that are declared superfund sites if they have contamination levels above L anywhere on the site or homes that are considered to have a lead hazard if levels above L exist anywhere in the house. Comprehensive measurements of all locations are never taken, rather a sample of locations is measured and then a determination is made as to whether the site or home is contaminated. While the sample of measurements provides an unbiased estimate of the average level of contamination, they provide an underestimate of the highest level of contamination. (Similar sources of bias arise when trying to estimate life cycle exposure to contamination by only collecting physical samples at selected interview times during a longitudinal survey.)

An example of how these two sources of bias interact is provided in Clickner et al. (2002). As part of the National Survey of Lead and Allergens in Homes, a representative sample of homes in the United States was selected and measurements were taken to determine how many had lead hazards. One source of a lead hazard is having any floor dust with lead loadings of greater than 40 $\mu g/ft^2$. Samples from four rooms were collected and the maximum was computed. Figure 4 (Figure C.8 from Clickner et al.,

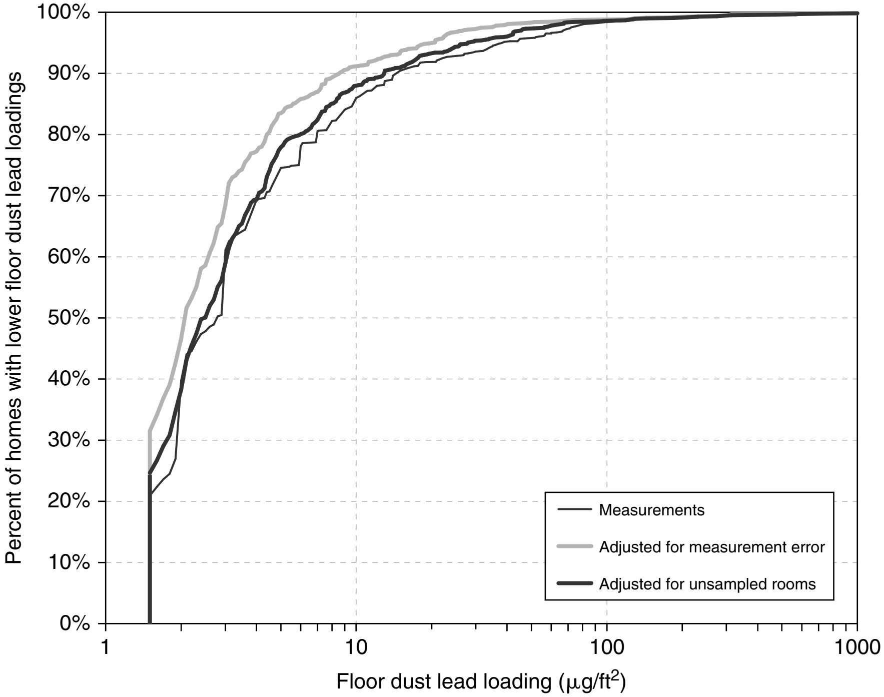


Fig. 4. Cumulative distribution of the maximum floor dust lead loading for Homes sampling and inference in environmental surveys.

2002) shows three curves. The thin black line shows the maximum of the measured dust levels. The gray line adjusts the maximum for the measurement error bias described above. This adjustment lowers the percent of homes that exceed the cutoff. The thick black line adjusts these maximums for the fact that only four rooms per house were measured. Since the maximum cannot go down if data from additional rooms were included, this adjustment increases the percent of homes exceeding the cutoff. Using the estimates after both adjustments yields an estimate of 4% of homes having floor dust lead loadings of 40 $\mu g/ft^2$ or more in one or more rooms. This is about 1% (one million) fewer homes than estimated using the actual measured floor dust measurements.

## 10. Conclusions

Sampling and inference in environmental surveys have much in common with other surveys. Thus much of what is contained in the earlier chapters of this book is relevant to environmental surveys as well. We have discussed some aspects of populations that need particular attention when designing and analyzing an environmental sample: focus on a broad population description, make use of the spatial context, use ancillary information, inadequate frames, difficult access to sampling locations, responses that are difficult and expensive to measure, evolving objectives, and the need to satisfy multiple objectives and stakeholders. We have outlined some methods that have been developed to address issues engendered by these aspects. This chapter is by no means a complete compendium, as there are other issues and methodology that we did not touch upon. The methods we did discuss are constantly being improved as we gain experience using them and software implementations become more readily available.

## Acknowledgements

The research described in this chapter was partially funded by the U.S. Environmental Protection Agency though cooperative agreement CR83-831682-01 to Oregon State University. It has not been subjected to the Agency's review and therefore does not necessarily reflect the views of the Agency and no official endorsement should be inferred. The conclusions and opinions are solely those of the authors. Mention of trade names or commercial products does not constitute endorsement or recommendation for use.

*Sample Surveys: Design, Methods and Applications, Vol. 29A*
ISSN: 0169-7161

DOI: 10.1016/S0169-7161(08)00020-5

20

# Survey Sampling Methods in Marketing Research: A Review of Telephone, Mall Intercept, Panel, and Web Surveys

*Raja Velu and Gurramkonda M. Naidu*

## 1. Introduction

Survey sampling methods play an important role in marketing research. The discipline of marketing itself draws its techniques from various social and physical sciences and any advances made in sampling methods in these areas almost always find an application in marketing research. Recognizing the importance of the topic, the first special issue (August 1977) of the *Journal of Marketing Research* was devoted to survey research. The articles in that special issue addressed three aspects of survey research, namely, sampling design, questionnaire preparation, and data collection. In an article that has appeared in an earlier volume, Velu and Naidu (1988) provided a survey of these aspects. Our objective there was to briefly review and update the aspect of sampling design with special focus on telephone, mall intercept, panel, and internet surveys. Because the design issues related to telephone sampling in particular the random digit dialing (RDD) methods are covered in a separate chapter 7 in this volume, we will focus on the other forms of surveys. A bibliography follows and, while not exhaustive, the listing of books and other references should provide a starting point for an iterative search. Although the coverage of topics is more relevant for United States, to the extent possible we provide information about the practices in other countries as well. To begin with, we have compiled a list of select institutions that actively engage in marketing survey research (Table 1).

Marketing researchers have been aware that subjective sampling procedures must be avoided in favor of probability methods of selection to make valid inferences about the target segments. Because of the inherent diversity of the marketing discipline, there has been a growing demand for all types of data necessitating more complex marketing surveys. Also, during the past three decades, the household (the nucleus of most consumer surveys) has undergone dramatic changes in terms of its composition and size. More women have joined the workforce, have become economically independent, and are making buying decisions. As the environment was changing, techniques for sample surveys were also changing. The high cost of personal household interviews has led to the development and use of more efficient sample designs and less expensive data

Table 1
A list of marketing research institutions

| URL | Institution | Marketing Research Services Offered |
|---|---|---|
| www.surveysampling.com | Survey Sampling International (SSI) | Leading provider of superior samples for mail, telephone, Internet, panel, B2B, B2C, surveys, RDD in 20 countries and internet panels in 17 countries. With global partners, and a reach of over 50 countries. Some 1500 organizations and 43 of the top 50 research organizations use their services. |
| www.createsurvey.com | Create Survey | Create Survey is a Web-based survey software that lets you build and run online surveys in the Internet. You may start using it right now or read more below. |
| www.surveysystem.com | The Survey System | The Survey System is the most complete software package available for working with telephone, online, and printed questionnaires. It handles all phases of survey projects, from creating questionnaires through data entry. The Survey System was designed specifically for questionnaires; so their software saves you time. |
| www.greenfield.com | Greenfield Online | Greenfield Online helps marketing research companies and consultancies connect with their consumer insights by programming and executing online surveys using their Internet-based online panel of prerecruited respondents. They couple access to survey respondents with executional excellence and quality. |
| http://us.lightspeedpanel.com | LightSpeedConsumer Panel | By taking the time to participate in Lightspeed Panel surveys, you have the power to let companies know exactly what you think. This helps you, the consumer, to develop and improve the products and services offered. |
| www.tnsglobal.com | TNS in North America | TNS is a market information group. It is the world's largest custom research company and a leading provider of social and political polling. It is a major supplier of consumer panel, TV audience measurement, and media intelligence services. It provides market information and measurement, together with insights and analysis, to local and multinational organizations. |
| www.e-focusgroups.com | e-focus Groups | e-FocusGroups offers solutions for all market research needs. It brings the benefit of more than 20 years of market research experience in a wide variety of industries, including consumer products, advertizing, pharmaceuticals, e-commerce, computer hardware, computer software, telecommunications, and banking, among others. |

(*Continued*)

Table 1
(*continued*)

| URL | Institution | Marketing Research Services Offered |
|---|---|---|
| www.forrester.com | Forrester | Forrester Research, Inc. is an independent technology and market research company that provides pragmatic and forward-thinking advice to global leaders in business and technology. For more than 23 years, Forrester has been making leaders successful every day through its proprietary research, consulting, events, and peer-to-peer executive programs |
| www.zoomerang.com | Zoomerang | Zoomerang pioneered online survey software in 1999 to give organizations like yours a powerful self-service alternative to conduct accurate comprehensive surveys with a minimum of cost and effort. Today, Zoomerang is the world's No. 1 source of online surveys, helping thousands of organizations in more than 100 countries. |
| www.web-surveyor.com | Web Surveyor | They are empowering people to make informed business decisions using their online data collection solutions. It provides online survey services that enable their customers to easily collect real-time feedback to drive their businesses. They ensure data security and confidentiality, a reliable survey hosting service, dependable survey software, and a responsive team of survey experts. |
| www.web-online-surveys.com | Web Online Surveys | This is an all in one service designed for people who are not computer experts and have the need to conduct surveys by themselves. |
| www.synovate.com | Synovate–Research Reinvented | Synovate is the world's most curious company. Their job is to learn what people like, and why they like the things they like. That knowledge helps product designers and manufacturers give people what they want. The work they do at Synovate is continuously stretching the definitions of conventional research. They operate across six continents, in 50 countries. |
| www.gartner.com | Gartner | They deliver the technology-related insight necessary to make the right decisions, every day. Gartner serves 10,000 organizations, including chief information officers and other senior IT executives in corporations and government agencies, as well as technology companies and the investment community. |
| www.vnu.com | Nielsen | The Nielsen Company is a global information and media company with leading market positions and recognized brands in marketing information, media information, business publications, and trade shows. The privately held company is active in more than 100 countries, with headquarters in Haarlem, the Netherlands, and New York, United States. |

(*Continued*)

Table 1
(*continued*)

| URL | Institution | Marketing Research Services Offered |
|---|---|---|
| www.imshealth.com | IMS Intelligence Applied | IMS is the one global source for pharmaceutical market intelligence, providing critical information, analysis, and services that drive decisions and shape strategies. |
| www.kantargroup.com | Kantar | Kantar is one of the world's largest research, insight and consultancy networks. They help clients to make better business decisions through a deeper understanding of their markets, their brands, and their customers. They help clients find *better ways to answer business questions*. |
| www.harrisinteractive.com | Harris Interactive | In an increasingly chaotic and competitive world, Harris Interactive can provide clarity and confidence. They believe that market research helps our clients understand the drivers of decision making and can strengthen enterprise equity. Providing clients with this accurate knowledge will help them achieve measurable and enduring performance improvements. |
| www.jdpower.com | J.D. Power Consumer Center | Since 1968, J.D. Power and Associates has been conducting quality and customer satisfaction research based on survey responses from millions of consumers worldwide. It has developed and maintains one of the largest, most comprehensive historical customer satisfaction databases for various products and services. |
| www.opinionresearch.com | Opininon Research Corporation | At Opinion Research Corporation, they provide objective, fact-based decision support, they earn their confidence with our fresh ideas and perspectives, grounded in rigorous research methods and business savvy. |
| www.dentsuresearch.co.jp | Dentsu Research On-line | Dentsu Research, a specialist in market research, has served as the eyes and the ears of team Dentsu, collecting and analyzing the latest in consumer information. Now, over 30 years later, marketing research remains the core of their work, providing any and all services clients require. |
| www.infores.com | IRI—Information Resources Inc. | Driving the transformation of the consumer packaged goods (CPG), retail, and healthcare industries, only IRI provides a unique combination of real-time market content, advanced analytics, enterprise performance management software, and professional services. |
| www.npd.com | NPD Group | The NPD Group, founded in 1967, is the leading global provider of consumer and retail market research information for a wide range of industries. They provide critical consumer behavior and point-of-sale information and industry expertise across more industries than any other market research company. |

*Note*: There are several other vendors and due to space limitations, they are not listed here.

collections methods such as the use of telephone and mall intercept interviews and the use of Web surveys. At the same time, the public has become increasingly concerned about invasion of privacy and the maintenance of confidentiality of the information obtained.

Some developments (see Frankel and Frankel, 1977) of interest to marketing researchers include: (i) techniques related to the manipulation of sampling frames, (ii) techniques related to respondent selection, (iii) methods for minimizing the total survey error, and (iv) improving the quality of nonprobability sampling. Broadly we organize the discussion of various forms of surveys around these areas. We shall focus briefly in Section 2 on sample frames and procedures for telephone household surveys, a topic that received a great deal of attention over the last several decades. Judgment or nonprobability sampling procedures, still viable in marketing research, are convenient to carry out and are less expensive compared to other methods. In Section 3 we shall comment on mall intercept surveys, which are used increasingly by several marketing research firms. The consumer panel studies are reviewed in Section 4. With the advent of the Internet in the mid 1990s, the Web survey has become quite popular because of its ease of implementation as well as its cheaper cost. We briefly review this area in Section 5.

## 2. Telephone surveys

The telephone is an important tool for the collection of marketing survey data in the United States. Although it has been used in the past mainly for short follow-up interviews, usually for clarifying the information provided in personal or mail interviews, marketing researchers had resorted to using the telephone due to the increasing cost of other forms of surveys. A distinct advantage of the method is accessibility to the respondent. Some major disadvantages are the limited time a respondent may want to spend with a physically absent interviewer and the inability of the respondent to actually "see" the product in question as in surveys where the interviewer can display the product and obtain observational data.

In the United States telephone numbers have three parts: a three-digit area code, a three-digit central office code or prefix, followed by a four-digit suffix. The list of all area code–central office code combinations currently in service can be obtained from the telephone companies. With the introduction of mobile phones, these combinations have exponentially increased in recent times. Numbers to exclude from such a list are those of (i) the telephone company central offices (such as 555 used for directory assistance) and (ii) other central offices used solely by government or businesses (such as 866). We shall refer to groups of consecutive numbers starting with 0, 00, or 000 within the suffix as "banks of numbers." For the operational convenience of the exchange, only certain banks of numbers are assigned to users.

### *2.1. Sampling frames for telephone households*

The sampling unit for most marketing investigations has been primarily a household and it is implicitly assumed when telephone sampling is used that a single telephone serves a single household. This is not necessarily the case in practice. Some households have more than one telephone and more than one number. With the call forward option, business calls are sometimes automatically transferred to home phones. It is estimated that more than 94% of the households in the United States can be contacted either via land line or via mobile phone (see Tucker et al., 2007).

Households with mobile telephones are different from households with land telephones, as shown by several demographic and economic variables. These demographic and economic differences are expected to manifest in attitudinal differences as well. We will comment on sampling issues related to mobile phones later in the chapter. What we describe below mainly applies to land lines only.

There are basically two kinds of sampling frames used for telephone surveys. The rest are minor variants of these two frames. One is the list-assisted frame. The list can come from the telephone directories or from previous surveys. The other is the set of all possible four-digit suffixes within the existing central office codes. The latter is used in RDD methods. There are some advantages and disadvantages in both frames. The most important drawback of this frame, however, is that it excludes working telephone numbers that are not listed in the directory. Also, telephone directories are outdated, on the average, by at least 7–8 months. The percentage of unlisted numbers varies by regions with roughly 30% of numbers in large metropolitan areas of the United States unlisted. Households with unlisted telephone numbers tend to differ from households with listed telephone numbers on key demographic characteristics (see Moberg, 1982). Brunner and Brunner (1971) found significant differences between the two groups on certain product ownership, usage, and purchase patterns.

The disadvantage of the second frame is the large number of nonworking telephone numbers that may be sampled with unrestricted random sampling. In the United States only a fraction of dialings will connect with a usable residential household. The effort to identify these numbers adds considerably to the cost of a survey. Waksberg (1978) reports that this spade work is done by marketing research firms, and the more "useful" sampling frames are developed by these firms at considerable expense and are not available to the general public. Most researchers cannot afford to duplicate such a costly task. It is important to narrow the frame used for RDD. The designs to be discussed in what follows are expected to reduce the proportion of unused numbers sharply.

To emphasize the inherent differences between the two frames and their variants, it is useful to mention the problems in determining the status of a given number. Dialing a working number can result in (i) a completed call, (ii) unanswered rings, (iii) a busy signal, or (iv) wrong or no connection because of misdialing or technical problems. Unless the call results in a contact, it is impossible to determine whether the number belongs to a household or a nonhousehold. In RDD sampling, a nonworking number is not always easily determined. Dialing such a number can result in (i) a recorded message stating that the call cannot be completed as dialed, (ii) no connection, (iii) unanswered rings, or (iv) connection with a number other than that was dialed. The last possibility introduces biases in RDD sampling, because the telephone system equipment is not normally designed to receive a nonworking number. Note that the households reached in this manner have a greater probability of inclusion in the sample.

### 2.2. *Telephone sample designs*

Telephone sample designs can be broadly divided into list-assisted and RDD methods. We shall briefly discuss these designs and finally discuss the concept of dual frame designs.

### 2.2.1. List-assisted methods

*2.2.1.1. Direct selection from directory.* This is the most basic of all the directory assisted methods. A sample of directory lines is selected using either systematic or simple random sampling. One could also use cluster sampling for easy execution. The cluster consists of a randomly selected line and the next $k$ lines. To avoid actually counting lines, directory column inches can be used. This method yields an equal probability sample of all listed numbers with a minimal percentage of wasted dialing due to non-working numbers. A disadvantage of the cluster sampling method is that names listed together in a directory might belong to the same community, religion, etc., and if they are homogeneous with respect to the variables being estimated, the design is inefficient as compared to simple random sampling of lines. The major disadvantage of the directory method is that it does not give any chance for unlisted working telephone numbers to appear in the sample. The bias may be significant in certain surveys and the following procedures are proposed to correct partially for the bias.

*2.2.1.2. Addition of a constant to a listed number.* A number is randomly selected from the directory and an integer, either fixed or randomized (between 0 and 9), is added to the directory number. This gives a chance for inclusion of possibly unlisted numbers in the sample. Some variants of the above mentioned procedure involve randomization of the last $r$ (2, 3, or 4) digits or a directory number. Two drawbacks of these procedures are as follows: (i) when $r$ increases, the number of wasted dial rings will increase, and (ii) all telephone numbers do not have an equal chance of inclusion, because the probability of selection of a number would be proportional to the number of directory listed numbers in the same $r$th bank. If the numbers are not in the directory, they automatically eliminate the possibility that numbers which follow them will be in the sample. A method suggested by Sudman (1973) to correct for (ii) is described in the following section.

*2.2.1.3. Sudman's method.* A random sample of listed numbers is selected and the last (usually $r = 3$) digits are ignored. This results in banks of numbers selected with probability proportionate to the number of listed numbers in the bank. Calls are made using RDD within the bank until a predetermined number of households with *listed* numbers have been reached. The predetermined number is fixed so that the resulting sample is self-weighting. If we let $N$ = total number of household telephones, $N_{\mathrm{L}}$ = number of telephones among $N$ that are listed, $n$ = sample size, $m$ = number of selected banks of working numbers, and $N_{\mathrm{L}_i}$ = number of listed telephones in the $i$th bank, then

$$\text{probability of inclusionof a number in the sample} = \left(N_{\mathrm{L}_i}\frac{m}{N_{\mathrm{L}}}\right)\left(\frac{N_{\mathrm{L}n}}{NmN_{\mathrm{L}_i}}\right) = \frac{n}{N}. \tag{1}$$

Remark. This probability is exact (and the sample is self-weighting) only if (a) the proportion of listed households numbers in the $i$th bank is equal to the overall proportion ($N_{\mathrm{L}}/_{N}$) of listed household numbers and the predetermined number of sampled listed households in a bank is $n/m$, or if (b) the predetermined number of sampled listed households in a bank is fixed as $nN_iN_{\mathrm{L}}/NmN_{\mathrm{L}_i}$ where $N_i$ is the number of household telephones in the $i$th bank.

The first bracketed term indicates the probability of inclusion of bank $i$ in the sample and the second term, that of selecting a number within the bank. The procedure is unbiased and self-weighting. As Waksberg (1978) points out, this method also has several problems. Ascertaining whether a number dialed is listed or not can be difficult. For example, in a national survey, the procedure requires the use of a large number of telephone directories. Finally, because the numbers are clustered, a large proportion of them may occur in relatively empty banks, resulting in unequal numbers of households per cluster.

### *2.2.2. Random digit dialing methods*

These methods are used to obtain equal probability samples of all telephone numbers both listed and unlisted. As mentioned earlier, an unrestricted application of the procedure will lead to the inefficient use of survey resources. Therefore, it is important to narrow the sampling frame by eliminating nonworking numbers. If information on nonworking numbers is available (e.g., which banks are not assigned), random digits within these banks could be excluded from the sample. Some telephone companies will provide information about working banks. However, this information is usually not available, forcing researchers to use directories to determine working banks. Typically those banks with less than three listed phone numbers are eliminated. The incidence of telephone households in the sample can be increased by eliminating the business telephones listed in the yellow pages of the telephone directory. It is evident that all these efforts require a considerable investment of time and, unless the frame is used repeatedly, the cost may be prohibitive for a small survey.

*2.2.2.1. Waksberg–Mitofsky design.* The (RDD) selection procedure proposed by Waksberg (1978) is as follows. Obtain from the telephone companies all area code–central office code combinations currently in service. Append all possible two digits and treat the resulting eight-digit numbers as primary sampling units (PSU). Randomly select a PSU and the next two digits. If the 10-digit number is for a residential address, the PSU is retained in the sample and if not, it is rejected. If retained, additional pairs of random numbers to identify the two last digits are selected within the same PSU and dialed until a set number of residential telephones are reached. This process is repeated until a predetermined number of PSUs are chosen. This design produces an equal probability sample of working telephone numbers. The procedure of selecting PSUs is similar to Lahiri's (1951) selection procedure for probability proportionate to size (pps), although the latter requires a prior estimate of cluster size. This procedure which selects PSUs with probability proportional to working numbers differs from Sudman's method which selects PSUs proportional to listed working numbers. The stopping rule for the Waksberg–Mitofsky design also refers to working numbers and is not restricted to listed numbers. It is important to note that this procedure uses a cluster size of 100, a practical advantage over a cluster of 1000.

A crucial problem in this procedure is the large value of the proportion of PSUs with no residential numbers. Because all possible choices of two-digit numbers are appended to area code–central office code combinations to arrive at the PSU, it is possible that a large number of PSUs may not contain any residential numbers. It is important to obtain an estimate of the proportion of PSUs with no residential numbers. This can be expected to be smaller for urban than rural areas. An estimate based on a national U.S. study is given by Groves (1978) as 0.65.

*2.2.2.2. Stratified element sample.* An alternative design is discussed in Groves (1978). The procedure initially groups together all central office codes in the same exchange and then groups together exchanges in the same area code. Size categories of the exchanges are then formed based on the number of central office codes in an exchange with the number of central office codes acting as a proxy to population density. Within each size category, exchanges are ordered geographically within an area code and similarly area codes are then ordered geographically. Given this ordering of the frame, a systematic sample of central office codes is drawn. A four-digit random number is generated and appended to a selected central office code, yielding a 10-digit sample telephone number. Groves (1978) observes that only about one-fifth of the numbers were confirmed a working household numbers, whereas in Waksberg's design a roughly threefold increase in identifying working household number is possible. The main attraction for using this design would be when there is a greater homogeneity among the prefixes. This design can be treated as a simple random sample when the stratification introduced based on the exchange size is rather weak.

*2.2.2.3. Dual frame sample design.* The two-stage cluster design, proposed by Waksberg (1978), is better than directory-based designs in terms of coverage rates and over stratified element sampling in terms of cost. However, the design requires a new selection from the same PSU for each nonworking number encountered, and thus adds to the cost of screening numbers to identify residences. It is difficult to distinguish nonworking numbers from unanswered residential numbers. Another problem is the low-response rates for telephone surveys attempted without prior contact. It is found that persons with listed numbers are more likely to cooperate than those with unlisted numbers. Groves and Lepkowski (1986) consider dual frame designs as proposed by Hartley (1962) to be useful when the target segment forms a majority of elements in one incomplete list frame (directory listings) but a minority in another complete frame (RDD generated numbers). The poststratified estimator suggested by Casady et al. (1981), which mixes the estimates from each of the two frames, is investigated by Groves and Lepkowski (1986) and Lepkowski and Groves (1986). If we let $p$ denote the proportion of the unlisted telephone population and $\theta$ denote a mixing parameter, the estimator of the mean is

$$\overline{y} = p\overline{y}_{\mathrm{UL,\,RDD}} + (1-p)\left[\theta\overline{y}_{\mathrm{L,\,RDD}} + (1-\theta)\,\overline{y}_{\mathrm{L,\,DL}}\right] \tag{2}$$

where $\overline{y}_{\mathrm{UL,\,RDD}}$ is the estimate for the unlisted population chosen by RDD, $\overline{y}_{\mathrm{L,\,RDD}}$ is the estimate of the listed population chosen by RDD, and $\overline{y}_{\mathrm{L,\,DL}}$ is the estimate of the listed population chosen from the directory frames cases. The cost advantage of the dual frame derives from the list frame in identifying the working numbers. Several survey research firms (see Table 1) maintain a computerized data bank of all published directories and in one test for the state of Michigan, Groves and Lepkowski (1986) report 88% of numbers on the list were found to be working numbers as compared to 59% for the selection of samples within the PSU in RDD design. From the form of the poststratified estimator, it can be seen that the crucial parameters are $p$ and $\theta$ which depend on the geographical region and the type of marketing research investigation. It is estimated that roughly 64% of contacted RDD sample households are in directory listings, but the proportion of RDD numbers not contacted but found in the listing is around 66%. At the national level it is not known what proportion of these noncontacted numbers are working residential numbers. This may be influenced by large metropolitan areas where a low rate of list frame coverage is known to exist. Thus, the dual frame design can result in increased coverage (than list frame) and also increased precision (than the cluster RDD) by following simple random/stratified element designs on the list frame, thereby avoiding homogeneity due to clustering. To evaluate the dual design more thoroughly, the marketing investigator must know several cost elements and the relative nonresponse bias. The nonresponse bias is typically measure by the difference between two group means, where only one group receives an advance letter. Based on a simulation study for the U.S. National Crime Survey, Groves and Lepkowski (1986) suggest optimal allocations between 35% and 80% to the list frame.

*2.2.2.4. List-assisted RDD methods.* The operational difficulties involved in implementing the Mitofsky–Waksberg method has led to increased use of list-assisted sample designs. Two main issues with the Mitofsky–Waksberg method were in replacing the nonresidential numbers and in variances being larger than a simple random sample or stratified random sample of the same size. The properties of the list-assisted methods were examined in detail by Casady and Lepkowski (1993). But the underlying structure of the telephone system has changed greatly since then. More area codes are now being assigned and there is a gradual decrease in the proportion of numbers that appear in directories. Thus, it has become increasingly difficult to identify the residential numbers. Tucker et al. (2002) evaluate relative efficiencies of list-assisted and Mitofsky–Waksberg designs and conclude that the relative gain in precision from list-assisted design has increased in the past decade.

### *2.3. Respondent selection in telephone surveys.*

There are a number of other issues to be addressed in telephone surveys. Some households have more than one telephone number, making it necessary to obtain this information during the interview so that appropriate estimation weights could be constructed. In any telephone survey, ambiguities exist about no answers, uncertain rings, busy signals, etc. Any stopping rule for classifying these is bound to introduce some bias in sample selection. A more serious problem from a marketing researcher's point of view is that the person answering the telephone is not necessarily the same person who makes the purchase decisions. As shown in the literature on consumer behavior, buying decisions result from an interaction of all family members. To retain the characteristics of a probability sample, the person to be interviewed should be selected at random. We discuss a few approaches to the problem in the following.

A selection procedure suggested by Kish (1967) in the context of area probability samples requires all eligible respondents within a household to be listed by sex and by age within sex categories. The interviewer then selects one respondent using a random number table (see Kish, 1976, Section 11). This procedure is difficult to use in telephone surveys where most refusals to participate occur at the beginning of the interview. The procedure is time-consuming and could present problems establishing rapport. For example, asking for the number of adult males in residence could be perceived as insensitive to single women living alone. Because rapport with the respondent is so vital to telephone surveys, Troldahl and Cater (1964) adapted the Kish format but based the selection on only two easy-to-answer questions: (i) How many adults live in your household, counting yourself, and (ii) how many of them are men? Using four selection matrices rotated randomly over the sample, a respondent is selected. This procedure does not significantly reduce refusals when compared to the Kish strategy (see Frey, 1983, p. 80). Bryant (1975) suggested dropping one of the four matrices every second time it appears in the rotation. This would result in the selection of more male respondents and the procedure takes into account increases in one-person households and households headed by women. It must be noted that these alternative strategies assign unequal probabilities of selection to some eligible respondents such as middle-aged adults. Another variation used by Groves and Kahn (1979) is to modify (ii) "how many of them are women?" A recent investigation by Czaja et al. (1982) reveals no major differences in cooperation rates and demographic characteristics across the three models.

Two procedures reported recently seem to be effective in terms of operational use and eliciting higher response rates. Basically, these two avoid asking household composition questions before beginning the interview. The first procedure is suggested by Hagan and Collier (1983). The designated respondent is predetermined to be one of four possibilities: oldest man, youngest man, oldest woman, or youngest woman. After the initial introduction, interviewers simply ask for the designated respondent (randomly chosen and printed on the interview form a priori) and when a respondent of that designation does not live in the household, the opposite sex is interviewed. In single-person households, the age designation is irrelevant. Based on a national study, the authors suggest that this procedure is an improvement in terms of lower refusal rate. The second procedure given by O'Roourke and Blair (1983) selects the adult who had the "most recent birthday." This is a probability selection method and ascertaining the birthday is considerably easy. Comparing this with Kish's procedure, based on a survey, the authors found the major difference in refusal rate occurred at the preselection stage. Once the respondents agreed to participate, it did not matter which procedure was used to continue the interview.

Rizzo et al. (2004) provide a less intrusive method for selection of within-household members. It uses the fact that about 85% of households in the United States have less than two adults. Thus, this method randomly selects either the screener respondent or the other adult. Other than gathering information on the number of adults in the family, the procedure does not call for any information. The procedure operates as follows. Let $N$ be the number of adults: if $N = 1$, the respondent is selected; if $N > 1$, randomly sample the respondent with probability equal to $1/N$. If $N > 2$ and if the screener respondent is not selected, then use the Kish method. This is a probability sampling method and does not result in self-selection biases.

### *2.4. Randomized response techniques in telephone sampling*

The randomized response technique originally introduced by Warner (1965) to obtain the estimates of behavior that is usually underreported and is found to be useful for personal interviews. A randomizing device is used to choose a statement and the respondent is asked to provide a response to the one selected. The interviewer is neither shown the outcome of the device nor is informed of which statement is answered. The most difficult aspect of a telephone application of the randomized response technique for sensitive questions is the provision of a randomization device. As Stern and Steinhorst (1984) observe, there are two main problems: (i) the device is not readily available to many respondents, and (ii) the complexity of instructions necessary to provide a satisfactory distribution may inhibit respondent cooperation. Also, suggestions from a "faceless" voice to flip a coin may be regarded as foolish by some respondents. However, an advantage of using a respondent-supplied randomizer is that it eliminates the respondent's suspicions that the interviewer has "fixed" the randomizer. A potential disadvantage is that it does not provide a known probability distribution. The technique continues to be used widely in social sciences. See Van der Heijden et al. (2000).

There are several randomizers suggested in the literature including credit card number, street address, occurrence of events, etc. (see Orwin and Boruch, 1982). The one that is tested on a limited basis is the last digit of randomly selected telephone numbers. This provides a known distribution for both the selection of sensitive and nonsensitive questions and the generation of surrogate answers (see Stern and Steinhorst, 1984). Although this method is considered to be successful on the issue of response privacy, the nonresponse is still high. This method also requires both the interviewer and respondent to have access to the same telephone directory. Each geographical area served by a different telephone exchange and telephone directory would be sampled as a separate stratum. At a national level, this may create some operational problems. Other randomizers such as the last digit of street address are supposed to overcome this problem, but in the absence of a known distribution of the last digits, they are not statistically attractive to use.

### *2.5. Locating a special population using RDD*

In many instances, the researcher may be interested in locating a subclass of the total population. Blair and Czaja (1982) show how Waksberg's two-stage cluster design can be modified, if it is known that this special population clusters geographically. This modification takes advantage of the fact that the telephone central office codes are assigned to well-defined geographic locations. It works as follows: select a simple random sample from all possible telephone numbers. These numbers are then called and only those working residential numbers of a household with the appropriate special characteristics are retained. The first eight digits of each retained number are then defined as a PSU. Using each retained telephone number as a random start in the PSU it created, numbers are then sequentially generated and screened. This procedure is continued until a certain cluster size is identified.

As Waksberg (1983) notes, this procedure has some serious statistical implications in which many situations may reduce the efficiency. But in the case of special populations, PSUs could exist in which it is not possible to reach the predetermined cluster size even if the 100 numbers are used. The special population households associated with clusters that are smaller than the specified cluster size have a lower probability of selection than the rest of the special population. Hence, to produce an unbiased estimate for the total population, we must adjust for unequal probabilities which increase the sample variance (see Kish, 1967, p. 430).

### *2.6. Ring policy in telephone surveys*

Each telephone call is composed of 2-second rings followed by 5 seconds of silence. Survey research firms on the average allow six rings per call, thus the amount of time taken to reach a potential respondent is on the average 37 seconds. Smead and Wilcox (1980) questioned how long the phone should be allowed to ring based on a telephone survey using the members of a major university consumer panel. Ten rings and three callbacks were used. The average answer time for the 219 respondents was 8.7 seconds with a standard deviation of 6.3 seconds. The answer times followed a gamma distribution and suggested that only four rings (or 23 seconds) were necessary to reach 97%.

### 2.7. *Telephone sampling: other uses*

The use of telephone interviewing is widespread because of its major cost advantages. However, there are still many situations that require face-to-face interviewing, particularly those that deal with special subgroups of the general population. This involves screening, and the rarer the group the more costly the screening. However, in general, telephone screening costs are lower than face-to-face screening. Sudman (1978), based on a realistic cost model, has shown that telephone screening will be an optimum procedure unless (i) the degree of homogeneity is small, (ii) the density of interviews is low, and (iii) locating and screening costs are small relative to interviewing costs. From the discussion in Section 1, it follows that (iii) could be an important consideration in using RDD. However, directory-based telephone screening might be cost effective.

Many survey research firms have databases constructed from the telephone directories supplements with auto registration data. These are useful for mail samples. Information collected from other sources such as census records are sorted by area code and telephone exchange that provides a faster way to reach a target population such as low income families, Hispanic groups, etc. The yellow page listings are used for business samples, because the directory category headings are broad and easy to use by marketing researchers.

Computer-assisted telephone interviewing (CATI) was used first by market research agencies in the private sector. The concept was proposed by the American Telephone and Telegraph Company to measure customer evaluation of telephone services. CATI is now very popular in other types of organizations as well. Interview responses are quickly processed and by accumulating counts of key respondent characteristics while interviewing, quota targets, that is, desired sample sizes in strata in RDD sampling, can be tracked. Adding visual monitoring to telephones from supervisory terminals, CATI provides efficient control in the interview process (see Nichols and Groves, 1986). Also see the discussion in Chapter 8 in this volume.

### 2.8. *Recent developments in telephones surveys*

Cell phones, pagers, faxes, modems, Internet, call forwarding, voice mail, and other convenient services offered to phone subscribers are creating increasing challenges to researchers to contact the public for telephone interviews. The explosion of telephone area codes as a result of these new products creates a much bigger challenge to researchers to draw representative samples from their target population. According to a Lockheed Martin Study, United States will run out of new area codes by 2010. This implies evolving challenges for telephone survey sampling methodology. It is estimated that more than 25% of U.S. households have more than one land line. Households with children, Internet access, home-based businesses, and the difficulty to identify multiple phone line households create new challenges to draw a random sample of households.

Telephone Consumer Protection Act (TCPA) prohibits calls to wireless and assess a penalty for each such call. Land lines are household-based whereas cell (wireless) is population-based. Of the 9% of U.S. households that have no land lines, around 2.5% do not subscribe to a phone, and 6.5% have only wireless service. Cell phone numbers proliferate into RDD samples due to call-forward their land lines to wireless services. Wireless service-only households tend to be young males (less than 35 years), educated, employed, renter, earning less than $40,000 annually, and have no children. Wireless also does not have 911 services and when they get married and have family, they may opt for a land line. Some agencies such as survey sampling international (SSI) use software, wireless ID that reduces sampling risk by identifying potential wireless phones. Merging of two overlapping and incompatible sampling frames and households with multiple phone lines create potential cover bias.

The recent Cell Phone Sampling Summit II sponsored by Nielsen Media Research was convened to discuss how the cell phones are treated in RDD surveys. It is estimated that approximately 70% of the U.S. households have cell phones and it is growing. The telephone frame can be partitioned into three components: (a) land-line telephone exchanges, (b) cellular telephone exchanges, and (c) mixed-use exchanges. It must be noted that cellular telephone numbers are located in all those components of the frame. In addition to the issues discussed in telephone sampling earlier, the design should explicitly consider,

> "Weighing for unequal probability of selection, including whether a cell phone is a personal device reaching only one potential respondent or a household device reaching more than one potential respondent."

Because cell phone usage is on the rise among the teenagers, it is possible to reach ineligible persons when surveying adults and thus RDD cell phone calls may result in a wastage. These and other recent developments are to be carefully studied. We summarize a few studies that have addressed these issues later.

Tucker et al. (2007) report the telephone service and usage patterns in 2004 based on the information obtained from Current Population Survey (CPS). As observed earlier, standard RDD techniques usually exclude the cell phones, thus resulting in undercoverage. It is estimated that 6% of the households have only cell phone service. The percentage of one-person households that are cell-only (8.1%) is somewhat higher than that of large households (5.5%). Cell-only households are more likely to be renters than owners of homes. If the distribution is sliced by age approximately 20% young adults (18–24) are cell-only users. The data indicate that among those households that have both cell and land line, very few receive any calls in cell phones. Tucker et al. (2007) suggest using individuals as sampling units rather than households. But this can cause problems for households with multiple members who may share a single cell phone.

Brick et al. (2007) discuss the feasibility of cell phone surveys in United States. The contact rates across various time periods were the same for cell samples, whereas the rates for land samples were lower during weekdays. The refusal rate for cell sample is generally much higher and efforts to follow up also do not result in success. The text messaging was not effective in raising the cell response rate.

## 3. Fax surveys

In the mid 90s, there was a growing interest in conducting surveys via fax. Faster delivery was the main reason put forth along with the possibility that it may give the impression to the responder that the matter is important. Dickson and Maclachlan (1996) conduct a study to compare the mail surveys with fax surveys. They estimate that the cost per returned questionnaire in the fax was less than one-fourth of the cost for the mail surveys. The selection bias due to ownership of fax machines was not addressed. It is not known what percentage of households have fax machines and even if they have, what percentage of them keep them on. With the increasing use of scanners and the internet, fax surveys are not likely to take off.

## 4. Shopping center sampling and interviewing

Interviewing shoppers in shopping malls started in the early 60's when the development of totally enclosed shopping centers provided researchers access to a large number of shoppers from a wide geographic area. Prior to the mall intercept, surveys were mostly conducted in supermarkets, discount stores, train stations, and places where large concentrations of people could be found. More than 170 malls have permanent market research facilities, some of which are equipped with interviewing stations, videotape equipment, and food preparation facilities for conducting taste tests. A large number of malls permit intercepts on a temporary basis but may prohibit interviewing because they see it as an inconvenience to their shoppers.

The two major advantages of a mall intercept interview are cost and control and it has many of the advantages associated with personal interviewing. Also, it is the only way to conduct most taste tests and ad tests requiring movie projectors or videotape equipment. However, there are a number of disadvantages. The important one is that shoppers are frequently in a hurry and may not respond carefully. It may be difficult to maintain a controlled interviewing environment in the presence of the respondent's children, relatives, etc. Despite these problems, mall intercept interviews are increasingly used in market research. It is estimated that, of those who had participated in any form of a survey, 18% were contacted through mall intercept interviews compared to 12% through personal interviews (see Gates and Solomon, 1982). Because of the administrative efficiency, it has some potential for growth.

### *4.1. Sampling issues*

Samples for most shopping center interview are selected haphazardly and do not reflect the general population. The effect and sources of biases are not properly understood and are not taken into account. If the investigation is at the early stages of product development, it may not be necessary to follow rigorous sampling procedures. But if the objective is to generalize to the population, it is important to follow rigorous sampling schemes. Shopping center sampling can be compared to sampling mobile populations. The major interest in studies related to mobile populations has been in estimating the size of the population, but little attention has been paid to sampling time and location. Sudman (1980) provides some procedures that take these aspects into account.

The key assumption in the mall samples is that all households have a nonzero (but not equal) probability of begin found in a shopping center. The assumption may not be realistic and the bias introduced for some special groups such as lower income or older households may be substantial. Second, because the probability of selection is a function of the frequency of visits, that frequency must be estimable. This may strain respondent memory and may introduce some biases.

Sudman's procedure works as follows: First, select the shopping centers using the same basic random sampling procedures used in the selection of locations in a multistage area probability sample with probability proportional to a size measure such as total annual dollar volume. The optimum number of shopping centers and the number of respondents can be determined using the formulae for area cluster samples,

$$n_{\text{opt}} = \left[ \frac{C_1}{C_2} \left( \frac{1-\rho}{\rho} \right) \right]^{1/2} \tag{3}$$

where $C_1$ is the set up cost at a shopping center, $C_2$ is the cost per interview and with a total budget $C = C_1 m + C_2 mn$, where $m$ is the sampled number of shopping centers, $n$ is the number of interviews per shopping center, and $\rho$ is the intraclass correlation coefficient between shoppers within shopping centers. Because $C_1$ is generally much larger than $C_2$, large samples are selected from each center; but the heavy clustering increases the sampling variance.

The respondents can be selected either when they arrive at the center or as they move around within it. For the latter, we require information on how much time they have spent in the center because persons spending more time shopping have a higher probability of selection. To select an unbiased sample of entrances, it is important to know the fraction of customers the entrances attract from previous counts. This size measure can be used to sample entrances with probability proportional to size and is much more efficient than sampling them with equal probability. Though the less-used entrances will be sampled fewer times than the more heavily used entrances, the sampling rate would be higher at the less-used entrances if a self-weighting sample is desired. Establishing rules for within shopping center sampling is more difficult than entrance sampling. Identical traffic patterns in all parts of the center cannot be assumed because the location of discount stores is more likely to attract customers different from those who shop at fashion centers.

It is important to use careful time sample procedures, to avoid biases against certain types of customers, for example, working women who mostly shop in the evenings and weekends. Selecting an eligible time period with equal probability is not an efficient design. The solution is identical-sampling of time periods with probabilities proportionate to the number of customers expected in the time period. Sudman (1980) suggests forming time–location clusters, based on past data and selecting these clusters with probability proportional to past size.

The above mentioned procedures are far more sophisticated than those procedures used in the past. There are still problems in their implementation and generalizability. We suggest using the dual frame concept. For each shopping center, we may obtain trade area maps showing geographic areas from which stores draw their trade, because shopping centers generally attract those households nearest to it. These maps are sometimes drawn from shopper surveys (see Blair, 1983) intended for a different use by the retail merchants. With such a map, we may have a sampling frame from which we can draw an independent sample by telephone that can be combined with the mall sample. For a related discussion, see Bush and Hair (1985).

It is known that sampling bias may occur when the individuals spend different lengths of time at the survey location. Because most surveys are conducted away from the entrances, individuals who spend more time at the mall are more likely to be sampled. Such samples are known as length-biased (Cox, 1969). Recreational shoppers are likely to be overrepresented in the sample. Nowell and Stanley (1991) report a study on the bias of length of stay and suggest correcting for the bias using the procedures given in Cox (1969). The key factor appears to depend on whether individuals can accurately estimate the time they spend at the mall. Nichols et al. (1995) report that the length of time spent in the mall is different for Hispanics. Generally they spend more time traveling to the mall, but spend less time in it. Thus, both frequency bias and length of stay bias need to be considered for the shopping mall estimates.

## 5. Consumer panels

The panel has become an important tool for monitoring market factors ever since Jenkins (1938) and Lazarsfeld and Fiske (1938) used them to study brand preferences and reader reactions to a magazine (*Women's Home Companion*). Since then, the use of panels to study the purchase behavior of nondurable consumer goods has gained importance in North America and some Western European countries. See Hardin and Johnson (1971) for various applications of panels in marketing research. Marketing Research Corporation of America (MRCA) followed with a panel of 7500 households in 1941 and focused on the consumer purchase behavior of grocery, health and personal care, and textile products. Today, the use of panels in marketing studies is much more widespread and there are hundreds of consumer and industrial panels mostly located in North America and Western Europe. Nevertheless, some of the initial sampling problems related to panels still remain. This section will briefly review some of these problems from a sample design perspective. The problems related to panel sample design are not usually covered in discussions of sample survey methods. Sudman and Ferber (1979) identified three critical areas likely to induce bias in panel sample design. These are as follows: (i) bias created by initial refusals, (ii) bias created by subsequent mortality, and (iii) bias created through conditioning. A brief discussion of these areas follows. It must be recognized that there are other critical areas, such as aging of the panel and possible changes in the population that are not represented in the sample, which are not discussed here.

A consumer panel measures purchases of a product at any given point over a period of time. This has been used to measure market trends, seasonal effects, and the effects of marketing strategies. Panel data from the *Chicago Tribune*, National Panel Diary (NPD), National Family Opinion (NFO), Marketing Research Corporation of America (MRCA), Intercontinental Marketing Services (IMS), etc. focus on different product lines and industries. The majority specialize on consumer products, mostly nondurables distributed through grocery stores, whereas industrial panels such as those from IMS focus on hospital equipment, supplies, and doctor's prescriptions. Alternatively, store audits are used to estimate market size and trends (A. C. Nielsen) and with the advent of electronic scanners of Universal Product Codes (UPC), purchase data have become much more reliable and offer extensive detail on product/brand purchases as well as profiles of sample buyers. Information Resources, Inc. with headquarters in Chicago provide Infoscan and Behaviorscan services to business clients. Each panel member receives a member identification card that is presented to the store clerk at the time of checkout. All purchases are electronically recorded, eliminating the need for written diaries. This method has distinct advantages, as its popularity is growing both in North America and abroad (Information Resources has operations in Australia, Canada, France, Great Britain, Japan, and West Germany). These sources also study consumer brand preferences and brandswitchings over a period of time. Panel data have been extensively used in the formulation and evaluation of pricing strategies (see Montgomery, 1971). Segmentation by usage, package size, effectiveness of "marketing mix" variables have been studied by, among others, Blattberg and Sen (1976). Models are developed to predict market penetration based on repeated buying rates (see Eskin, 1973). With the information provided by panels on both purchasing and media exposure, efforts were made to estimate the effectiveness of advertizing particularly for new products (see Nakanishi, 1973). Carefoot (1982) and Information Resources, Inc. have used scanners to evaluate the effectiveness of advertizing. MRCA's panel data have been utilized to sense changing food habits leading to the modification of existing products and the development/introduction of new products to better serve the consumer.

### *5.1. Bias created by initial refusals*

Refusals, noncooperation, and nonresponse are to be expected in any survey. The level of cooperation attained is dependent on recruiting methods used and the nature of tasks required by the panel members. Often higher rates of cooperation are achieved if the expected effort from the respondent is lower. Panels recruited by face-to-face contact tend to have higher rates of cooperation than those recruited by telephone or mail. Oversamples are drawn initially to balance demographic variables such as geography, household size, income, education of the head of household, etc. Even if the panel fits all these demographics, there is no assurance that the panel results are bias free if willingness to cooperate on a panel and purchase of a product are related to a variable such as lifestyle. Panel cooperation seems to be closely associated with family size; for example, households with two or more members tend to cooperate more readily than single-person households. From the studies of the U.S. Department of Agriculture (1953) and additional investigations ("Panel bias reviewed," 1976), the following patterns emerge:

- Single-person households have a higher tendency to be noncooperators or "not-at-homes." They have less interest in food purchases and maintain records on an irregular basis.
- The older the housewife (after 55 years), the lower the chances of joining the panel. This may be related to education and the ability to keep records.
- Homeowners are more likely to cooperate than tenants. This again may be related to household size.
- Working wives are less likely to join the consumer panel than nonworking wives.
- Panel cooperators tend to be more "price conscious" than noncooperators.
- The income distribution of panel members and that of the U.S. population tend to be very similar except at the lower end where a smaller percentage of lower income households are represented in the panel.

Except for household size, the differences between cooperators and noncooperators tend to be negligible with respect to demographic profiles. However, the differences could be significant with respect to socio-psychographic characteristics such as organization, record keeping, and price consciousness. With new developments (Infoscan and Behaviorscan) the need to keep records by the panel members is eliminated, reducing potential errors in reporting, recall, and record keeping. Atwood consumer panels in Great Britain and Germany show no significant differences between panel members and the general population with respect to readership of magazines and newspapers and selected psychological and buying variables (Sudman and Ferber, 1979).

In summary, the evidence from the United States and European studies indicates that some biases in consumer panels such as household size, age of the housewife, and level of education of the head of household are possible. In panels requiring less effort, the refusal rate is lower resulting in lower sample bias. Panels that require more effort and those recruited by mail or telephone often tend to have a higher percentage of noncooperators resulting in higher bias.

The ratio method of estimation has often been used to obtain better estimates of the population. Under-representation of smaller households or a specific geographic region is overcome by the application of suitable poststratified weights in deriving the population estimates.

### *5.2. Bias due to attrition/mortality/formation of new households*

A panel should be representative of a target population. Though the population itself may not change drastically from year to year, some changes do occur over time. Dissolution of old households, formation of new households, household moves, etc. are examples of changing population characteristics. Potential problems are as follows: (i) panel member dropouts, (ii) household moves, (iii) household dissolutions, and (iv) new household formations. We will discuss each of them briefly.

(i) Dropouts: Panel dropouts or attrition is often estimated to be 5–10% from one period to the next in the United States. Charlton and Ehrenberg (1976) reported that 88% of their limited sample completed the 25-week panel. Farley et al. (1976) reported a 43% dropout rate from the waves of interviewing spanning 18 months. Personal situations, such as illness in the family, birth of a child, enlistment in the army, etc., are often the reasons for dropout. Two methods have been used to overcome this problem. An oversample could be made in anticipation of an expected dropout rate. However, in practice, it may not be possible to maintain large oversamples (European panel operators tend to follow this procedure). Besides, this would lead to sampling bias. The second method is to replace the dropout household with a new household of similar characteristics by a method of imputation in the field. The problem of noncooperation of a newly selected household is similar to that of initial recruiting. A prepared list of substitute households is searched until a replacement is found. Even if replacements are representative with respect to selected socioeconomic and demographic variables, they could differ on behavioral variables such as purchase quantity, degree of brand loyalty, private brand proneness, etc. Winer (1983) suggested that replacements be made with due consideration to selected behavior variables.

Sobol (1959) and Bucklin and Carman (1976) demonstrated that attrition introduces potential bias in panel-based market research. Hausman and Wise (1979) have designed a model of attrition and proposed a maximum likelihood method of estimation of parameters. They estimated the parameters in the presence of attrition as well as bias due to attrition. Winer (1980, 1983) and Olsen (1980) developed procedures for estimation of attrition bias in the absence of replacement of dropouts.

Maintaining a representative panel is not easy. Most panel operators recognize the importance of suitable compensation and effective communication with panel households as essential factors in keeping morale high and turnover rate at a minimum.

(ii) Household moves: When a household moves, it is a generally accepted principle to follow it. The only exception is if the panel is confined to a specific geographic area and the move takes the household out of that target area. Following the panel wherever they go ensures continuous representativeness of the panel including the patterns of mobility inherent in the population.

(iii) Household dissolutions: In the event that all members of a panel household die, the household is often replaced with a similar household. If one of the spouses dies and the other joins a nursing home, the household is dropped from the panel.

(iv) New household formations: The panels are continuously monitored as to the size of the household. If a new household is formed through marriage, the new household is recruited with probabilities inversely proportional to the number of persons who will constitute the new household. Thus, in the case of new households resulting from a marriage, half the split-offs are recruited. This way the panel recruits younger households to maintain their representatives in the population.

### *5.3. Bias created through conditioning*

The term "conditioning" refers to stimuli in a broad sense and includes all contacts between panel operators and panel households such as initial recruiting calls, instructions/training, diary keeping, compensation, and newsletter or other forms of communication whether personal or mail. Sudman and Ferber (1979) classified the effects of the stimuli into three categories: immediate, short-term, and long-term. These effects could be in terms of purchase behavior affecting brand choice, store choice, quantities purchased, number of shopping trips per unit time, expenditures on a product per unit of time, etc. For example, keeping a "time-use" diary might cause a person to use a different pattern of time utilization than the "usual." Besides changes in behavior, it might also change attitudes and beliefs affecting future behavior.

Studies focusing on the immediate effect of the acceptance of an invitation to join the panel on a household have used "recall" techniques to assess the differences in purchase behavior before and after joining the panel. The results, however, were inconclusive. The effect of short-term conditioning seems to be evident bases on empirical studies. A 1973–1979 study conducted by the Survey Research Laboratory at the University of Illinois on medical diaries found that first month reportings were 14% higher than the subsequent records of the following 2 months. Similarly, Sudman (1962) found that a panel diary method used to collect data on 10 product purchases over an 8-week period reported that first week purchases were 20% higher than the 8-week average and second week expenditures were 8% below the average. The experiences of U.S. Bureau of Census (1972–1973) also support the evidence of the existence of a short-term conditioning effect on the behavior of panel households. As a result, many practitioners ignore the first period as "trial" data or omit it in the trend analysis.

Substantial evidence exists that a special stimulus can result in major changes in reported purchase behavior. A sticker reminder in a diary and a postcard reminder to record all soft drink purchases resulted in an increase or more than 30% in reported purchases. A similar study on reporting purchases of citrus products (special form included for reporting) showed that the experimental group had a significantly higher incidence of purchase records of citrus products during the first month than the control group. However, the initial conditioning effect seemed to have disappeared in later months.

Some researchers have speculated that keeping diary records could sensitize households over time and cause them to be better shoppers. One panel study indicated that an average household made 2.7 trips per week for grocery shopping during the first 3 months of data collection period and 2.6 trips per week in the next 3 months. The differences are not statistically significant, and any conditioning effect is negligible. Ehrenberg (1960) using a British consumer panel and Cordell and Rahmel (1962) using A. C. Nielsen panel for television viewing habits concluded that there may be a slight short-term effect of panel conditioning but it disappears over the long term.

Long-term effects on households serving as panel members is of major concern as they could develop fatigue or become uninterested in keeping diaries. Interestingly enough, there is no evidence to support such a hypothesis. Ehrenberg (1960) described several studies and pointed out that over a 10-year period the Atwood consumer panel compared "old" and "new" panel members and found no significant differences. The general conclusion was that the length of panel membership did not systematically affect the reported results. Any "conditioning" that may exist in the early period of panel membership is likely to wear off or stabilize over a reasonably short time.

Some form of compensation is very common for most continuing panels and is often in the form of money, gifts, or other forms of motivation (participation in lotteries, etc.). The amount or value of compensation seem to vary widely depending on the type of respondent. For most consumer nondurables, the compensation has been in the range of $10–$60 a year. For physician panels, the compensation was several hundred dollars. Both European and Japanese panels seem to receive better compensation than those in North America. Ferber and Sudman (1974) and Sudman and Ferber (1971) reported that the households receiving compensation provided better quality data than those who did not. Their conclusion was that compensation in sufficient amounts is necessary to ensure initial and continuing cooperation as well as quality of reporting. There is no evidence that the form of compensation has any major impact on cooperation (Ferber and Sudman, 1974).

### *5.4. Consumer panels: other issues*

A study by Grootaert (1986) on the estimation of household expenditures in Hong Kong using the panel diary method suggested the use of multiple diaries—each member of the household maintains a separate diary of daily expenditures. This method resulted in more accurate reporting of expenditures particularly on "personal" products such as clothing, shoes, and services. The reporting arrangements depend on family structure, size, and decision making process within a household. As such, the results are not usually generalizable to other countries.

With high-tech electronic methods of data collection using scanners, the need to maintain written diaries is diminishing. As increasingly more retail stores are equipped with UPC scanners, data collection using panel method has become increasingly important for various marketing experiments. This has led to what is called "single source" research where many promotional experiments can be tested out by following the panel members from their TV sets to checkout counters.

It is easy to measure accurately the effect of promotional campaigns via this high-tech research. Information Resources, Inc. (IRI) monitors 3000 households in eight small town markets. The microcomputers record when the television is on and which station it is tuned to. IRI sends out special test commercials via cable channels. The single source research has it drawbacks. The size of the panels is still relatively small because of the high-cost nature of data collection and hence it is doubtful how generalizable the results would be to the entire market. Second, how do we know viewers are actually watching the test commercials. The change in the buying behavior is also questionable when the panel members are probably conscious of being in the panel. Brand loyalties are somewhat difficult to change by a short-term advertizing. But this research may be useful for new products (see Kessler, 1986).

### *5.5. Recent developments in consumer panels*

*International household consumer panels* are maintained by various commercially oriented survey research companies. SSIs surveyspot (U.S. Panel) covers North America whereas Opinionworld offers collective panel for Europe. SSI offers proprietary panels in more than 40 countries and in early 2007, it added China to the list of countries offering consumer panels. Though the literature on panels initially focused on consumer/household panels, now panels are extended to commercial and professional panels. Commercial panels are often used to track movement of goods and services at different stages of distribution to monitor trends. For example, a panel of pharmacists is used to track or monitor trends in prescription drugs.

## 6. Web surveys

The online world has become as important to Internet users as the real world (http://digitalcenter.org). "The internet has been a source of entertainment, information, and communication since the web became available to the American public in 1994." During the past decade Internet has become the primary vehicle for conducting marketing research. Web surveys, Internet panels, E-focus groups, web advertizing research, etc. have replaced traditional methods of conducting marketing research. Internet has also become rich source of secondary data and become universally accessible by anyone from anywhere and brought down the cost of conducting marketing research more effectively, with higher speed, and ever declining costs of unit information. Further developments, as noted in the sections that follow, will have profound effect on tools, and methods employed in marketing research over the next 10–15 years.

### 6.1. Internet penetration

Since Internet accessibility to the public in 1994, it has made rapid strides (see Tabel 2) to become a very powerful platform and changed the way we do business, and the way we communicate. It is the universal source of information. In fact, Internet is the most democratic of all mass media. With a very low investment, any business irrespective of its size can have a web page and reach a very large market, directly, fast, and economically. With a small investment almost anybody can have access to the world-wide web. The number of internet users in December 1995 was 16 million representing only 0.4% of the world population. This has grown to 361 million or 5.8% of the world population by December 2000 and then to 1.018 billion or 15.7% of world population by December 2005 and to 1.093 billion or 16.6% of world population in December 2006 (www.internetworldstats.com accessed on Feb. 12, 2007). This represents an annual growth of some 46.8% since 1995 and reaching a moderating annual growth of 25% during the past 5 years.

Although the annual growth of Internet users may moderate from the past rates of growth, it is a fair assumption to forecast 15–20% annual growth between now and 2010. That translates to 1.91–2.27 Bil Internet users by 2010. This tends to imply that Internet research can be representative and effective as other traditional methods with the fast growth of Internet populations. As reported recently by researchers, the problems of conducting Internet research must be effectively addressed and resolved, just as the problems with traditional research. (Ilieva et al., 2002; Kellner, 2004; Mathy et al., 2002; Schillewaert and Meulemeester, 2005)

English is the language of some 30% of Internet users followed by Chinese and Spanish by 14% and 8%, respectively. Japanese and German occupy fourth and fifth ranks with 7.9% and 5.3%, respectively. The top five Internet users languages account for nearly two-thirds of Internet user population. (www.internetworldstats.com) Table 2 presents SSI Internet Samples by country and official language. It can be seen that economically well-developed countries tend to have higher and similar levels of penetration.

Table 2
Internet users by region and penetrations by their respective populations

| Region | No. of Internet Users (million) | Internet Penetration (% of Population) |
|---|---|---|
| Asia | 389 | 10.5 |
| Europe | 313 | 38.6 |
| North America | 232 | 69.4 |
| Latin America | 89 | 16.0 |
| Africa | 33 | 3.5 |
| Middle East | 19 | 10.0 |
| Australia/Oceania | 19 | 53.5 |

*Source*: www.internetworldstats.com accessed on Feb. 12, 2007

### 6.2. *Web surveys: issues*

The power of the Web appears to have both positive and negative sides. Because it is relatively inexpensive to conduct surveys on the Internet, any business organization irrespective of its size can avail this opportunity. On the other hand, the proliferation of Web surveys makes it difficult to evaluate the quality of the surveys. Couper (2000) has observed: "It has become much more of a fragmentation than a bifurcation ("quick and dirty" versus "expensive but high quality," as was originally predicted) with vendors trying to find or create a niche for their particular approach or product." The Web surveys must be evaluated like other surveys in terms of their sampling, coverage, nonresponse, and error properties.

#### 6.2.1. *Coverage and sampling error*

The construction of sample frame for Web surveys that will lead to selecting probability samples is not easy. The sampling frames are often incomplete and the coverage error is probably the most serious as only about 42% of the population have used the Internet even in the United States. Although it is expected to grow, it is not clear how much this percentage will be in the future. It may be constrained by the interest of the population in information sources.

The problem is not only who has access to the Internet but also the demographic and behavioral difference of the population base between those who have access and those who do not. The National Telecommunications and Information Administration (NTIA) report generally identifies that income, race, education, and household composition all play a role in having Internet access. Thus, the challenge for Web survey researchers is to find ways to reach the target population or otherwise, the inferences from survey results could be very restrictive. Because of the coverage issue, the sampling errors are likely to be high and skewed.

#### 6.2.2. *Nonresponse error*

The nonresponse error depends on the rate of nonresponse and on the difference between respondents and nonrespondents on the variables of interest. When the sampling frame itself cannot be defined, the problem becomes even more acute. Couper et al. (1999) summarize the response rates in e-mail surveys and observe that the response rates in e-mail surveys are lower than the rates for mail surveys. Several reasons are attributed to this gap: lack of personalization as in mail surveys, technical difficulties in using the Internet, and confidentiality concerns.

#### 6.2.3. *Measurement error*

The Web is more flexible for constructing survey instruments, such as adding visuals, etc., and therefore provides many options in its form and in its content. There is no definite conclusion on the ideal form of the surveys; it is clear that it all depends on the target population. In longitudinal surveys it is possible that the response over time may be biased. Given that the sampling frames are not easily defined for Web surveys, the statistical adjustments that could be made need to be studied more carefully.

### 6.3. *Types of web surveys*

Couper (2000) provides a neat summary of various types of Web surveys (see Table 2 of his paper). The surveys could be broadly classified as based on nonprobability and probability methods. Although the nonprobability methods are similar to other media surveys such as telephone and mail surveys, we focus here on probability methods. They take only two forms: restrict the sample to population with web access and thus limiting the generalizability of the survey results; use other methods to reach broader population via RDD-type of tools. These are briefly summarized in the following sections.

### *6.3.1. Intercept surveys*

These are targeted toward visitors to a Web site and are used mainly for eliciting product-related opinions and in general to acquire customer feed back. Typically, systematic sampling is used and cookies are used to track the visits and for the timing of the exposure to the survey.

### *6.3.2. List-based samples*

Here, we begin with a list of households with Web access and invite a select sample to participate in the survey with proper checks for avoiding duplications. Although this may cover only a portion of targeted sampling frames, yet useful estimates of samples error, etc. could be derived.

### *6.3.3. Prerecruited panels*

Panel members are selected using probability sampling methods such as RDD and are recruited for surveys on the Web. Because the selection has a probability basis, the quantification of various types of errors are possible. However, it is recognized that the nonresponse errors, etc. could be still different for Web and further research must be done to better understand the dynamics of nonresponse between the Web and other modes.

### *6.3.4. Probability samples*

The basic approach here is to first define the appropriate sampling frame and then provide Web access to those who are recruited but do not have access. Although this is much more scientific than the other methods, the initial response rate to the recruiter interviews has been somewhat poor. But this approach essentially solves the problem of representative coverage and Web access.

### *6.3.5. Mixed-mode Design*

These designs combine various modes of reaching the targeted group. Mixed-mode surveys provide an opportunity to overcome the weaknesses of each method, but deployment of mixed-modes of data collection raises many challenges including the possibility that some respondents may give different answers to each mode. Dillman (2007) identifies five situations for use of mixed-mode surveys.

- Collection of the same data from different members of a sample.
- Collection of panel data from the same respondents at a later time.
- Collection of different data from the same respondents during a single data collection period.
- Collection of comparison data from different populations.
- Use one mode to prompt completion by another mode.

Use of mixed-mode surveys may enhance the possibility of improving response rates and reduction of nonresponse and coverage errors. Though there is no compelling evidence for choice and sequence of mixed-modes to be employed, it appears prudent to start with a method that is least expensive such as web surveys and then move towards mail, telephone, and if still necessary employ personal interviews. When multiple modes of data collection are employed it is essential to make a deliberate effort to deliver equivalent stimulus regardless of whether it is delivered aurally or visually.

### *6.4. Online Panels*

The conduct of surveys via online panels has gained prominence in recent times because of easy access to the internet. But the recruitment of panel member in some cases is not based on known sampling methods. Harris Interactive for examples recruits panel members on a voluntary basis, but collects extensive demographics for postsurvey adjustments. Knowledge Network uses the telephone RDD methodology to recruit the panel members, but the survey information is collected via the internet. In a study related to health surveys, Baker et al. (2003) found that these online panels are no different from other panels in terms of response rate, attrition, etc.

#### *6.4.1. Some panel sources*

Many suppliers of marketing research services offer online panels (www.iri.com; www.acnielsen.com; www.lightspeedresearch.com; www.mra-net.org, and others). Information Resources Incorporated offers a behavior scan system whereas AC Nielsen offer scan track and specialty panels such as the African-American consumer panel (www.targetmarket.com). Foreign vendors such as Marsc panel management (www.marsc.co.uk), and Intage Inc. (www.jmra.net.or.ip) offer consumer panels in the U.K. and Japan, respectively. Several vendors such as Forrester (www.forrester.com), Lightspeed (www.lightspeedresearch.com), Marketmakers group (www.marketmakersgroup.com), Perez (www.perez.com), Robert Thale Associates (www.robertthaleassociates.com), and others offer B2B panels as well as consumer panels. Future Information Research Management (FIRM) (www.confirmit.com) operates a database of global 5000 for use of online B2B surveys and panels.

## 7. Conclusion

With the advent of Internet, we expect that a great number of market research surveys will be carried out through the Web. Because of the coverage issues we expect increased use of mixed-mode surveys. The surge in the number of surveys is bound to affect both the nonresponse rates as well as the quality of the response. The Web provides a unique opportunity to customize the surveys but this comes with a price of increase complexity to generalize the results of the surveys because, the web universe is still evolving. Several articles that have appeared in the *International Journal of Market Research* emphasize the various issues pointed out in this chapter. It is obvious that until the issue of Web sampling frame is clearly understood, the quantification of sources of survey errors will continue to have bias.

*Sample Surveys: Design, Methods and Applications, Vol. 29A*
ISSN: 0169-7161

DOI: 10.1016/S0169-7161(08)00021-7

21

# Sample Surveys and Censuses

*Ronit Nirel and Hagit Glickman*

## 1. Introduction

For many people, the simultaneous use of the terms census and sample survey seems contradictory. This chapter highlights the past use of sample data in census projects, as well as describes innovative developments in census methodology that accommodates sample data to various degrees. We start with a brief presentation of the main features of a census and the new trends in census methodology.

Censuses provide a core of official statistical data, around which demographic analyses, survey estimates, and administrative data are calibrated. A population census has been defined recently as "the operation that produces at regular intervals the official counting (or benchmark) of the population in the territory of a country and in its smallest geographical sub-territories together with information on a selected number of demographic and social characteristics of the total population" (United Nations Economic Commission for Europe [UNECE], 2006, p. 6, no. 19). The essential features of a census, as specified by the Commission, include universality, simultaneity of information, and individual enumeration. These features imply that there is a (a) well-defined census population, (b) reference date for all census data, which is usually referred to as Census Day, and (c) accurate data pertaining to individuals with regard to place of residence and other sociodemographic characteristics on Census Day are collected. Thus, for a person to be enumerated correctly in a census, nontrivial eligibility criteria must be met.

*National eligibility*. Each person should have one and only one usual place of residence, which defines the country he or she belongs to. A person living continuously in one country for more than a predefined period of time (e.g., for over one year on Census Day) is considered a "usual resident" of that country and is included in the target census population. A person living outside the country for longer than that period is not eligible for enumeration in the census. An illegal work migrant, for example, may be eligible for enumeration in the census if he or she has resided continuously in the destination country for more than, say, one year. The crucial role of the time reference is also noteworthy. For example, a baby born one day after Census Day is considered ineligible for enumeration, whereas a person in the target population who died one day after Census Day is eligible.

*Local eligibility*. Within a country, every individual should be counted at his or her usual place of residence on Census Day. The definition of a "usual residence" specifies, for example, criteria for people who divide their time between two places of residence (e.g., people working away from their family's place of residence; children in shared custody). The issue of place of residence also interacts with the time reference. People who moved one day before Census Day are considered ineligible at their former address, even if they lived there for 30 years, and are eligible at their new address.

For detailed recommendations on eligibility issues, see UNECE (2006) and United Nations (2006).

In the past, censuses have involved nationwide area enumeration (door-to-door data collection). In recent decades, however, it has become evident that there is more than one way to conduct a census with the essential features mentioned above. First, various collection methods have been developed within the traditional door-to-door framework, including self-enumeration and mail back and/or mail out options. Another group of new methodologies includes censuses that are based on administrative sources, with or without supplementary fieldwork data. The third group of innovative census-taking methods is based on appropriate accumulation of *sample* survey data that cover the census population over a predefined period of time.

It is not difficult to understand why many countries invest in developing new census methodologies. To begin with, advances in information technology have enhanced the role of administrative data in managing official statistics, including construction of registers such as population and housing registers. Enhancement of record linkage procedures has made it possible to accumulate data from various sources. Concomitantly, the demand for more detailed information has increased, whereas the willingness of individuals to respond to questionnaires has declined. Thus, many new approaches focus on improving quality and timeliness of census outputs while reducing the response burden. Finally, some countries expect the new methodology to reduce costs and enable expenses to be distributed more evenly over time.

Regardless of the methodology used, census counts are subject to different types of errors, which include coverage, content, and operational errors. Of those, coverage errors are the most crucial. There are two types of coverage errors: undercount and overcount. Undercount occurs when an eligible person is omitted from the enumeration, and overcount occurs when an ineligible person is erroneously enumerated. The definition of coverage errors depends on the geographical scale of interest. For example, some countries define a coverage error only at the national level, whereas other countries consider a person who is enumerated incorrectly at the level of a geographical region (e.g., in the wrong province) as contributing to undercount in the correct region and to overcount in the region of enumeration. Errors at various geographic scales within a given country can be of interest as they may relate to different uses of census data.

Because the census is an important and costly operation, evaluation of its coverage errors has become a "state-of-the-art" procedure in census-taking. Thus, many countries conduct a postenumeration survey (PES) immediately after enumeration (whether the enumeration is field-based or administrative), with the objective of estimating census coverage rates. Furthermore, some countries use PES estimates to adjust census counts for coverage errors. Section 2 describes the use of sample surveys to estimate coverage errors. The section begins with a description of model-based undercount estimation (Section 2.1) and extends the approach to estimation of undercount and overcount (Section 2.2). To conclude, we present design-based approaches to coverage evaluation (Section 2.3). In that section, corresponding sample designs are presented together with illustrative examples.

The PESs are generally large surveys possibly comprising hundreds of thousands of households. As such, some countries conduct operations that evaluate the PES. This "second-order" evaluation becomes a "first-order" evaluation of the census output when the PES results are used to statistically adjust the census counts. Although the main source of bias in "raw" census counts are errors pertaining to coverage, the main sources of bias in a PES or in adjusted counts relate to measurement errors, modeling, and processing. Owing to the dearth of comprehensive investigations on this topic, Section 3 attempts to provide a conceptual framework for possible uses of sample data to evaluate statistical adjustment of census counts. In light of the growing diversity of census-taking methods, we focus on broad principles rather than on specific solutions. The first step we propose is to analyze the remaining uncertainty in the adjusted counts (Section 3.1). The second step is to identify potential errors resulting from this uncertainty (Section 3.2), and the last step is to design different evaluating operations, including "Evaluation Follow-Up" (EFU; see Section 3.3).

Section 4 deals with an entirely different types of census that is based on a system of sample surveys. This approach adopts the principles of a rolling sample design, which is briefly described in Section 4.1. The remainder of Section 4 focuses on a description of the rolling census in France, which is the only country that has decided to carry out a sample-based census to date. In the concluding section, we describe sample surveys carried out in conjunction with a census (Section 5). These are the "long-form" surveys, in which comprehensive socioeconomic information is collected in the framework of the census. The U.S. Census is presented as an illustrative example of the main methodological features of such surveys (Section 5.1). A relatively recent development of the long-form concept is referred to by the UN a "traditional census with yearly updates of characteristics" (UNECE, 2006). In this type of census, only short-form data are collected, and the long form is replaced by a set of annual samples from which socioeconomic data are collected during the intercensal years. This idea was implemented in the American Community Survey (ACS), which is described in Section 5.2.

## 2. The use of sample surveys for estimating coverage errors

A population census is exposed to different types of errors, including coverage, content, and operational errors (UNECE, 2006). Of these, coverage errors are the most serious because the main objective of a census is to provide a full and accurate count of the population. Let $C$ be the census count and $N$ be the true population count. The census net coverage error $D$ is defined by

$$D = N - C.$$

A positive value of D indicates net undercount and a negative value indicates net overcount. Coverage errors arise due to omissions or erroneous enumerations of people in the census. In the past, when enumerators conducted door-to-door enumeration, the most common coverage problem was undercount of dwellings and of individuals within dwellings. In recent years, interest in correcting overcount errors has grown, as several countries base their censuses on administrative registers, and as door-to-door enumeration and form collection has been replaced by various combinations of data collection through the mail and internet. Registers are often subject to inaccuracies in both directions (undercount and overcount) because of delayed updates or lack of reporting. Mailing and internet responses are exposed to duplications and fabrications, as well as to difficulties in understanding census eligibility criteria. Because the census is a large, central, and costly operation carried out once every 5–10 years, evaluation of coverage errors has become a "state-of-the-art" procedure in census-taking. Thus, many countries conduct a PES immediately after the census enumeration activity that estimates the census coverage rate. Furthermore, some countries use PES estimates to adjust census counts for coverage errors. In this section, we describe several coverage models and the corresponding PES sample design, as well as design-based evaluation programs.

### 2.1. *Dual system estimator–based estimation of undercount*

To begin with, we consider a census list that is exposed only to undercount. To estimate the extent of undercount in the census list, another source of information is required, namely, another full or sample-based enumeration. The most known model for estimating the size of a closed population using two incomplete enumerations is the capture–recapture model. This model has been used since the 19th century in many disciplines, such as wildlife management, epidemiology, physics, criminology, software testing, and, of course, demography (see, e.g., reviews by Chao, 2001; Schwarz and Seber, 1999). Variants of the problem include the case where two enumerations attempt to count all members of the population (nonsample case) and the case where one of the enumerations is sample-based (census sample). Data may be obtained from field data collection or from a list frame such as a register. In the census-taking literature, these models are referred to as dual system estimators (DSEs). For a comprehensive review of literature on capture–recapture modeling in census methodology, see Fienberg (1992) and Chao and Tsay (1998).

#### 2.1.1. *The dual system model*

We will begin with a description of the standard DSE (Peterson, 1896; Sekar and Deming, 1949; Wolter, 1986). Consider a closed population $\Omega$, which comprises $N$ individuals residing in a given geographical area at a specific time. Assuming that $N$ is fixed but unknown, the problem is to estimate $N$. Suppose, for the time being, that we have made two attempts to count the entire population and have obtained two lists of identified individuals. After matching the two lists, a $2\times2$ table is set up, as in Table 1. Table 1 is commonly referred to as *The Dual System Estimation Table*.

The entries in the table relate to the number of people counted in list A and list B, $Y_{11}$; the number of people counted in A but not in B and in B but not in A, $Y_{10}$ and

Table 1
The dual system estimation table for the nonsample case

| | | Census List B | | |
|---|---|---|---|---|
| | | Counted | Missed | Total |
| Census List A | Counted | $Y_{11}$ | $Y_{10}$ | $Y_{1+}$ |
| | Missed | $Y_{01}$ | $Y_{00}$ | $Y_{0+}$ |
| | Total | $Y_{+1}$ | $Y_{+0}$ | $Y_{++} = N$ |

$Y_{01}$, respectively; and the number of people missed in both lists, $Y_{00}$. Note that $Y_{00}$, the marginal totals $Y_{+0}$, $Y_{0+}$ and the population total $Y_{++} = N$ are unobservable, and therefore need to be estimated. Let $p_{ab}$ be the probability of inclusion in the *ab*th cell, $a, b = 0, 1, +$. The estimation procedure is based on three major assumptions: (A1), lists A and B are created as a result of $N$ mutually independent trials (autonomous independence); (A2), counting probabilities are homogeneous across individuals (heterogeneous independence); and (A3), the event of being counted in list A is independent of the event of being counted in list B (causal independence). With these assumptions, the probability of being counted twice is the product of the marginal counting probabilities, $p_{11} = p_{1+}p_{+1}$, and the maximum likelihood estimators of the probabilities that a person will be counted in list A and in list B, $p_{1+}$ and $p_{+1}$, respectively, and of the total population $N$, are

$$\hat{p}_{1+} = \frac{Y_{11}}{Y_{+1}}, \quad \hat{p}_{+1} = \frac{Y_{11}}{Y_{1+}}, \quad \hat{N} = \frac{Y_{1+} \cdot Y_{+1}}{Y_{11}} = \frac{Y_{1+}}{\hat{p}_{1+}}. \tag{1}$$

It can be seen that the total population estimator $\hat{N}$ is expressed as the number of individuals counted in the first list divided by the estimated counting probability of this list $\hat{p}_{1+}$. Given assumptions (A1)–(A3), these estimators are strongly consistent with asymptotic normal distribution (see, e.g., Alho, 1990).

In many applications of the dual system methodology, it is not realistic to assume that both lists are based on a counting procedure that aims to count the entire population. Wolter (1986) provides a detailed description of an alternative model where only one list, for example, the first one, is a nonsample list. The second list is based on a sample of people that is selected from the target population for possible inclusion in the second list. In the context of a census, the first list would be the full enumeration census list and the second list would be provided by a postenumeration undercoverage sample survey.

Suppose that the underlined geographical area is divided into $M$ plots known as enumeration areas (EAs). A simple random sample of $m$ EAs is chosen, and the data collected for list B consist of an enumeration of the population living at the sampled areas only. At this stage, it is assumed that both the census and the survey count only eligible people. For this census-sample case, the maximum likelihood estimators of $N$ and the marginal capture probabilities are

$$\tilde{p}_{1+} = \frac{Y_{11}^U}{Y_{+1}^U}, \quad \tilde{p}_{+1} = \frac{Y_{11}^U}{Y_{1+}^U}, \quad \tilde{N} = \frac{Y_{1+} \cdot Y_{+1}^U}{Y_{11}^U} = \frac{Y_{1+}}{\tilde{p}_{1+}}, \tag{2}$$

where the superscript $U$ indicates sampled EAs. The population size is estimated as the total number of individuals counted in the census list divided by the sample-based estimator of the counting probability $\tilde{p}_{1+}$. The parameter $p_{1+}$will also be referred to as the undercount parameter.

When a more complex sampling design is used for sampling EAs, predictors of $Y_{11}$, $Y_{1+}$, and $Y_{+1}$ are calculated according to the particular sampling scheme. In that case, the population size is estimated by substituting the appropriate design-based predictors in (1), $\tilde{N} = (Y_{1+}\hat{Y}_{+1})/\hat{Y}_{11}$, where $Y_{1+}$ is the census count as before. For example, in the U.S. Accuracy and Coverage Evaluation (ACE) Survey described in Section 2.2.2 (U.S. Census Bureau, 2004), $Y_{+1}$ is essentially predicted by $\hat{Y}_{+1} = \sum_{k \in S} w_k$, where $w_k$ reflects the inverse of the probability of selection of person $k$ in the sampled EAs, as well as adjustments for missing data and other operational problems. A similar predictor is calculated for $Y_{11}$.

A slightly different estimation approach was taken by the United Kingdom 2001 One Number Census (ONC). Here, the classical DSE (1) was applied at the EA level, and a ratio estimator was used to estimate the size of the entire population.

$$\tilde{N} = \tilde{R} Y_{1+} \text{ where } \tilde{R} = \sum_{i \in U} \hat{N}^i \Bigg/ \sum_{i \in U} Y^i_{1+} = \sum_{i \in U} \frac{Y^i_{1+} Y^i_{+1}}{Y^i_{11}} \Bigg/ \sum_{i \in U} Y^i_{1+}, \tag{3}$$

where $i$ indicates EAs, see Brown et al. (1999).

Wolter (1986) derived an expression for the asymptotic expectation and variance of $\tilde{N}$ given in (2) by applying the standard Taylor series method, $\mathrm{E}\tilde{N} = N + C$ and $\mathrm{Var}\, \tilde{N} = N \cdot C$, where

$$C = \frac{(1 - p_{1+})(1 - p_{+1})}{p_{1+} p_{+1}} + \frac{1 - f}{f} \frac{1 - p_{1+}}{p_{1+} p_{+1}}, \tag{4}$$

and $f = m/M$ is the sampling fraction. When other estimation schemes are adopted (e.g., the ONC ratio estimator (3)), the variance is usually estimated by resampling methods.

The estimators described above rely heavily on model assumptions. Some modified versions of those estimators have attempted to deal with the potential failures of these assumptions. Bias resulting from failure of the heterogeneous independence or causal independence assumptions (assumptions (A2) and (A3) in Section 2.1.1, respectively) is referred to as correlation bias. Heterogeneity in counting probabilities is often handled by poststratification. Typical stratification variables include geographic units, age × gender groups, and other socioeconomic variables. Thus, it is assumed that the heterogeneous independence assumption is satisfied within poststrata, and a DSE is computed for each stratum. Huggins (1989) and Alho (1990) further generalize the DSE to cases in which counting probabilities vary for different people (e.g., cases when some heterogeneity still remains within strata). The individual counting probabilities are estimated through logistic regression using relevant explanatory variables. The model predicts the propensity that a person will be counted in the census and in the sample (for further insights on this topic, see also Haines et al., 2000, and a presentation by Bell, 2007).

Causal independence is not satisfied if the act of someone being included in the census affects his or her probability of inclusion in the coverage survey. This can happen, for example, when data collection for the PES and census enumeration are not conducted at completely different times or if some information about the first enumeration is available at the second enumeration. Let $\theta = Y_{11}Y_{00}/Y_{10}Y_{01}$ be the odds ratio in Table 1, then $E\theta \cong 1$ under the assumptions of the model. However, in the presence of correlation bias, the total population $N$ can be estimated as

$$\hat{N}(\hat{\theta}) = \hat{N} + (\hat{\theta} - 1)Y_{10}Y_{01}/Y_{11}, \tag{5}$$

where $\hat{\theta}$ is a predictor of $\theta$ and $\hat{N}$ is the DSE estimator (1) (Bell, 1993). Additional independent (external) data are required to predict $\theta$ (e.g., administrative data or demographic estimates). If those data are available, we may have a good demographic total estimate at the national level. Plugging this estimate in place of $\hat{N}(\hat{\theta})$ in (5) yields a prediction of $\theta$. If we are interested in corrected DSEs within poststrata, either external total estimates for these strata are required or some assumptions on $\theta$ can be made. Since no accurate external estimates are available at subnational levels in many instances, the second alternative is usually adopted. The simplest assumption is that $\theta$ is constant across all poststrata. In that case, a synthetic estimate is obtained using strata-specific census and PES counts combined with a national estimate of $\theta$. Other possible assumptions that yield corrected DSEs have been proposed by Bell (1993, 2001), Elliott and Little (2000), and Brown et al. (2006). Methods based on a third enumeration, known as triple system estimators, are discussed by, for example, Darroch et al. (1993) and Zaslavsky and Wolfgang (1993).

Finally, we note that the dual system model is based on the multinomial distribution. An alternative model for the target population capture process is based on the Poisson distribution (e.g., Cormack and Jupp, 1991). The main advantage of the Poisson model is its amenability to standard maximum likelihood theory. Log-linear models are also discussed extensively in the capture–recapture literature (e.g., Rivest and Levesque, 2001), including models that accommodate heterogeneous capture probabilities. As far as we know, the Poisson and log-linear models have not been used in a census context.

### *2.1.2. Principles of sample design and an illustrative example*

To estimate undercount using the DSE, undercoverage postenumeration surveys have been conducted in many countries, including Australia (Australian Bureau of Statistics, 2007), Italy (Cocchi et al., 2003), Turkey (Ayhan and Ekni, 2003), and the United Kingdom (Brown et al., 1999). Typically, two-stage or multistage stratified area samples are used in those surveys. To keep the description simple, we present a two-stage design that has the following features:

*Target population.* Ideally, the PES target population should be the same as the census target population. In practice, however, some countries exclude population groups such as people living in nonprivate dwellings or in remote areas.

*Primary sampling units.* These first-stage units are relatively large geographical units such as municipalities.

*Secondary sampling units.* The primary sampling units (PSUs) are partitioned into smaller plots that typically comprise several dozens of households or dwellings. These may correspond to existing administrative units such as postcode areas, geographical units such as blocks, or often EAs defined specifically for the census. The main rationale for selecting an area-based cluster sample is the lack of reliable sample frames that are independent of the census enumeration, and that can be used to select units such as dwellings, households, or individuals. In addition, an area sample might be preferred for operational reasons.

*Sample frames*. When address files or building registers exist, they may be used to design and select the secondary sampling unit (SSU) sample. These lists can be on a national scale or can comprise a combination of local lists. As mentioned before, in many situations, only a PSU-level list is available and the second-stage sampling is based purely on geographic area maps. It should be emphasized that the frames should not depend on census information.

*Stratification of PSUs*. Undercount is generally not homogenous across PSUs. For example, it is expected that undercount will be higher in areas with a high immigration rate or with a high proportion of young people. Therefore, many countries stratify PSUs by characteristics that were found to be important determinants of undercount to attain sample efficiency. For example, a national Hard-to-Count (HtC) score was constructed in the U.K. ONC based on the previous census information. The index distinguished among PSUs by their expected level of census coverage.

*Sample size and sample allocation*. The overall sample size usually balances accuracy requirements, cost, and other considerations such as using the PES to collect additional socioeconomic data (see Section 5). Some countries aim at sample sizes ranging from 200,000 to 400,000 households (e.g., Italy and United Kingdom) and others have smaller samples ranging from 20,000 to 40,000 households (e.g., Australia and Turkey). Sample allocation to strata usually aims to minimize the variability of the DSE in key areas and population groups. Hence, sampling fractions are generally unequal in different strata. As such, sampling rates are typically expected to be larger in areas with higher undercount rates, as can be seen in Eq. (4). Since counting probabilities are unknown at the time of sample design, the design should be robust to deviations in the hypothesized distributions. Hence, simulation studies (e.g., Brown et al., 1999), sensitivity analyses (e.g., Nirel et al., 2003), and pilot surveys are used to design the sample.

*The U.K. 2001 Census Coverage Survey*. The 2001 ONC project aimed to identify and adjust for omissions in the 2001 Census. Undercount was evaluated on the basis of a PES known as the Census Coverage Survey (CCS). The objective of the CCS was to provide undercount estimates at the subnational level by age groups and gender and was to allocate the undercount to small areas. The CCS was a stratified two-stage sample, where the primary strata were estimation areas comprising approximately 500,000 people. There were 101 estimation areas in England and Wales, eight in Scotland, and three in Northern Ireland. The PSUs were the 1991 Census Enumeration Districts (EDs).

Using the HtC distribution, three strata corresponding to three levels of enumeration difficulty (lowest 40%, middle 40%, and top 20%) were defined within each estimation area. Within these strata, PSUs were further stratified by size groups of the key age-gender distribution. The size strata were formed on the basis of a design variable, which captured the age-gender structure of the EDs using babies, young males, and elderly female age-sex groups. A sample of EDs was selected within those strata, and postcodes (SSUs) were selected within each selected PSU, with probability related to the mean number of addresses per postcode within the selected ED. A sample of about 19,500 postcodes was selected, which consisted of about 370,000 households (including about 320,000 in England and Wales alone). The sample size aimed at a 1% relative error rate for the EA level estimates, and a relative rate of 0.1% for the national counts. Data was collected by face-to-face interviews as compared to the Census self-completion questionnaire. For more details, see Brown et al. (1999) and Abbot and Marques dos Santos (2007).

## 2.2. DSE-based estimation of undercount and overcount

As we have pointed out, the problem of census overcount has become as important as undercount in census methodology. Surprisingly, for example, in the U.S. 2001 census, the estimated net undercount was $-0.5\%$, that is, the census data essentially revealed a net overcount rather than a net undercount. In this section, we will describe two approaches to estimating overcount, together with a dual system for estimating undercount. The first approach suggests a design-based estimate of overcount rate. The second approach is based on the multinomial-Poisson model, which extends the dual system multinomial model and includes both undercount and overcount parameters.

### 2.2.1. Extending the dual system estimator

In the undercount scenario described in Section 2.1, two sources of information were required to estimate undercount—census enumeration and a sample of EAs. We refer to this sample as the undercoverage sample or in short, the U-sample. Consistent with the DSE assumptions, the U-sample is drawn independently from the census list, for example, an area sample. If the census list comprises $Z$ enumerations, of which $X$ are ineligible, then $Z = Y_{1+} + X$. We also assume, as before, that the U-sample list does not include ineligible persons.

To predict $X$, a second sample is selected, that is, the overcoverage sample or the O-sample. The objective of this sample is to identify erroneous enumerations that result from counting ineligible people in the census. Therefore, the sampling frame is now the census list, which is assumed to consist of eligible and ineligible people. To reduce cost and simplify data collection, it is convenient that the O-sample comprises the same EAs as the U-sample. Specifically, the O-sample consists of all people who are enumerated by the census in the same plots as those that are selected for the U-sample.

Assume, as before, that a simple random sampling scheme of $m$ plots out of $M$ has been selected. The design-based adjustment of the standard DSE for overcount first shrinks the census count, $Z$, by the O-sample estimate of the share of correct enumerations in the list, $(Z^O - X^O)/Z^O = Y^O_{1+}/Z^O$, where the superscript $O$ indicates O-sample counts. The predicted number of eligible people is then expanded by the U-sample estimate of the counting probability $\tilde{p}_{1+}$ of (2). In sum, we obtain

$$\tilde{N}_D = Z \frac{Y^O_{1+}}{Z^O} \frac{1}{\tilde{p}_{1+}} = \frac{Z \cdot Y^O_{1+} \cdot Y^U_{+1}}{Z^O \cdot Y^U_{11}}. \tag{6}$$

This design-based approach to estimating overcount together with the DSE has been used in U.S. Censuses (e.g., Hogan, 1993, 2003; Mulry, 2007, see below) and in Switzerland (Renaud, 2007a). Variance estimates for $\tilde{N}_D$ can be obtained by resampling methods such as stratified jackknife methods (U.S. Census Bureau, 2004, Section I, Chapters 7–14).

A model-based approach for estimating and adjusting for overcount is based on five basic assumptions. The three assumptions of the classical DSE model (A1)–(A3), plus two additional assumptions: (A4), the number of ineligible people counted by the census in an EA, $i$ is distributed according to the Poisson distribution, with expectation $\lambda N^i$ for $\lambda > 0, i = 1, \ldots, M$; and (A5), all EA counts of eligible and ineligible people are mutually independent. Note that assumption (A4) states that the rates of ineligible counts are homogeneous across EAs. This multinomial-Poisson model was proposed by Glickman et al. (2003) for the 2008 Israeli Integrated Census.

When $Z = Y_{1+} + X$ represents the number of eligible and ineligible people counted in the census list as before, the estimators derived from the model are

$$\tilde{p}_{1+} = \frac{Y_{11}^{U}}{Y_{+1}^{U}}, \quad \tilde{p}_{+1} = \frac{Y_{11}^{U}}{Y_{1+}^{U}}, \quad \tilde{\lambda} = \frac{X^{O}}{Y_{1+}^{O}/\tilde{p}_{1+}} \quad \tilde{N}_M = \frac{Z}{\tilde{p}_{1+} + \tilde{\lambda}}. \tag{7}$$

Thus, the estimate of the undercount parameter $p_{1+}$ is the same as in (2), and the overcount parameter $\lambda$ is estimated by the share of ineligible persons out of the eligible persons on the census list, with an adjustment for undercount. Note that since $p_{1+}$is the expected share of eligible persons in the list and $\lambda$ is the expected share of ineligible persons, their sum amounts to the expected size of the list, $Z$, divided by the size of the target population. In that way, we obtain the estimator $\tilde{N}_M$ in (7). Note that for the sample design considered here, the census list corresponding to sampled EAs was the same for the U-sample and the O-sample.

The expressions for the asymptotic expectation and variance are similar to those derived by Wolter (1986), for the case with no overcount in the census list. We obtain $E\tilde{N} = N + C$ and Var $\tilde{N} = N \cdot C$, where

$$C = \frac{(1-p_{1+})(1-p_{+1})}{p_{1+}p_{+1}} + \frac{1-f}{f}\frac{1-p_{1+}}{p_{1+}p_{+1}} + \frac{1-f}{f}\frac{\lambda}{p_{1+}+\lambda}\left(\frac{p_{1+}}{p_{1+}+\lambda} - \frac{1-p_{1+}}{p_{1+}}\right). \tag{8}$$

The last term in the right side of Eq. (8) represents the contribution of overcount to the variance of the census population estimator.

It should be noted that the share of eligible persons estimated in (6) is defined with respect to the total number of census enumerations (eligible and ineligible), whereas the overcount parameter $\lambda$ estimated in (7) is defined only with respect to the number of eligible persons.

### *2.2.2. Sample design and illustrative examples*

The O-sample typically consists of all people who are enumerated by the census in the same EAs selected for the U-sample. Thus, the sampling units and stratification variables are defined according to principles similar to those described in Section 2.1.2. Operationally, data collection for the two samples is linked in the following schematic stages:

(1) U-sample data collection;
(2) Construction of U-sample list;
(3) Matching U-sample and O-sample lists; and
(4) Follow-up of unmatched cases (overcoverage-follow-up, OF).

Hence, the OF fieldwork is limited to people listed in the sampled EAs who were not linked to the U-sample list. These may be ineligible people, eligible people missed by the U-sample enumerators (U-sample undercount), or false nonmatches. Sample size and sample allocation take into account expected overcount patterns in addition to expected undercount patterns. Note that whereas the U-sample unit is a household or a dwelling, the enumeration unit in the OF is typically an individual.

The following is a description of the U.S. 2001 Census Accuracy and Coverage Evaluation and the Israeli Integrated Census paradigm for the 2008 census. Two examples illustrate estimators (6) and (7), respectively.

*2.2.2.1. U.S. census 2000 accuracy and coverage evaluation.* The ACE consisted of an undercount sample (the Population sample, or P-sample) and an overcount sample (the Enumeration Sample, or E-Sample). The PSU was a block cluster comprising about 30 housing units. A national sample was selected in three phases:

(a) In the first phase, block clusters within states were classified into three size strata (small, medium, and large) and a fourth American Indian Reservation stratum. The size of the block clusters was based on preliminary census files, and an initial sample of about 30,000 block clusters was selected. For these block clusters, lists of housing units were created by field work.

(b) In the second phase, the field-based lists and updated census address lists were used to substratify the first phase sample within the large and medium strata in each state (reduction strata). A subsample of block clusters was selected, with equal selection probabilities within second-phase strata and possible differences in selection probabilities across strata. The second phase sample consisted of 11,303 cluster samples with about 850,000 housing units.

(c) In the third phase, a subsample of housing units was selected in block clusters consisting of 80 or more housing units. This phase elicited a final sample of about 301,000 housing units from the 11,303 block clusters selected in the second phase. The respective E-sample list consisted of 311,000 housing units (approximately 700,000 people).

Data were collected for the P-sample by means of computer-assisted personal interviewing (CAPI). The P-sample list was matched to the census list in the sampled blocks or in adjacent blocks, using computerized or computer-assisted clerical matching procedures. All the unresolved cases in the P- and E-samples (e.g., nonmatches, possible matches) were sent for follow-up interviews. About 50,000 people were included in the P-sample follow-up and about 143,500 people in the E-sample follow-up (for a detailed account, see U.S. Census Bureau, 2004).

The form of the DSE used in each estimation poststratum of the ACE was essentially as follows:

$$\tilde{N} = Z \times \frac{\hat{Y}^E_{1+}}{\hat{Z}^E} \times \frac{\hat{Y}^P_{+1}}{\hat{Y}^P_{11}} \times \phi, \tag{9}$$

where $Z$ is the total number of census enumerations, $\hat{Z}^E$ is the estimated number of census enumerations from the E-sample, $\hat{Y}^E_{1+}$ is the estimated number of eligible enumerations from the E-sample, $\hat{Y}^P_{+1}$ is the estimate of the total population from the P-sample, $\hat{Y}^P_{11}$ is the estimated number of enumerations from the P-sample that match to the census, and $\phi$ is a correlation bias adjustment factor applied for male adults. All four E-sample and P-sample estimators in (9) are basically expansion estimators adjusted for missing data and other operational problems. The middle expression in (9), $\hat{Y}^E_{1+}/\hat{Z}^E$, is the census overcoverage correction factor, while the last expression, $\hat{Y}^P_{+1}/\hat{Y}^P_{11}$, is the census undercoverage correction factor (see, e.g., Mulry, 2007).

*2.2.2.2. The integrated census paradigm planned for the 2008 Israeli census.* The basic idea of the Integrated Census (IC) is to replace the traditional nationwide field enumeration with an "enumeration" of the Population Register (PR) augmented by survey data for estimating and adjusting for coverage errors. Note that enumeration of the PR means collecting the data from the PR files. Estimates of the coverage parameters are obtained through two coverage surveys: The U-survey is based on an area sample and provides estimates of undercount rates, and the O-survey is based on a sample of people from the PR and provides estimates of overcount rates. Notably, U-sample enumeration is "blind" to the PR. Thus, the enumerators do not know if and where a person in their area is listed in the PR.

In the Israeli administrative-statistical system, the country is divided into statistical areas (SAs), which comprise 3000–4000 residents on the average. The aim of the IC is to provide population estimates by age and sex subgroups within SAs. In preparation for the IC, SAs are divided into EAs, which include about 50 households each on the average. All PR records are geocoded and clustered into the above-mentioned EAs and a random sample of EAs is then selected within each SA. The U-sample comprises all eligible people who live in the sampled EAs, and the O-sample includes all people who are listed in the PR in the same EAs.

The planned sample for the first IC will comprise about one-fifth of the population (about 400,000 households). However, the sampling fraction within SAs will vary in accordance with accuracy requirements and on the level of the coverage parameters. Sample allocation is basically extracted from the variance estimator (8), and estimates of the coverage parameters are plugged in by matching the previous census with the PR, as well as on the basis of other demographic data (e.g., percentages of children, young people, elderly people, religious people, and new immigrants—for further details, see Nirel et al., 2003).

The U-sample fieldwork will start one day after Census Day and will last from four to six weeks. The U-sample file will then be matched to the PR and a list of the people remaining in the O-sample EAs will be created, which includes the people in the O-sample file but not those in the U-sample file. To complete the O-sample fieldwork, all the people in the remainder list (the overcoverage-follow-up sample) will be traced and interviewed to determine their status. The OF-sample is expected to include approximately 20% of the O-sample on average.

### *2.3. Other approaches to estimation of coverage errors*

Several countries evaluate coverage errors through PESs, which provide direct estimates of overcount and undercount rates by usual weighting methods rather than through DSE-based estimates. A prime example of this approach is the Canadian Coverage Error Measurement Program. A traditional census is conducted in Canada every five years, with modifications in the data collection methodology introduced in each cycle. Thus, in the 2001 census, questionnaires were distributed by enumerators, completed by household members, and returned by mail (mail out). In the 2006 census, an additional online option was offered for completion and return of the census questionnaire. The Canadian evaluation program will be illustrated below on the basis of the 2001 census. The program comprised the following four studies (Statistics Canada, 2004a):

(1) The Dwelling Classification Study (DCS), which focused on undercount due to misclassification of dwellings as unoccupied and due to nonresponse;
(2) The Reverse Record Check (RRC), which estimated total undercount and overcount that was not included in the other studies (AMS and CDS below);
(3) The Automated Match Study (AMS), which focused on overcount of people who were counted more than once within the same area; and
(4) The Collective Dwelling Study (CDS), which dealt with overcount of people who are counted in noninstitutional collective dwellings as well as in private dwellings.

Due to lack of space, we provide here a brief description of these samples (for a detailed account, see Statistics Canada, 2004). The main features of the above samples are highlighted in Table 2. In the DSC, enumerators returned to unoccupied and nonresponse dwellings in the sampled EAs in an attempt to determine whether or not the dwellings had been occupied on Census Day. The estimates from this survey were the only ones that were fed back into the census database. Approximately 223,000 people were added to the census database.

The RRC was sampled independently from six sampling frames that cover the entire census target population. The total sampling fraction was approximately 0.2% although the sampling fractions varied between and within frames. The main frame was the 1996 census (providing 74% of the total RRC sample), and the other frames included people who had been missed in the 1996 RRC, as well as files of births, immigrants, nonpermanent residents, and health care beneficiaries. An effort was made to identify people who appeared in more than one frame. The 1996 census frame was stratified by province, gender, age, and marital status. Sample allocation was based on past coverage and tracing rates, which yielded higher sampling fractions for strata with a high percentage of hard-to-count people. An intensive tracing operation provided telephone numbers and other contact details for approximately 50% of the people in the RRC sample, and data were collected primarily by computer-assisted telephone interviewing. Notably, this study was completely independent of the census operation.

Table 2
The main features of the sample design in the Canadian 2001 Census Coverage Error Measurement Program (for a detailed description, see Statistics Canada, 2004)

| Study | Target Population | Sample Size | Sample Design |
|---|---|---|---|
| Dwelling Classification Survey (DSC) | Nonresponse and unoccupied dwellings | 1399 enumeration areas | Urban: single-stage stratified sample<br>Rural: Two-stage sample |
| Reverse Record Check (RRC) | People who should be enumerated | 60,653 persons | 1996 Census: single-stage stratified sample with varying sampling rates |
| Automated Match Study (AMS) | Matched pairs of households within the same region | 17,275 pairs of households | Single-stage stratified sample |
| Collective Dwelling Study (CDS) | Usual residents in noninstitutional collective dwellings | 4500 residents | Single-stage stratified sample. Sample size was related to population size |

Although the last two studies did not involve field operations, they are described here to provide a complete presentation of the program. The objective of the AMS was to identify duplicate households within the same geographic region. Automated record linkage identified exact and near matches, and the pairs were then stratified by variables such as geographic proximity and level of similarity. A questionnaire review was conducted for a random sample of pairs. Finally, the CDS reviewed questionnaires completed by a sample of usual residents in noninstitutional collective dwellings who had reported an alternative address on their census form.

For each study, weighted estimates of undercount and/or overcount were calculated. These estimates were combined arithmetically to provide the overall estimates. For instance, the estimated undercount and overcount rates for the 2001 Census were 0.0395 and 0.0096, respectively, yielding a net undercount rate of $\hat{D}/\hat{N} = 0.0299$ (SE = 0.0014).

## 3. The use of sample surveys to evaluate statistical adjustment of census counts

The coverage sample surveys discussed in the previous section estimate the bias in census enumeration due to undercount and overcount. In many countries, the primary use of these postenumeration surveys (PESs) has been to evaluate the quality of the census and to gain insight into coverage issues for future censuses. In several countries, however, such surveys are considered or even used to improve the accuracy of the census counts by statistical adjustment. In the United Kingdom, for example, the 2001 ONC comprised full enumeration as well as a PES, which was known as the CCS (Section 2.1.2), and covered a sample of approximately 370,000 households. Using DSE, census counts at the local authority district level were adjusted for undercount (Office for National Statistics, 2000).

Another interesting example is the U.S. ACE survey, which was conducted following the 2000 Census. The idea that census counts may be adjusted by sample survey data led to some controversy regarding the usefulness of such a procedure. Freedman and Watcher (2003, 2007) argue that "error rates in the adjustment are comparable to if not larger than errors in the census." In 2003, the Census Bureau decided not to use the ACE results for the population base of the intercensal estimates due to "technical limitations" of the ACE Revision II estimates (Mulry, 2007; U.S. Census Bureau, 2003b).

This section deals with evaluation of adjusted census counts. The primary aim of such evaluation is to assess whether potential errors that may be introduced by a PES are small enough so that its use for adjustment will improve the census accuracy. Some error components can be incorporated in the adjusted estimates. For example, when dual system models are used, correction for correlation bias can be included in the adjusted estimates through a correlation bias factor (see Section 2.1.1). This correction was applied in the U.S. ACE (Bell, 2001; Mulry, 2007), as well as in the U.K. CCS (Brown et al., 2006). Model and sampling errors can be expressed through variance estimations, as in the Israeli IC (Glickman et al., 2003; Section 2.2.1).

However, the adjusted estimates can still be subject to biases resulting from errors in data collection and data processing. Fieldwork in PESs attached to traditional censuses may take place a few weeks or even a few months after census day. This may lead to recall problems, for example, it is possible that people will not remember the exact date they moved from their census day address or that they will even have problems remembering birth or death dates around census day. For other types of censuses such as a register-based censuses combined with coverage surveys, fieldwork for the undercount survey may start on census day. However, differences between register variables and field variables, as well as differences in data collection procedures (administrative versus fieldwork), can introduce new and additional types of errors. In light of that situation, the following section focuses on some of the error components that require additional evaluation.

We start by analyzing known and unknown counts at the end of the data collection process in a census that comprises full enumeration and coverage surveys. We then discuss possible errors in the adjusted estimates and suggest additional data collection with the objective of evaluating the adjusted estimates. We conclude the section by describing evaluations of coverage estimates based on sample surveys that were carried out by the U.S. Census Bureau after the 1990 and 2000 censuses.

### *3.1. Known and unknown counts in the DSE and extended DSE paradigm*

The first step is to analyze the various pieces of information provided by a full enumeration and coverage surveys. This analysis will highlight the missing information and help to design a program for evaluating potential biases in the final estimates. Table 3 summarizes the counts obtained by an extended DSE paradigm. We extend the subscript notation of Section 2 to include a third data source and apply it to eligible counts $Y$ and ineligible counts $X$. Thus, $Y_{101}$ denotes the number of eligible people counted in the full enumeration, who were not on the undercoverage list but were in the OF. Similarly, $X_{100}$ is the number of ineligible people who were counted in the full enumeration but not in either of the two surveys. In contrast to the previous section, we do not assume here that the undercoverage survey is free from overcount. Let us follow the stages in data accumulation for this paradigm, which are given as follows:

(1) *Full enumeration.* The first step obtains a full enumeration list, either through field collection or administrative "collection." At this stage, we observe $Z$, which includes eligible and ineligible people. We assume that this list is incomplete and note that $Z = Y_{11} + Y_{101} + Y_{100} + X_{11} + X_{101} + X_{100}$. However, the break down of Z is not known at this stage.

Table 3
Data summary for a sampled area in an extended DSE paradigm, including an undercoverage and an overcoverage-follow-up surveys

| | | *Undercoverage Survey List* | | | | | | *Total* |
|---|---|---|---|---|---|---|---|---|
| | | ***Eligible*** | | | ***Ineligible*** | | | |
| | | Counted | Missed | | Counted | Missed | | |
| | | | *OF Survey List* | | | *OF Survey List* | | |
| | | | Counted | Missed | | Counted | Missed | |
| *Census List* | Counted | $Y_{11}$ | $Y_{101}$ | $Y_{100}$ | $X_{11}$ | $X_{101}$ | $X_{100}$ | $Z$ |
| | Missed | $Y_{01}$ | $Y_{00}$ | | $X_{01}$ | – | | |

*Note*: Shaded cells indicate unknown counts.

(2) *Undercoverage survey.* Once the survey data are collected and matched to the full enumeration, we obtain (a) the count of matches $Y_{11} + X_{11}$; (b) those in the full count and not in the survey $Y_{101} + Y_{100} + X_{101} + X_{100}$; and (c) those who participated in the survey but not in the full list $Y_{01} + X_{01}$. The additional information we seek, namely the omissions of the full enumeration ($Y_{01}$), is included in the sum $Y_{01} + X_{01}$. However, this information is masked by erroneous survey enumerations (e.g., fabrications) $X_{01}$. In addition, $Y_{00}$ is unknown. Note that although in theory $X_{11}$ may be positive, that is, both the full enumeration and survey do not identify a person as ineligible, it is believed to be either equal to zero or negligible.

(3) *Overcoverage-follow-up survey.* Data collection and matching of OF data with previous data shed some light on omissions in the undercoverage survey through $Y_{101}$. Respondents in the OF also provide information on ineligibles through $X_{101}$, but the remaining uncertainty due to OF survey nonresponse is still included in $X_{100}$. Actually, the sum $Y_{100} + X_{100}$ is known, but the break down of eligible and ineligible respondents is unknown. To complete the table, we note that by definition, $X_{000} = 0$ because the OF list is extracted from the census list.

In sum, after completing the data collection and matching, the unknown counts that remain are $Y_{00}$ and a break down of $Y_{01} + X_{01}$ and $Y_{100} + X_{100}$ (see shaded cells in Table 3). The other counts in the table are considered "known." Note that a DSE paradigm that does not include an OF survey is a special case, with $Y_{10} = Y_{101} + Y_{100}$ and the appropriate $X$'s equal zero.

### 3.2. *Error components*

The second step in developing an evaluation framework is to analyze the sources of missing data, on the one hand, and the sources of errors in the known counts, on the other. Table 4 summarizes typical sources of errors leading to undercount and overcount. It is important to note that the types of errors and their relative weight depend on the country, on the census methodology, and on data collection methods.

We start by discussing errors pertaining to the unknown counts (see Table 3). The main sources for these are nonresponse in the OF survey affecting the break down of $Y_{100} + X_{100}$, nonresponse and inadvert omissions of dwellings in the undercoverage survey affecting $Y_{00}$, and erroneous enumeration of buildings that do not belong to a sampled area, as well as fabrications that affect the break down of $Y_{01} + X_{01}$. Another group of errors is those caused by model biases, when missing data are imputed using a coverage model (e.g., DSE) and nonreponse imputation models.

Other errors relate to "known" counts in Table 3. Those are subject to measurement and matching errors. One group of measurement errors deals with census eligibility in the national and local dimensions. People who do not belong to the census population (ineligible) may be erroneously counted. Such cases include visitors who came to the country for a limited period (e.g., for less than three months) and happen to be present on Census Day. Similarly, there are eligible people who might not be counted, for example an illegal work emigrant who has been living in the country for more than a certain period (e.g., for over one year). Errors with regard to Census Day Eligibility also include babies born after Census Day who were counted in the census or people who died after census day and were not counted (see Fig. 1).

Table 4
Typical sources of undercount and overcount in adjusted estimates

| Error Type | Sources of Undercount | Sources of Overcount |
|---|---|---|
| *Unknown cells* | | |
| Nonresponse | Closed dwellings | |
| | U-survey and OF survey refusals and noncontact | |
| Misclassification | Erroneously classified as an empty or nonresidential dwelling | Erroneously classified as a residential or noninstitutional dwelling |
| Geocoding/address | Buildings or dwellings erroneously deleted by enumerators from area | Buildings or dwellings erroneously added by enumerators in area |
| Other | | Undercoverage survey fabrications |
| Modeling | DSE and Extended DSE assumptions not met | |
| | Nonresponse imputation errors | |
| *“Known” cells* | | |
| National eligibility | Forgotten | Counted |
| | Long-term visitors | Short-term visitors and tourists |
| | | Citizens living abroad |
| | Born before Census Day | Born after Census-Day |
| | Died after Census Day | Died before Census Day |
| Local eligibility | | Counted more than once |
| | Students at parents’ home (depending on definition) | Students in dormitories and in parents’ home |
| | Not counted | |
| | Outmovers after Census Day | Inmovers after Census Day |
| | People with multiple residences | People with multiple residences |
| | Children in shared custody | Children in shared custody |
| | No permanent address | |
| Geocoding/address | Erroneously omitted from maps of area | Erroneously geocoded to area |
| Other | | OF fabrications and multiple response |
| Matching | False nonmatches due to insufficient data | False matches |
| | False nonmatches due to geocoding to wrong area | |

Census residence is another difficult eligibility issue. At the time of the PES interview, some people may have already moved away from their census address (outmovers) and others may have moved in. Since spatial accuracy is a key issue in censuses, incorrect records of census residency can lead to substantial bias. Part of the problem for outmovers is that information about them is obtained by proxy (e.g., from neighbors) and may thus be unreliable. Other common problems in measuring census residency pertain to short-term visitors, people with vacation homes, college students, and children in shared custody. All these problems can result in erroneous enumerations (people counted in two locations) or in omissions (not counted in any location).

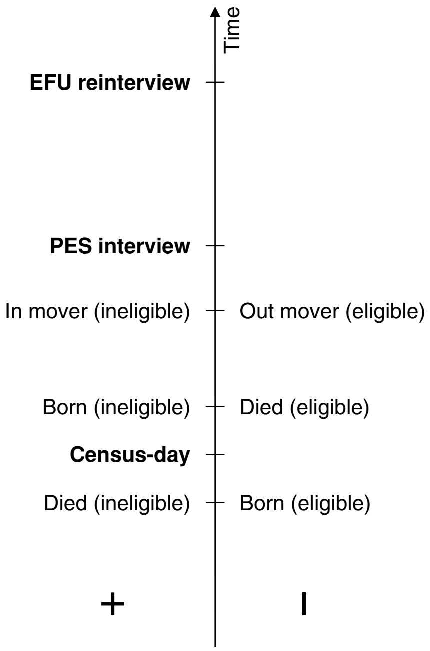


Fig. 1. Schematic illustration of eligibility errors resulting in erroneous enumerations (+) and omissions (−).

Another group of errors relates to measurement of geographical locations. In field operations, enumerators can mistakenly classify a dwelling as empty or nonresidential. Alternatively, there may be errors in the geographic association of a building to an EA (geocoding): buildings can mistakenly be added or omitted from an EA because of erroneous geocodes. Countries that do not have a register of buildings or dwellings are particularly susceptible to geocoding errors. However, address and building registers are also susceptible to risks such as duplication (e.g., a building on the corner of two streets can be entered with both addresses) or omitted addresses. The last group of measurement errors relates to fabrication and multiple response (mail, interviewer, internet), which results in erroneous enumerations.

Finally, record linkage procedures are used to produce the DSE and extended DSE tables (Winkler, Chapter 14). These procedures are applied to identify the people who were enumerated in both the census and the survey, or who were enumerated in one and not in the other. Two types of error can occur when this procedure is used. The first type of error can be made when people enumerated in both the census and the survey are not linked, and results in false nonmatches. The second type of error can be made when people are erroneously matched and results in false matches. Linkage errors can result from insufficient matching information, erroneous data leading to false nonmatches, imputation errors, or geocoding errors.

An important example of error component analysis is the U.S. Census Bureau total error model for PES estimates (Hogan and Wolter, 1988; Mulry and Kostanich, 2006; Mulry and Spencer, 1991). The model attempts to estimate systematic errors remaining after adjustment. It incorporates modeling, sampling, data collection, measurement, and processing errors. The idea is to break the overall bias of the empirical census estimate into major error components and to try and estimate each component separately. Specifically, the model includes components such as matching errors, errors in census address reporting, fabrication errors, errors in measuring erroneous enumerations, correlation bias, and sampling errors. For an interesting analysis of potential problems in a PES, see also Hogan (2003).

### *3.3. Evaluation follow-up*

The last step is to design the required operations. Some evaluation operations, such as expert clerical reviews and analyses based on demographic estimates and administrative data, can be carried out in the office. Evaluation of errors such as Census Day residence and missing data require a reinterview. We focus here on evaluation components that are carried out by means of additional sample surveys, which will be referred to as EFU. Notably, reinterviews are only useful if they provide more accurate data than the census and the PES. Therefore, when we consider such an operation, it is important to balance the need for information with the plausibility of obtaining accurate and consistent data. A key feature is a questionnaire with specific questions, which aim to elicit more accurate recall and understanding of information relevant to determining all dimensions of census eligibility.

Depending on the census methodology and data collection methods, a system of EFU surveys can be designed to evaluate the expected errors. Those surveys may include the following:

*Dwellings/households sample.* The first proposed operation is to conduct reinterviews in a sample of dwellings that were classified by the PES as empty, closed, or nonresidential, as well as in a sample of dwellings with residents who were not linked to the census enumeration or who were linked with low probability. This sample is intended to add information on PESs omissions, as well as to correct classification errors and false nonmatches.

*Buildings/address sample.* Another possible operation involves buildings that were deleted from a sampled area or added to it by a PES enumerator, as well as buildings with residents who were interviewed at an address that differs from the census address. This sample can shed light on geocoding errors, as well as on errors in census residence, and errors in address registers.

*Sample of individuals.* Finally, a sample of individuals from the census list, including OF nonrespondents, nonmatches within a household (especially students and babies) or other nonmatches (e.g., people who reported no change of address), and people in hard-to-count groups such as young male bachelors, residents of institutions, and households with members who have more than one place of residence or households comprising a divorced/separated parent and children. This sample attempts to supplement information on OF missing data, Census Day recall problems, matching errors, and fabrications.

To conclude this section, we will describe some evaluation operations and error component estimation that were carried out by the U.S. Census Bureau after the 1990 and 2000 PESs. These evaluation studies were mostly aimed at assessing errors in particular aspects of the PESs operations and estimation. They were not intended or designed to provide corrections to the undercount estimates. The first net undercount estimate derived from the 1990 PES was 2.1%. A sample of whole households and partial household nonmatches in 919 block clusters and a sample of matches from the same blocks were selected to assess P-sample data collection errors. Moreover, in the E-follow-up survey, interviews were conducted among the same 919 block clusters to assess E-sample errors. The results showed that the net error in the DSE estimate was approximately 0.5% (Mulry and Spencer, 1993). Based on the results presented by Mulry and Spencer (1993, Table 1), the main error components that reduce the undercount estimate are matching error (0.21%), P-sample collection error (0.31%), E-sample operations error (0.25%), and ratio-estimator bias (0.11%). The main components that increase the estimate were E-sample collection error (−0.17%) and model bias (−0.29%). The contribution of fabrication errors, imputation errors, and sampling errors to bias in the 1990 census was negligible. The revision of the 1990 PES estimates involved developing a new poststratification, redoing matching for 104 influential block clusters, and correcting two computer processing errors that affected the estimation of erroneous enumerations (Hogan, 1993).

The ACE survey that followed the 2000 Census provided three estimates of net census undercount. The original net undercount estimate published in March 2001 was approximately 1.2%. Evaluation of this estimate, which involved reinterviews and/or rematching of approximately 1/10 of the P- and E-samples, is known as the EFU. The EFU evaluated Census Day residency by reinterviewing people who were included in the P-sample and people who were included in the person follow-up (PFU) samples with unusual living situations, moving status, etc. The ACE Revision Preliminary Estimate published in October 2001 indicated a net undercount of 0.06% (Thompson et al., 2001). However, the final ACE Revision II Estimate published in March 2003 revealed a net undercount of −0.5% (i.e., a net overcount of 1.3 million, Mulry, 2007). The Revision II evaluation was motivated by the first revision evaluations but involved additional work. It revealed that the reduction of 1.7% in the net undercount was mainly due to census duplications that fell in the E-sample but were not detected as ineligible (−1%), E-sample coding corrections (−0.9%), P-sample duplications (−0.4%), and correlation bias (+0.6%, U.S. Census Bureau, 2003a, p. 31, Table 12). Duplicates were identified by a nationwide search rather than by the limited area search conducted earlier. Coding corrections pertain to conflicting addresses provided by the PFU and the EFU. Other errors included dwellings registered under two different addresses in the address file. In sum, the post-ACE evaluations showed that the main failure of the ACE was inaccurate measurement of the Census Day residence.

## 4. The use of sample surveys for carrying out a census

The search for new census-taking methods has yielded an entirely different type of census, which is based on continuous cumulative sample designs and utilizes the principles of rolling sample design (see Section 4.1 below). The main advantage of a sample-based census is that it provides more frequent and timely estimates of large national domains. Such estimates supply information on temporal variation, in contrast to the "once in ten years" census. The main drawback of the sample-based census is that it does not provide a detailed geodemographic "snapshot" on a particular date so that comparisons between domains are much more complicated. Furthermore, the issue of coverage becomes much more complicated in a census that is rolled over time because of the population movements in time and space. Thus, the likelihood of coverage errors may increase substantially in a sample-based census (for further discussion of the merits and drawbacks of sample-based censuses, see Office for National Statistics, 2003).

### 4.1. Rolling samples and some extensions

The concept of a rolling sample was proposed and developed by Leslie Kish in a series of papers (e.g., Kish, 1998). A rolling sample design jointly selects a set of $k$ mutually exclusive (not overlapping) periodic samples, each of which is a probability sample with a fraction $f = 1/F$ of the entire population. One sample is interviewed at each time period, and the accumulation of $k$ periods yields a sample with a fraction $f' = k/F$. The main aim of the rolling sample design is to provide detailed estimates in temporal as well as spatial dimensions. Specifically, Kish emphasizes the need for adequate annual estimates at the national and major regional/domain levels. By keeping the samples mutually exclusive, maximum efficiency of the accumulation is attained and the estimation procedure is simple. Note that the rolling sample design assumes that each sample is a representative sample of the relevant regions and domains. This means that if the PSUs are clusters, they should be smaller than the target domains. Moreover, the PSUs should form a rolling sample themselves. Therefore, a design in which all samples include the same PSUs, with a rolling sample within a PSU, has been referred to by Kish (e.g., 1999) as a *cumulated representative sample* (CRS) design. This design is not strictly a rolling sample.

Kish (e.g., 1990, 1999) further extended the concept of a rolling sample to a *rolling census* by taking $k = F$, that is, a sample with a cumulative fraction $f' = F/F = 1$. Thus, the rolling census replaces the simultaneous and complete enumeration of the population, carried out once in several years, by a continuous cumulative sample survey that covers the entire population over a time period $F$. For example, a rolling census comprising of 10 annual 1/10 samples yields full coverage of the population after 10 years. A rolling census can provide national and large domain estimates each year based on the latest sample, as well as smaller domain estimates based on appropriate accumulations of a number of samples.

The basic rolling census design is defined by the number of samples $F$ and by the choice of sampling unit. The number of samples is likely to coincide with the "regular" intercensal periods of 5 or 10 years. The natural sampling unit is a small, well-defined geographical area. Some modifications of the basic design can include unequal sampling fractions in different strata as well as some overlap between samples (Kish, 1998). Because of the complexity involved in conducting a census, it is expected that a rolling sample design will be combined with a CRS design or with other panel designs (Kalton, Chapter 5). For example, small local authorities can be sampled by a rolling design, whereas large local authorities can be included in all samples, with a rolling sample of addresses within them. The actual design is clearly determined by the census objectives, as well as by numerous statistical and nonstatistical constraints. A qualitative analysis of alternative designs is presented in the Office for National Statistics (2003).

### 4.2. An illustrative example: the French rolling census

At the time of the writing of this chapter, France is the only country that decided to carry out a sample-based census (e.g., Dumais et al., 1999; Isnard, 1999). The first sample was interviewed in 2004 and the first census will be completed in 2008. Because this is the only actual example of a sample-based census, we will describe its general design and estimation procedure.

The national statistical agency of France, INSEE, has decided that to comply with the requirement of timeliness on the one hand and budgetary constraints on the other, the census will enumerate 5/7 of the population over a five-year period (Durr, 2004; Durr and Dumais, 2002). The sampling unit is a "commune," which is a subdivision of a French territory and a local authority. The size of a commune varies from several dozen residents to over 300,000 residents. Therefore, the communes are first stratified by size. The *large communes* stratum includes all communes with 10,000 or more residents, and the s*mall communes* stratum comprises all communes with less than 10,000 residents.

A two-stage annual sample is selected as follows: in the large communes stratum, all communes are included in the sample, and 0.08 of the dwellings are selected for enumeration. In the small communes stratum, approximately 0.20 of the communes are sampled, and all dwellings in the sampled communes are included in the sample. Thus, large communes are surveyed every year and small communes are surveyed once during a five-year period. Because the small communes comprise approximately 50% of the population, the annual sampling fraction is about 0.14 ($0.5 \times 0.08 + 0.5 \times 0.2$) and the accumulated five-year fraction is about 0.70, as required.

The sample design in both strata is controlled by a multiannual rotating scheme. The large communes sample is drawn from the "inventory of located buildings" (RIL) list. The addresses in each large commune are divided into five balanced groups that have a similar distribution with regard to variables such as age, gender, and type of dwelling, based on the 1999 population census data. For year $i$, $i = 1, \ldots ,5$, a sample of addresses in the $i$th group is selected. All dwellings within a sampled address are enumerated. The sampling fraction of an address within group $i$ is approximately 0.40 so that 0.08 of the addresses (and hence of the dwellings) in a commune are sampled annually. The RIL is updated every year to account for new and demolished buildings, and the five sample groups are updated accordingly.

Small communes are divided into five representative groups within each of the 22 regions of France based on the same variables as those of the large communes. During year $i, i = 1, \ldots, 5$, the sample of small communes is comprised of the $i$th group in each region.

In sum, the French design is referred to as a "rolling census" but it covers approximately 70% of the population over a five-year period. The design extends Kish's ideas by combining a CRS design for the large communes with a rolling sample of small communes.

### 4.3. Estimation

The census design defines a feasible resolution of estimates in time and space. Specifically, estimation for small domains can involve a combination of direct (design-based) estimates and synthetic (model-based) estimates, possibly using auxiliary information from additional sources.

Let $Y$ be the outcome of interest, with annual estimators $\hat{Y}_i i = 1, \ldots, F$ and $\hat{Y}(\mathbf{W}) = \sum_{i=1}^{F} W_i \hat{Y}_i$, a census estimator with $\mathbf{W} = (W_1, \ldots, W_F)$, $\sum_{i=1}^{F} W_i = 1$. Kish (e.g., 1999) considers several basic weighting schemes: (a) using the last year only, where $W_F = 1$ and all other weights equal zero; (b) averaging all years with equal weights, where

$W_i = 1/F, i = 1, \ldots, F$; and (c) monotonically nondecreasing weights—$W_1 \leq W_2 \leq \ldots \leq W_F$. Clearly, cases (a) and (b) are special cases of (c).

The example of the French census will be used to illustrate a typical system of estimates that can be produced by a sample-based census. Consider years $F$, $F$-1, $F$-2, $F$-3, and $F$-4. Three types of estimates are provided: (a) population counts for year $F$-2 for every commune; (b) small area estimates based on data collected over the previous five years and pertaining to year $F$-2; and (c) national and regional estimates for the current year $F$ (Durr, 2004).

### 4.3.1. Commune population count for year F-2

For every commune, regardless of the year it was surveyed, an estimated population count is provided at the end of year $F$ for the beginning of year $F$-2. For a large commune, let $X_i$ be an auxiliary variable (e.g., number of dwellings in the RIL) during year $i$, and $\overline{X} = \sum_{i=F-4}^{F} X_i/5$ the average number of dwellings over the last five years. Let $\hat{Y}_i$ be the expansion estimator for population count of that commune based on the data for year $i$, and $\overline{\hat{Y}} = \sum_{i=F-4}^{F} \hat{Y}_i/5$ will be the average estimate. Accordingly, the estimated count for year $F$-2 is the ratio (synthetic) estimate given by

$$\hat{Y}_{F-2} = \overline{\hat{Y}} \frac{X_{F-2}}{\overline{X}}. \tag{10}$$

For a small commune, the estimate depends on the year it was enumerated. Denote by $Y_i$ the population count of a commune that is fully enumerated in year $i$. For a commune surveyed in year $F$-2, the estimate is equal to its population count, $Y_{F-2}$. For communes surveyed prior to year $F$-2, the estimates are extrapolated using a ratio estimate similar to the one in (10), yielding

$$\hat{Y}_{F-2}^{F-4} = Y_{F-4} \frac{X_{F-2}}{X_{F-4}} \quad \text{and} \quad \hat{Y}_{F-2}^{F-3} = Y_{F-3} \frac{X_{F-2}}{X_{F-3}},$$

where the superscript in $\hat{Y}_{F-2}^{(\cdot)}$ denotes the actual survey year. For communes surveyed after year $F$-2, the estimates are obtained by interpolation. The first value used for the interpolation is the actual count on the year surveyed. The second value is an estimate for year $F$-3, obtained from extrapolation of the previous count (five years before). We obtain

$$\hat{Y}_{F-2}^{F-1} = \alpha_{F-1} Y_{F-1} + (1 - \alpha_{F-1}) Y_{F-6} \frac{X_{F-2}}{X_{F-6}} \quad \text{and}$$

$$\hat{Y}_{F-2}^{F} = \alpha_F Y_F + (1 - \alpha_F) Y_{F-5} \frac{X_{F-2}}{X_{F-5}},$$

where $0 \leq \alpha_i \leq 1\, i = F - 1, F$, and is typically no smaller than 0.5. These estimates are further calibrated (Kott, Chapter 25) to the national and other large-scale estimates (for other versions, see Durr and Dumais, 2002).

### 4.3.2. Small-area estimates for year F-2

Every year, a file containing data for the previous five years will be constructed, including a sampling weight for every person and dwelling. This file enables expansion estimation for any geodemographic subgroup, subject to accuracy limitations. These estimates are taken to pertain to the midpoint of the period, for example, at the beginning of year $F$-2. For large communes, the weight can be extracted from (10) and is equal to $\varphi X_{F-2}/\overline{X}$, where $\varphi$ is the inverse of the respective inclusion probability. For a small commune, the weight is $\hat{Y}^i_{F-2}/Y_i$, corresponding to the estimates for year $F$-2.

*4.3.3. National and regional estimates for current year*

Each annual survey is a representative sample comprising about eight million people. Hence, usual survey methods (e.g., expansion estimates) enable reliable national and regional estimates for the current year. These estimates are used to calibrate the commune and small area estimates.

## 5. Sample surveys carried out in conjunction with a census

Censuses originally focused on enumeration of people. However, as the demand for more detailed social and economic data increased, there was pressure to include additional variables in the census questionnaire. Nonetheless, there is a delicate balance between the length of the census form on the one hand and data quality and response burden on the other. To solve this problem, many countries use two types of census forms: a *short form* with a few (about 10–20) demographic variables, and a *long form* with comprehensive information on topics such as housing, employment, education, income, immigration, fertility, and disability. The short form is completed for all people in the census population, whereas the long form is typically completed for a random sample of the population. In that way, the response burden is reduced for the population at large, and the census machinery can be used to supplement information based on the detailed data collected from the population that filled out the long form. Because long-form data collection is carried out during the same time period as the census, it is considered to be one of the census outputs (United Nations, 2006).

Overall sampling rates for the long-form sample vary for different countries and range from 5 to 20% of the population. Because of the high sampling rate compared with current annual or panel surveys, the long-form sample provides detailed snapshot information. Another strong link between the long-form sample and the census is that calibration of sample estimates to the census counts is straightforward because the two forms share the same geodemographic data.

The simplest selection method is to sample every $l$th unit systematically (e.g., dwellings, households), where $l$ is the inverse of the sampling fraction. One in every five and 1-in-10 are common rates. Variable rates are also possible. In Brazil, for example, municipalities with more than 15,000 residents are sampled with a 1-in-10 rate, whereas smaller municipalities are sampled with a 1-in-5 rate. Another sampling scheme is to select geographical areas (clusters) and survey all households within a selected area. An example of such one-stage cluster sample is the undercoverage survey of the Israeli Integrated Census (Section 2.2.2). This survey aims to estimate coverage errors and collects the long-form data at the same time. The EAs are sampled within SAs, and sample allocation is predominantly determined according to a coverage model that typically yields differential sampling rates for different SAs.

### 5.1. An illustrative example: The U.S. Census 2000 long-form survey

As an illustrative example, we present some main features of the sampling and estimation procedures used in the long-form survey in the U.S. Census 2000. The sampling frame was the Decennial Master Address File, and the target overall sampling fraction was 1/6. This rate was achieved through systematic sampling with variable rates within four size strata. The rates ranged from 1-in-2 for the smallest size stratum with less than 800 housing units to 1-in-4 and 1-in-6 for the 800–1200 and 1200–2000 size strata, respectively, and 1-in-8 in the ≥2000 stratum.

To obtain sampling weights, weighting areas with at least 400 people were defined within the four size strata. These areas were generally similar to the census tabulation areas. The initial weight was the inverse of the sampling fraction, and the final weight was obtained by iterative proportional fitting methodology, otherwise known as raking (Kott, Chapter 25). The long-form estimates were calibrated to the census counts in dimensions such as household type and size, and race by gender and age groups. Each stage of the raking procedure was adjusted for one dimension and lasted until a predefined stopping criterion was attained. Hefter and Gbur (2002) indicate that the difference in percentages between the census counts and the weighted sample totals ranged from −5.09 to 6.84% for single race groups (see Table 2 in Hefter and Gbur, 2002).

It is interesting to note that the direct variance estimates were calculated by the successive difference replication (SDR) methodology, which takes advantage of the systematic sampling of housing units. For an estimated total $\hat{Y} = \sum w_j y_j$, the basic successive difference estimator (Wolter, 1984) is

$$\operatorname{Var}(\hat{Y}) = (1-f)\frac{n}{2(n-1)}\sum_{j=2}^{n}(w_j y_j - w_{j-1}y_{j-1})^2,$$

where $w_j$ is the final weight for person $j$ and $f$ is the sampling fraction. The replication version is based on replicate samples and is simple to use

$$\operatorname{Var}(\hat{Y}) = (1-f)\frac{4}{R}\sum_{r=1}^{R}(\hat{Y}_r(\mathrm{SDR}) - \hat{Y})^2$$

where $\hat{Y}_r(\mathrm{SDR})$ is the estimate for the $r$th replicate using the appropriate replicate weights (Fay and Train, 1995; Gbur and Fairchild, 2002). The variance estimates produced by the U.S. Census Bureau do not calibrate each replicate to the census counts. Schindler (2005) examined a raked-SDR method and argued that it best reflects the sample design and estimation procedure. According to Schindler, the raked-SDR produces smaller variance estimates than other methods such as SDR and Jackknife. Finally, design factors by state and size strata are calculated by comparing the SDR standard errors to simple random sampling standard errors based on a 1-in-6 sample.

### 5.2. Rolling the long-form survey

A relatively recent development of the long-form idea has been referred to by the UN as a “traditional census with yearly updates of characteristics” (UNECE, 2006). This type of census is based only on the short-form data, and the long form is replaced by a set of annual surveys, where socioeconomic data are collected continuously during the intercensal years. The census with yearly updates aims to provide more frequent information for small domains.

To date, the United States is the only country that has decided to use a large continuous survey to obtain data on the long-form topics on a regular basis instead of using the traditional census long form. Interest in intercensal information in the United States goes back to the 1940s (see Section 3 in Alexander, 2002) and provided the basis for the ACS. The ACS methodology was tested in nationwide surveys from 2000 to 2004. Full implementation of ACS then began in 2005. These tests were carried out separately from the census and provided data for comparison between the long-form survey and ACS estimates. The main benefits of the ACS compared to the long-form sample are timely and frequent estimates. Another benefit is higher data quality in terms of completeness of response. The main weakness is larger estimation errors due to factors such as smaller sample size and less accurate population controls for adjusting the survey weights (National Research Council, 2007).

The ACS basically uses a monthly rolling sample design (Section 4.1). Approximately 250,000 addresses are surveyed each month, corresponding to an average sampling fraction of $f = 1/F = 1/480$ (a total of approximately 120 million addresses). The survey uses $k = 60$ mutually exclusive monthly samples, yielding 5-year average estimates with approximately 1-in-8 sampling rates, as compared to the 1-in-6 long-form average rate. The ACS uses a systematic sample of addresses, and the sample is selected in two stages. In the first stage, a "super sample" is selected, using a constant rate in all strata, which equals the largest sampling rate required for any one stratum. In the second stage, samples from the first sample are selected to give the desired fraction for each stratum. This design simplifies the handling of stratum-specific sampling rates and their dynamics over time.

Five-year accumulations of ACS data provide products similar to those obtained by the census long form. Average annual estimates are provided for areas with over 65,000 residents, and average three-year accumulations are provided for areas with over 20,000 residents. More frequent estimates for small domains can be obtained using small area methods (Lehtonen and Veijanen, Chapter 31; Datta, Chapter 32), see Malec (2005). For multiyear estimates, the ACS accumulation of samples over the years approximately averages the annual estimates with equal weights, contrary to Kish's (1998) inclination to increase $w_i$ with $i$. For a comprehensive description of the ACS methodology, see U.S. Census Bureau (2006b), and for a discussion on alternative estimation approaches, see Breidt (2007).

## 6. Concluding remarks

Developments in survey methodology used for censuses can largely be attributed to technological progress, which influences data quality, timeliness, and the cost of direct data collection. The CAPI, use of personal computers, handheld devices, and global positioning systems (GPS) can greatly improve the accuracy of key census variables, in particular, eligibility topics. Advances in internet technologies and in encryption also enable online self-interviewing, which can improve response rates in the long run. It is beyond the scope of this chapter to address this topic and the interested reader can find an overview in UNECE (2006, Chapter II). Regarding data processing, we have mentioned that progress in record linkage capabilities allows census surveys to be matched with PES on a national level and, thus, reduces matching errors and duplication.

Another important topic that has not been covered in this chapter concerns estimation procedures for small subgroups. Although estimators of coverage errors for national and subnational counts were described in Section 2, of this chapter, PESs cannot provide direct coverage estimates for small groups due to sample-size limitations. To derive adjusted counts for small areas, estimates obtained by small-area techniques can be used (Datta, Chapter 32). Suppose that the coverage factor for a given poststrata is $1/(\tilde{p}_{1+} + \tilde{\lambda})$ (Eq. (7)). Assuming that there is homogeneous coverage within poststrata, a synthetic adjusted census count for a subgroup $g$ is given by $Z_g/(\tilde{p}_{1+} + \tilde{\lambda})$, where $Z_g$ is the census count for subgroup $g$. The small-area estimates are generally calibrated to the national and subnational direct estimates using methods such as the generalized regression (GREG) (Kott, Chapter 25) or prediction regression (PREG) (Bell et al., 2007). For further details on these procedures, see Brown et al. (1999), Dick and You (2003), Office for National Statistics (2000); see also U.S. Census Bureau (2004, Chapter 8).

Finally, we mention the issues of missing data on the one hand and multiple responses on the other. A unique feature of census-related surveys is the need to impute eligibility status at the national and local levels. In Israel, for example, 5–10% of the listings in the Population Register are emigrants, many of whom are not traced by the O-survey. Imputation models based on administrative data as well as on sample data estimate the propensity of a missing case to be an emigrant. At the other extreme, the census and PES might provide two different census addresses. In such cases, decision rules should determine which address is more likely to be the correct one. Another important census issue is household counts. Entire households can be missed or individuals within households can be missed. The ONC project, for example, developed an imputation methodology which provided a fully imputed census file that is consistent with the adjusted census counts. The methodology involved three steps: missed household imputation, missed persons within households imputation, and "pruning and grafting" of imputations to the adjusted census counts (Office for National Statistics, 2002; Steele et al., 2002; see also U.S. Census Bureau, 2004, Chapter 6).

We have presented two directions in the development of sample survey methodology for censuses. The first direction tightens the relationship between censuses and sample surveys and focuses on evaluation of the census counts, with or without adjustment. The second direction moves away from the "snapshot" census and involves large-scale continuous sample surveys that either replace the census altogether or collect detailed and timely socioeconomic data to supplement the census data. A third direction that may be pursued in the future is to shorten the census cycle from 5–10 years to 2–5 years by basing full enumeration on administrative data and adjustment through coverage surveys.

*Sample Surveys: Design, Methods and Applications, Vol. 29A*
ISSN: 0169-7161

DOI: 10.1016/S0169-7161(08)00022-9

22

# Opinion and Election Polls*

*Kathleen A. Frankovic, Costas Panagopoulos and Robert Y. Shapiro*

## 1. Introduction: the reasons for public opinion and election polling

Public opinion polls are widely used to learn about the political attitudes, voting, and other behavior of individuals, by asking questions about opinions, activities, and individuals' personal characteristics (e.g., Abramson et al., 2007; Asher, 2007; Erikson and Tedin, 2007; Glynn et al., 2004; Traugott and Lavrakas, 2004). Responses to these questions are then counted, statistically analyzed, and interpreted.

Historically, academicians and government researchers in the United States engaged in opinion polling have called themselves "survey researchers," many with interests in psychologically oriented attitude research (Converse, 1987). There is also a separate and more visible group of "pollsters," originally involved in commercial research and in journalism, whose poll results on political and social matters are reported widely in the news media (Converse, 1987; Frankovic, 1998; Moore, 1992; Rogers, 1949). Others conduct proprietary polling for political candidates, political parties, or other clients (Eisinger, 2003, 2005; Jacobs and Shapiro, 1995; Stonecash, 2003). Today, survey researchers and pollsters have become synonymous, although the polls they do can vary in their purpose, type, scope, and quality (see Chapter 21).

This chapter describes these aspects of opinion and election polls, focusing largely on the United States, but providing some international comparisons and discussions of similar aspects and uses of polling. It begins by reviewing the public and private uses of opinion and election polling. Next, it summarizes the general methodological issues in polling that require attention in doing public opinion research. It then examines the cases of preelection polls, "exit polls" on election days, and, briefly, postelection polling. The last sections consider other methods of interview-based opinion measurement and the challenges ahead for opinion and election surveys as interest in public opinion and polling continues.

* The authors are listed alphabetically. This article is dedicated to our late colleague and friend, Warren J. Mitofsky, who would not have been shy about critiquing what we take full responsibility for here.

### *1.1. Polling for public consumption*

Polling has been motivated by wide interest and curiosity about public opinion in general and especially about voting during election periods. This is clear in the history of "straw polls" described later. Ordinary citizens are curious about this; in turn, journalists who write about politics try to appeal to their audiences' interests in candidates and issues, especially in the latest elections. This has been apparent from the early days of informal "straw votes" or "straw polls" in the United States to the big expansion of national polling in the 1970s as valid and reliable surveys could be done by telephone (Blankenship, 1977; Nathan, 2001).

Contemporary opinion and election polls, especially those done for public consumption, have many sources. By the end of the 19th century, politicians, academics, market researchers, journalists, and government all had begun serious data collection. In many states local leaders kept "poll books" registering the preferences of every registered voter. These "poll books" had their greatest value in times of political movement and uncertainty. An 1880 Republican poll of 26,000 Indiana Civil War veterans showed that 69% would vote Republican that fall; by 1888, only 30% of those same individuals would (Jensen, 1971, p. 26). In 1886, the social survey movement began in England, and social welfare workers collected data on poverty, housing, and crime. Sociologists developed tools for measuring opinion. In the United States, newspaper market research departments were established and national advertizing campaigns for brands like Ivory soap became prominent in the 1880s. Political advertising on a mass scale soon followed. In 1888, the Republican National Committee placed presidential campaign ads on New York City streetcars for the first time, changing campaign tactics from persuasion to marketing (Jensen, 1971, p. 159). The U.S. Census first used an early version of the Hollerith computer card for storing and analyzing data in 1890.

Opinion and election polls also date from the 19th century in the United States. In 1824, straw poll counts appeared in partisan newspapers – along with suggestions that the public might not agree with political leaders in their choice of presidential candidates (Smith, 1990). In some cases, counts of candidate support were taken at public meetings. In others, books were opened for people to register their preference. Some newspapers praised the technique. The *Niles Weekly Register* said of a count taken at a public meeting, "This is something new; but an excellent plan of obtaining the sense of the people" (*Niles Weekly Register*, May, 1824).

In 1896, the *Chicago Record* sent postcard ballots to every registered Chicago voter, and to a sample of 1 in 10 voters in eight surrounding states. The *Record* mailed a total of 833,277 postcard ballots, at a cost of $60,000; 240,000 of those sample ballots were returned. The *Record* poll found Republican William McKinley far ahead of the Democrat, William Jennings Bryan. McKinley won; and in the city of Chicago, the *Record*s preelection poll results came within four one-hundredths of a percent of the actual election-day tally.

By the 1920s, even papers whose editorial pages were clearly partisan were apparently comfortable reporting straw polls that indicated their editorial choice in an election was losing the contest for voters. Between 1900 and 1920 there were nearly 20 separate news "straw polls" in the United States. The *Literary Digest* established a poll in 1916. In 1920, it mailed ballot cards to 11,000,000 potential voters, selected predominantly from telephone lists. In later years, car registration lists and some voter registration lists were added to the sampling frame. Although the *Digest* touted its polling as impartial and accurate, it was also tied to an attempt to increase the magazine's subscriber base.

These news straw polling operations involved outreach to as many groups as possible and included huge numbers of interviews (often conducted on street corners). In its 1923 mayoral election poll, the *Chicago Tribune* tabulated more than 85,000 ballots. In the month before the election, interviews were conducted throughout Chicago, and results published reporting preferences by ethnic group (including the "colored" vote), with special samples of street car drivers, moviegoers (noting the differences between first and second show attendees), and white collar workers in the Loop.

But the real emergence of preelection polls as we now know them came in the 1930s. In 1935, both George Gallup and Elmo Roper began conducting a new kind of news poll: Gallup for a consortium of newspapers, and Roper for *Fortune* magazine. The stated goals were both democratic and journalistic. Gallup co-authored a book called *The Pulse of Democracy* (Gallup and Rae, 1940), while *Fortune*'s editors in their very first poll report in June 1935 explicitly linked impartial journalism and polls: "For the journalist and particularly such journalists of fact as the editors of *Fortune* conceive themselves to be, has no preferences as to the facts he hopes to uncover.... He is quite as willing to publish the answers that upset his apple cart of preconceptions as to publish the answers that bear him out" (*Fortune*, 1935).

Gallup and Roper (as well as Archibald Crossley, whose polls for the Hearst newspapers began in 1936) interviewed only a few thousand adults, unlike the tens of thousands in the *Tribune*'s canvass or the millions answering the *Literary Digest* polls. However, samples were selected to ensure that regions, city sizes, and economic classes were properly represented.

Although not a true probability sample, they were far more representative than the larger street corner or postcard polls (cf. Chapters 1, 16, and 20). And in that first test in 1936, the so-called scientific polls successfully predicted a Roosevelt victory. In fact, Gallup not only predicted a Roosevelt victory, but also predicted the *Literary Digest's* mistake. The flaws in the *Literary Digest*'s procedures (using lists that in nearly every state were biased in favor of the economically better-off during an economic depression) are fairly obvious today. First, by almost always limiting the sampling frame to those owning telephones and automobiles, lower-income voters were excluded, even though by 1936 social class would matter more than state or region in a person's presidential choice. Second, only about two million of the *Digest*'s ten million or so postcard ballots were returned, limiting the polling count to those who both received a questionnaire and bothered to respond. This led to a "selection bias" and the *Digest* overestimated support for Roosevelt's opponent, Republican Alfred Landon (see Squire, 1988). Although the new pollsters embarrassed the *Digest*, they too underestimated Roosevelt's margin of victory, suggesting biases in their methods as well. Those biases recurred in 1940 and 1944, foreshadowing the polling debacle of 1948 (see below and Converse, 1987).

By May 1940, 118 newspapers subscribed to the Gallup Poll. *Fortune* surveys appeared monthly. Between 1943 and 1948, at least 14 state organizations were conducting their own polls, using methods approximating those of the national pollsters (Frankovic, 1998).

The very first question asked in the first Gallup Poll was: "Do you think expenditures by the government for relief and recovery are too little, too great, or about right?"

Totally 60% of respondents said "too great," which was similar to what Americans thought about spending on "welfare" in the National Opinion Research Center's General Social Survey more than 50 years later (see Davis et al., 2005; Gallup, 1972, p. 1; Page and Shapiro, 1992, Chapter 3). By 1940, Gallup asked if the public approved of how President Franklin Roosevelt was handling his job, as well as which problems facing the country Americans believed were most important. These questions continue to be asked and are widely cited today.

Preelection polling began in other western democracies in the period between the wars and expanded afterwards (early European survey data are cited in Cantril with Strunk, 1951). There was a Gallup Institute in Great Britain beginning in 1938, and one in Canada starting in 1935. In Britain, "Mass Observation," which collected information about everyday life in Britain from 1937, used a different approach, creating a national panel of volunteers to reply to regular questionnaires (Hubble, 2005).

As democracy spread, so did polling and preelection polls. After the surrender of Japan, the U.S. occupying forces instituted public opinion polling, and the techniques established in the United States were adapted to accommodate at least some Japanese traditions (Worcester, 1983).

Wendell Willkie, an American businessman, used market researchers in his unsuccessful 1940 presidential campaign. Franklin D. Roosevelt was the first president to receive ongoing polling information – though from a distance, compared to later presidents – from Hadley Cantril, whose Office of Public Opinion Research at Princeton University had both an academic base and collaborated with the Gallup Organization. Roosevelt's use of polls during the period leading to the United States' entry into the war showed how polls provide strategic information for leading – or manipulating – public opinion (Page and Shapiro, 1992, Chapters 5 and 9).

At the same time, government agencies expanded their use of survey methods and large-scale polling beyond normal Census operations, beginning with the Department of Agriculture in 1939 (see Chapter 18). Survey research was used for public policy and public administration purposes, in attempts to improve wartime agricultural policies, control prices, understand race relations, stimulate war bond drives, measure the morale of civilians, and the well-being, outlook, and opinions of members of the armed forces (Converse, 1987; Gosnell and David, 1949; Stouffer et al., 1965 [1949]; Truman, 1945; Wallace and McCamy, 1940). The State Department, unknown to Congress, had the National Opinion Research Center (NORC, the first independent national survey research center, see below) survey the public's opinions toward American foreign policy during the Cold War (Eisinger, 2003; Foster, 1983; Page and Shapiro, 1992).

University-based surveys had their beginnings in 1939–1940, when Columbia sociologist Paul F. Lazarsfeld founded the Office of Radio Research, later renamed the Bureau of Applied Social Research. Lazarsfeld et al. conducted the first sophisticated survey study of decision-making in presidential elections in 1940 in Erie County, Ohio, followed by a study of the 1948 election in Elmira, New York (Lazarsfeld et al., 1944; Berelson et al., 1954). Two large and widely known research centers developed the long-term capacity to do national surveys: NORC, established in 1941 by Harry Field at the University of Denver and moved to the University of Chicago in 1947; and the University of Michigan's Survey Research Center (SRC, Institute for Social Research [ISR]), established in 1946 under the direction of Rensis Likert and Angus Campbell, who had been involved in government survey research.

Later, many other universities established their own survey research centers, especially after telephone surveying became more common (Sudman and Bradburn, 1987). The localized election studies begun at Columbia were subsequently continued on a larger national scale elsewhere. In 1948, Angus Campbell and Robert Kahn at the University of Michigan conducted the first national election study in the United States – or anywhere. This small study led to the long-term series of election studies conducted first by the Survey Research Center and the Center for Political Studies at Michigan, formalized as the American National Election Studies (ANES or NES) in 1977 with National Science Foundation (NSF) support, expanding control to a larger academic community. In 1972, NORC became the home of the ongoing, NSF-funded General Social Surveys (Davis et al., 2005).

At the outset, interest in polling was related to the original "democratic" philosophy associated with it – that individuals' opinions could be added up to constitute the will of the people that ought to have influence on political leaders and governing (Gallup and Rae, 1940). Critics argued that this simple definition of public opinion was misguided: public opinion as a concept and influence involved processes of purposive group interactions and communications that reached the attention of government, and in this process not all opinions counted equally (Blumer, 1948; Rogers, 1949). Ironically, the results of the early academic studies supported the views of the critics of polling. These studies showed that the public was not very knowledgeable and well informed about politics, was not influenced in expected ways by mass communication, and voted in ways related to seemingly mindless social characteristics and interpersonal influences. Partisan attachments were no more than equivalent to team loyalties (Berelson et al., 1954; Campbell et al., 1954, 1960; Lazarsfeld et al., 1944). The electorate voted neither on the basis of issues nor awareness of the competing political ideologies of the day (Converse, 1964). This spurred further debates about the capability, "rationality," and overall "democratic competence" of public opinion and the American electorate (cf. Althaus, 2003; Glynn et al., 2004, Chapter 8; Page and Shapiro, 1992; Zaller, 1992).

By 1948, pollsters had gained a national audience far beyond that of their press releases and print media subscribers. Elmo Roper gave weekly Sunday radio talks on CBS, and both Roper and Gallup were televised. The front-page headline of the *Washington Post* on Election morning read, "Dewey Deemed Sure Winner Today," although the last preelection Gallup Poll had given the Republican Thomas E. Dewey only a five-point lead over the incumbent Democrat Harry Truman (49.5–44.5%).

Poll predictions of an easy win for Dewey made Truman's victory all the more devastating for the polling community. Gallup lost some subscribers. In one study that interviewed 47 editors who had used polls before the election, half said they would no longer do so (Merton and Hatt, 1949).

There were selection biases in the 1948 polls. Interviewers used their own judgment in interviewing quota samples of the public. An academic investigation conducted under the aegis of the Social Science Research Council resulted in a greater emphasis on probability selection of respondents (see Mosteller et al., 1949, and Section 3 later). In addition, the final 1948 published polls were conducted weeks before the election, and as a result, they were not in the field to pick up any shifts in the electorate up until the day people voted. Beginning in 1950, Gallup made its sampling methodology more rigorous, providing interviewers with more strict instructions regarding the selection of respondents.

But polling was now embedded in American politics. As an extension of its campaign activities, the Republican Party provided poll results to the White House after Eisenhower's election. George Gallup and later Louis Harris routinely reported their latest polls to presidents. Harris worked as John F. Kennedy's political consultant and pollster. These president–pollster relationships became regularized with Lyndon Johnson and his pollster, Oliver Quayle; Nixon and his polling operation run by H.R. Haldeman, with pollsters David Derge, Robert Teeter, and members of the Opinion Research Corporation; and Gerald Ford continuing with Teeter (see Eisinger, 2003; Jacobs and Shapiro, 1995, 1995–1996).

Although their wartime survey operations were disbanded, government agencies' interest in survey and Census data continued after the war (Converse, 1987; Sudman and Bradburn, 1987). United States government sponsored political polling was devoted, interestingly and lawfully (as there were prohibitions against polling that could be construed as partisan), only to long-term surveying in foreign countries through the United States Information Agency (USIA; see Crespi, 1999). The tradition of public administration and policy-related survey research continued, including state and local survey research to improve policy formation and implementation, and to facilitate feedback on the performance and effectiveness of government bureaucratic agencies, policies, and programs (e.g., see Desario and Langton, 1984; Kweit and Kweit, 1984, Van Ryzin et al., 2004a, 2004b).

Polling in the United States, especially telephone surveys, expanded substantially in the 1970s with the development of Random Digit Dialing (see Chapter 7) and of computer-assisted telephone interviewing (CATI) allowing polls to be done quickly and relatively cheaply. Household telephone penetration reached 94%, easing interviewing over the phone, which cost less and enabled centralized control over interviewers. Sampling by random digit dialing was also easier and cheaper than the enumeration and sampling procedures used for in-person interviewing. Concerns about interviewer access to respondents in urban areas also motivated the use of phones (see Blankenship, 1977; Nathan, 2001). In other countries, data collection also moved from in-person to telephone when possible.

Journalistic organizations that wanted to poll could centralize data collection with phone banks, occasionally utilizing advertizing or classified ad department offices at night and on weekends. Improved newsroom computer technology, especially the arrival of PCs in the 1980s, meant that polls in response to breaking news could be conducted and reported within a half hour of an event's occurrence. The same expansion occurred for polling by political parties, candidates, and advocacy and interest groups.

But Vietnam and Watergate, by increasing cynicism among journalists and the public, were as important as technology in the development of news surveys during the 1970s. Extending the trend that began in the Progressive era of the 1890s, there was an even further shift away from partisan influence over the press. Journalists again believed it important to bypass party leaders and go directly to the public for its opinions. Media polling made journalists less vulnerable to manipulation by political parties, candidates, organized groups, and others who used poll data for their own purposes (see Gollin, 1980a, 1980b, 1987; Jacobs and Shapiro 1995–1996). The news media could now verify claims by others about the state of or changes in public opinion.

CBS News and NBC News had conducted occasional polls for many years, and had used computers and some statistical modeling in election coverage as early as 1952

(Bohn, 1980), but serious polling units were organized only in time for the 1976 election. CBS News formed a contractual arrangement with *The New York Times* in 1975. NBC News worked on its own in the 1976 election, but joined with the Associated Press from 1977 to 1983, and later with the *Wall Street Journal*. ABC News joined forces with the *Washington Post* in the early 1980s; that partnership continues. CNN first joined with *Time* magazine in 1989, and worked with *USA Today* and the Gallup Organization from 1992 to 2005. Partnerships by different media outlet have occurred in state and local polling as well.

As the ease of fielding surveys increased and costs dwindled, public pollsters increasingly probed the public's vote intentions and preferences on candidates for lower offices. State-level polls routinely asked respondents' views about candidates for statewide office, such as gubernatorial and U.S. Senate candidates. Some survey organizations, such as SurveyUSA, which use automated polling, Interactive Voice Response (IVR, see below), on a large scale, have also fielded congressional-district and even municipal-level surveys that ask respondents about candidates in those races.

The historical literature on polling in other countries is less extensive and cumulative than in the United States (but see Worcester, 1983, 1987, 1991). Polling spread to Mexico and Latin America after World War II through the work of Joe Belden and others. After the fall of Communism, polling in Russia and Eastern Europe came through sociological institutes and partnerships. There are now polling companies in nearly every country, with preelection polls conducted wherever there are elections. This suggests that polling has expanded its reach in tandem with democracy and democratization worldwide. There are also many collaborations, including a European Social Survey, modeled after the NORC General Social Survey, and the broader International Social Survey Programme (ISSP). The Eurobarometer, established in 1973, has been used by the European Commission to measure public opinion in all EU member countries. The Eurobarometer concept has been extended to other continents: the Latinobarometro, established in 1995, the Afrobarometer, established in 1999, and the Asian Barometer Survey (formerly the East Asian Barometer Survey), begun in 2000. The broadest international survey collaboration to study elections is the Comparative Study of Electoral Systems (CSES).

In countries other than the United States, there have been wide variations in direct government involvement in political polling. No systematic comparison across countries has been made to date, although comparative research has slowly begun (e.g., Nacos et al., 2000; Worcester, 1983). In Canada, like the United States, polling for the government originated during World War II. In addition to consulting available Gallup polls, government officials initiated regular opinion polling of their own to study public morale, attitudes toward living costs and rationing, and opinions about Canada's involvement in the war and expectations about the postwar economy (Page, 2006). This began a tradition of government polling and politics in which polls are used for purposes of determining what issues should be high on the agenda and for determining communication and policymaking strategies for promoting policies (see Jacobs and Shapiro, 2000). In the former Soviet Union, the collection of data on media use and other "sociological" topics led to the collection of other opinion data (Mickiewicz, 1981).

As public polling became institutionalized, there was a concurrent increase in professional survey organizations. The American Association for Public Opinion Research (AAPOR) was founded in 1947, followed soon after by the World Association for Public Opinion Research (WAPOR). These organizations provided meeting places where both pollsters and academic researchers could discuss and debate issues in survey research (Sheatsley and Mitofsky, 1992). The National Council on Public Polls (NCPP), an association of public polling organizations with a particular interest in the polls produced for public consumption in the United States, was founded in 1969. The European Survey Research Association was founded in 2005 to provide communication among European survey researchers. There are also organizations that are more concerned about the market research industry, including CASRO (Council of American Survey Research Organizations) and CMOR (Council for Marketing and Opinion Research), now merged with the Marketing Research Association, in the United States, and ESOMAR (originally known as the European Society of Opinion and Marketing Research and now branded as a "world organization").

In addition, archives collect and make available polling results beyond the original data collection and publication. The Roper Center (now at the University of Connecticut) was established in 1947, and pollsters like Roper and Gallup deposited data sets there for academic and public use. Currently, nearly every major public polling organization does the same, including some international organizations. The Inter-University Consortium for Political and Social Research (ICPSR, at the University of Michigan) was established in 1962 and has data sets from academic social science studies, from the United States and elsewhere, as well as public polls. A list of some international archives can be found at: http://www.ropercenter.uconn.edu/data_access/data/data_archives.html.

### *1.2. Polling for private consumption: candidates and campaign strategy*

Public opinion and polling research increasingly became a major part of political campaigns in the United States in the 20th century, with the use of focus groups and other means of message testing, initial benchmark polls to provide basic public opinion information to candidates, periodic trend polls to compare to the benchmarks, and even more frequent – weekly or even daily – "tracking polls" (Eisinger, 2005; Jacobs and Shapiro, 2000; Stonecash, 2003). In this century, political campaigns have attempted to "micro-target" voters and engage in vote-getting activities in increasingly precise ways with the aid of both their own proprietary polling and other publicly available data (Hamburger and Wallsten, 2006; Jacobs and Shapiro, 2005). These polls have been used to find out how specific types of voters might be influenced by emphasizing particular issues or otherwise finding messages that will resonate with them, ultimately affecting their voting decisions.[1]

In the 1992 congressional campaigns, candidates perceived public opinion surveys to be a crucial source of information for gauging public opinion in House campaigns. On a scale that ranged from 1 (not important or not used) to 5 (extremely important), the mean score for the importance of polling was 3.5, second only to personal candidate contact with voters as a source of information (mean score = 4.4). Candidates

[1] The ways in which polls are used for this purpose are sometimes mistaken with a campaign tactic known as "push polls," that under the guise of an opinion survey is an attempt to talk to large numbers of voters (tens of thousands, in contrast to under 1500) in an effort to spread often misleading information designed to directly influence voters (see Asher, 2007).

apparently perceived polling information to be more reliable than information obtained from newspaper, radio or television sources, local party activists, mail from voters, and national party leaders (Herrnson, 2004). Moreover, polling is widespread even in low-salience, local-level elections; 41% of candidates in municipal elections conducted polls (Strachan, 2003).

Presidents' pollsters have been visible figures beginning with Patrick Caddell for President Jimmy Carter, and continuing with Richard Wirthlin for Ronald Reagan, Robert Teeter for George H.W. Bush, and Stanley Greenberg followed by Dick Morris for Bill Clinton (Eisinger, 2003; Heith, 2004; Jacobs and Shapiro, 2000; Murray, 2006). George W. Bush, despite his disparaging claims about political leaders using polls, appointed an academic public opinion expert, Peter Feavre, as a member of his national security team dealing with public support for the Iraq war.

American-style campaign politics, including heavy use of polling, has spread to other countries, and American consultants have routinely worked in both developed and developing democracies. One study found that nearly 60% of U.S.-based political consultants had worked abroad, especially in Latin America, postcommunist countries, and Western Europe (Plasser and Plasser, 2002). For example, Greenberg Quinlan Rosner, whose principal Stan Greenberg polled for Bill Clinton's 1992 presidential victory, worked on the campaigns of Britain's Tony Blair, Germany's Gerhard Schroeder, South Africa's Nelson Mandela, Israel's Ehud Barak, as well as in Bolivia, Honduras, Poland, and Mexico. However, the extent to which American electioneering styles have been adopted abroad has depended on the political, social, cultural, institutional, and regulatory environments in each country.

Polls also have been conducted by interest groups and other organizations. These groups also have polled their own members and supporters, as a way of maintaining contact with them. They have publicized their polling to promote their organizations' goals, just as candidates and parties have promoted their political objectives. Some organized groups have engaged in fundraising or membership solicitation in the context of mail or other political surveys – "soliciting under the guise" of survey research (SUGing) or "fund raising under the guise" (FRUGing) – which survey professional associations such as the American Association for Public Opinion Research (AAPOR) and the National Council of Public Polls (NCPP) have considered unethical uses of surveys (Traugott and Lavrakas, Chapter 3, 2004).

## 2. General methodological issues in public opinion and election polls

Methodological issues in public opinion and election polling can perhaps best be summarized under the general rubrics of *errors in sampling, measurement error, and errors in conceptualizing and specifying* what's being studied (Brady and Orren, 1992; see especially the full guide provided by Weisberg, 2005). Errors that are part of sampling go beyond the estimation of the margin of sampling error and include defining and adequately covering the population that is to be studied. Then comes the problems of nonresponse and the need to weight the data appropriately (see Chapter 8). Measurement errors include the potentially substantial effects of the *mode* of surveying used, question wording, question order or context effect (including the order in which response categories are offered), coding and data processing errors, problems in imputing missing data, and the effects that interviewers can have on responses to questions (see Chapters 9, 10, and 12; Asher, 2007; Schuman and Presser, 1981; Weisberg, 2005). Conceptual and specification problems bear on the ultimate validity of the opinion measures used. In some cases researchers assume respondents have genuine attitudes or that particular issues and attitudes have a high degree of salience, which may not in fact be the case. The researcher must consider the possibility of "nonattitudes" and the transitory nature of survey responses (see Asher, 2007; Bishop, 2004; Converse, 1964), and they should make use of insights to be gained from "cognitive interviewing" methods and the results of survey pretests (see Beatty and Willis, 2007; Sudman et al., 1996; Tourangeau et al., 2000). Even in cases where researchers are confident that survey responses reveal real attitudes, they may have difficulty.[2]

The different sources of survey error can overlap or interact with each other, and they can vary by the mode in which the survey is administered. Modes (or ways) of conducting surveys include in-person/face-to-face interviewing, mail and other paper surveys, telephone surveys, and most recently, on-line polling. Each survey mode has advantages and disadvantages in the way interviewers or data collectors interact with respondents (Asher, 2007; Dillman, 1978, 2007; Fricker et al., 2005; Weisberg, 2005). Nonresponse is a persistent and growing problem across modes, but it is more prevalent in those selected for lower cost (nonresponse error includes noncontacts and refusals, as well as problems of noncoverage of the desired population because of a limited sampling frame). If the attitudes and opinions of nonrespondents differ significantly from those of respondents in ways not easily corrected through demographic weights, nonresponse bias is introduced. In-person surveys normally have higher response rates. Their larger costs insure the close contact with respondents that typically lead to higher response rates.

But even when different surveys use the same mode there can be survey "house effects," the effects of the particular procedures, instructions to interviewers, and other rules for interviewing that are specific to individual survey organizations (Smith, 1978).

Survey respondents often appear to conform to "social desirability" pressures: self-reported rates of voting in elections are substantially (and inaccurately) higher than actual turnout rates (cf. Belli et al., 2001; McDonald, 2003a). Complications and bias may also arise from respondents' reactions to interviewer characteristics, including race, gender, age, and ethnicity. Evidence of interviewer effects dates back to the 1940s (Katz, 1942). Most notably, there have been significant race-of-interviewer effects (cf. Finkel et al., 1999). Black respondents may report more favorable attitudes about white candidates to white interviewers than they do to black interviewers (Anderson et al., 1988; cf. Finkel et al., 1999). Black respondents also demonstrate higher levels of political knowledge in telephone surveys when interviewed by black interviewers than when they

[2] One illustrative case arose in the 2004 U.S. presidential election in which responses to the exit polls indicated that more than 20% of voters chose "moral values" from a list of issues as the most important influence on how they voted. This led to a lively debate about how the exit poll respondents interpreted the meaning of "moral values" – whether this phrase referred to issues like abortion or the rights and behavior of homosexuals, whether it was a statement about the candidates' morality, or whether it was a fall-back answer if respondents did not think the other issues on the list were sufficiently salient to their vote choice (see Langer and Cohen, 2005).

are interviewed by whites (Davis and Silver, 2003). The race-of-interviewer effect on expressed voter preferences in preelection polls has been especially acute in elections with a white candidate running against a black candidate. In one study, white respondents were 8 to 11 points more likely to express support for the black candidate to black interviewers than to white interviewers (Finkel et al., 1999). In some past elections, white respondents have been far more likely to report an intention to vote for the black candidate compared to their behavior on Election Day, although there has been less evidence of this in more recent elections and no evidence that it happened in the 2008 election of Barack Obama (Hopkins, 2008). Gender and ethnicity-based interviewer effects have also been reported (Hurtado, 1994; Kane and Maccaulay, 1993). Interviewer effects have also been raised as a significant problem in conducting exit polls (discussed below; cf. Bischoping and Schuman, 1992).

Different modes of interviewing can mitigate or exacerbate survey problems.

### *2.1. Mail surveys*

Mail surveys have been an appealing mode of data collection, partly due to lower cost. Because mail surveys are self-administered, interviewers are not necessary, thus eliminating them as a source of bias. The assurance of anonymity and confidentiality may also encourage respondents to be more forthright, especially when being probed on sensitive or controversial issues (Asher, 2007; Dillman, 1978, 2007). Mail surveys can be an effective way to gather information from smaller, specialized population groups. The main drawback of mail surveys has been their high rate of nonresponse, which has tended to exceed the nonresponse rates for telephone and personal interviews. General population mail surveys have frequently achieved response rates under 5% (Dillman, 1978, 2007; Glynn et al., 2004). Monetary incentives, personalized notification letters that arrive before the survey, multiple correspondences with nonresponders, and prepaid return postage can boost response rates for mail surveys (Asher, 2007). Other problems have included having little assurance about who actually completed the survey, no opportunity for question clarification, and some uncertainty as to what order questions were answered in (pertaining to possible context effects). Mail surveys also tend to require long field periods for follow-up mailings.

### *2.2. Telephone surveys*

Most surveys in the United States since the 1970s have been conducted by telephone. Random Digit Dialing (RDD) techniques allowed researchers to select a random sample of households, helping to resolve the sampling challenges often inherent with other survey modes. Refinements to RDD made the process even more efficient.

One early innovation was the Mitofsky–Waksberg procedure developed by Warren Mitofsky (Mitofsky, 1970) and later refined by Joseph Waksberg (Waksberg, 1978). Residential phone numbers represented only about 20% of phone exchanges nationwide but tended to be highly clustered within working blocks of 100 consecutive numbers. Mitofsky–Waksberg, a form of area probability sampling, took advantage of this fact to exclude attempts to banks of phone numbers that were unlikely to be residences. The procedure randomly generates 10-digit phone numbers that can be called to identify working blocks. Random samples can be generated from those blocks identified as working. This procedure increased considerably the efficiency of sampling for telephone surveys while preserving the scientific integrity of the sample.

Telephone sampling was further refined to take advantage of list-assisted sampling techniques, using telephone book listings as a seed to generate random household numbers (Brick and Waksberg, 1991). Important refinements were made, involving the statistical theory underlying list-assisted methods and the effects of telephone system changes, including the lower proportion of telephone numbers assigned to residential units (see Tucker et al., 2002, 2007).

Registration-based sampling (RBS) allows survey researchers to draw random samples of registered voters from publicly available voter files rather than relying on the respondent's self-reported registration status. Pollsters for political campaigns have used RBS sampling extensively, despite the fact that registration-based samples have frequently suffered from incomplete coverage (due to unavailable telephone numbers) of the population of registered voters and inaccuracies in the voter file.

An advantage of sampling from registration lists has been that many lists contain additional information from public records that can be used to forecast voter turnout. Although the quality of information available has varied across jurisdictions, typical registration files have contained dates of birth, dates of registration, and (where relevant) party registration. Data on past voter turnout has been furnished by registrars or acquired by private vendors, and researchers can use this information to assign voting propensity scores to individuals on the list and thus draw samples of voters most likely to participate in the election. Some have found that registration-based sampling can improve the accuracy of election forecasts over RDD (Green and Gerber, 2006).

Computer-assisted telephone interviewing (CATI) has allowed polling to be faster, more efficient, and more accurate (Asher, 2007). CATI has also permitted randomizing the order of questions and even response categories for each question, diminishing the possibility of order effects. CATI has helped facilitate and expand experiments in public opinion research, through the National Science Foundation-funded Time-sharing Experiments for the Social Sciences Program, using both telephone and internet surveys (see TESS).

Telephone surveys have had several disadvantages, however, including growing rates of nonresponse, partly due to innovations like caller ID, answering machines, privacy managers, and increasing cell phone coverage (see below). Interviews have tended to last no more than 15 or 20 minutes, as respondents become fatigued faster than in other modes. Telephone surveys have also been vulnerable to interviewer effects, as the characteristics of the interviewer and their levels of training and motivation can influence the quality of the data collected.

### *2.3. In-person/face-to-face surveys*

Although in-person surveys became less common in public opinion and election polls well before the end of the last century, primarily due to cost and time considerations, they have remained an effective way to collect rich and complete public opinion information. The National Opinion Center's General Social Survey (GSS), for example, continued to be administered in-person. Any type of simple, stratified or systematic sampling of respondents dispersed over a vast geographic area would be impractical, so in-person surveys have typically used area probability samples that are essentially stratified multistage cluster samples (MCS).

MCS requires the country to be divided into geographic regions (four in the United States). Within each, a set of counties and standard metropolitan statistical areas have been randomly selected. Within these, primary sampling units (PSUs) of four or five city blocks are randomly selected and then four or five households randomly selected from each block. Random selection has sometimes been abandoned at the household level because complete lists of all household members may not be available, but interviewers have generally relied on some systematic method to select respondents within the household (Erikson and Tedin, 2007). The General Social Survey and the National Election Study continued to select respondents randomly at the household level. The response rate for the 2006 GSS was 71%; the preelection response rate for the 2004 NES was 66%.

One of the main disadvantages of in-person surveys is their higher cost, but these costs are necessary because these surveys are designed to achieve (and they are usually successful in achieving) higher response rates than other modes of data collection. On the plus side, nonresponse rates tend to be lower, and respondents are usually more engaged and forthcoming in the interview. Interviewers are also able to be more personal and interactive, use visual aids and clarify questions, and monitor nonverbal behavior. However, the potential for interviewer effects described earlier is maximized with surveys administered in-person.

### *2.4. Online surveys*

Online surveys offer many advantages, including cost efficiency and speed. Researchers can also develop and administer complex questionnaires and they inform respondents how far they have proceeded through the survey (something they can directly see in mail surveys). More important, researchers can include visual aids and they can conduct experiments with random assignment of "treatments," as well as provide graphic or multimedia enhancements. Item nonresponse can be reduced because respondents can be reminded and motivated to revisit incomplete items. Web-based surveys are self-administered, so interviewer effects can be avoided.

On the other hand, web-based surveys present serious methodological problems that may limit researchers' ability to make valid statistical inferences from them, including challenges in assembling sampling frames for probability sampling, coverage issues, and selection bias. Despite rapid penetration of the Internet into households, coverage issues remain especially acute in the United States, because access is restricted to about 60% of the population. Web-based surveys are also not immune to nonresponse. Other disadvantages include questionable security and authentication procedures and differences in format and presentation across computing systems and browsers (Dillman, 2007).

Since as yet there is no available listing of electronic mail (e-mail addresses) and not everyone has had online computer access, one approach in the United States has been to use RDD or in-person survey probability sampling methods to draw random samples of adults, giving respondents without internet access such access to form "panels" who participate in multiple surveys, thereby reducing the costs of drawing repeated samples. A more controversial variation of this method has recruited thousands of volunteers to join panels through invitations on internet sites; this approach has assumed that any selection bias can be corrected through sophisticated methods of "propensity score" weighting (Rosenbaum and Rubin, 1983; Taylor, 2000; Taylor and Terhanian, 1999). Online panels have been used in multiple countries in Europe and Asia, even some with relatively low internet penetration. Yougov was founded in 2000 in the United Kingdom, and has maintained an internet panel for research. It was successful in predicting the outcome of the 2001 Parliamentary election, performing better than conventional polls. It was not quite as good in the 2005 Parliamentary election, and it fared poorly in its efforts in the U.S. 2004 presidential election. To date, preelection polls from self-selected internet panels in the United States have done no better than telephone surveys, and often less well. But, given the problems facing telephone surveys, they have received a great deal of interest. There has been a need for more transparency and openness so that these surveys (like IVR, see below) can be evaluated fully (Blumenthal, 2005).

Probability sampling for Internet surveys originated with Willem Saris who developed this method in the Netherlands prior to the development of the Internet (Saris, 1998). It utilized "computerized self-administered questionnaires" (Couper and Nichols, 1998, p. 13) and was implemented by the Telepanel of the Netherlands Institute for Public Opinion (NIPO or Dutch Gallup). Respondents were provided with computers and modems, and were trained to download and fill out the questionnaires by computer. Upon completing the questionnaire, each respondent uploaded it for data collection and processing. The Telepanel idea was adopted in other countries. Most recently, in 2006, the US National Science Foundation funded a trial run of this method using in-person full probability sampling recruitment of respondents. The Dutch government also provided a major grant to support a large-scale Telepanel.

### 2.5. *Mixed or multiple mode surveys*

To date, mixed mode methods have not been used extensively in public opinion and election surveys. Mixed mode surveys offer respondents a choice of response modes. Multiple mode surveys have collected information from respondents using several survey modes. Some respondents may participate by mail, for example, while others in the same survey may be interviewed on the telephone. The methodological considerations described earlier for each survey mode continue to apply to mixed mode surveys respectively, and others may arise when combining samples of respondents interviewed using different modes, such as differences in responses depending on the method of data collection. Offering respondents a choice of response modes may increase rates of participation.

## 3. Preelection polling: methods, impact, and current issues

The earliest scientific preelection polls, in the 1930s, relied on in-person interviewing. Interviewers were sent to selected locations and instructed to interview a specific number of men and women, young and older people, higher and lower-status voters. Completed questionnaires were then returned for tabulation and reporting. The Gallup Organization developed a procedure to speed up the polling as Election Day drew near, in which interviewers telegraphed back the responses to specific questions.

As already noted, the early Gallup and other polls had significant successes in predicting presidential victors, at least until the 1948 election. Their methodology, particularly the reliance on the quota selection process administered by the interviewers themselves, had been questioned by government statisticians like Louis Bean, of the Department of Agriculture, Philip M. Hauser and Morris Hansen, of the Bureau of the Census, and Rensis Likert, from the Bureau of Agricultural Economics, and some academics. The post-1948 Social Science Research Council Report (Mosteller et al., 1949) also questioned the reliance on quota samples, and it observed that the pollsters had overestimated the capabilities of the public opinion poll.

There was another source of concern in making prediction from preelection polls. What the pollsters (and even some academics, like Paul Lazarsfeld) had discovered from the 1936, 1940, and 1944 presidential elections was that few if any changes could be attributed to the campaign (Lazarsfeld et al., 1944; on the history of "minimal" campaign and especially "media effects," see Klapper, 1960). In 1948, they believed that the lead held by the Republican candidate in the fall could not be affected by anything either candidate could do. Polling stopped several weeks before the election. After 1948, pollsters would poll much closer to Election Day.

Methodological changes tend to follow problems in election prediction internationally as well. For example, in Britain, preelection pollsters wrongly predicted a Labor victory in the 1992 parliamentary election. The Conservative Party won by eight percentage points. The British Market Research Association conducted an investigation, and found similar problems to those in the United States in 1948 (Jowell et al., 1993). Voters changed their minds at the last minute, and there were also problems with the sampling methods. Most British pollsters opted to continue in-person quota sampling for the next national election, although they did change their quotas. But others moved to implement greater changes, including the adoption of telephone polls. Similar issues were reported in 2002 preelection poll errors in France (Durand et al., 2004).

When preelection polls underestimated Ronald Reagan's victory margin in the 1980 election, pollsters decided that in future years they needed to continue interviewing through the night before the election (Hansen, 1981; Kohut, 1981; Mitofsky, 1981).

Differences in methods may apply to the designation of likely voters or allocation of undecided respondents (discussed below), but may also mean framing preference questions differently. Martin et al. (2005) found evidence of these differences in the preelection polls conducted in 2000.

### *3.1. The allocation of undecided voters*

Preelection polls are routinely adjusted from pure probability samples of the adult population. Some of those adjustments include the management of voters who refuse to give a preference when asked. The range of the number of undecided voters in preelection surveys can vary greatly across surveys and across stages of the campaign. Between 1988 and 1996, the proportion of undecided voters reported in polls ranged from 3 to 73% of the sample (Visser et al., 2000). Typically, 15% or more of the electorate may be undecided during a presidential campaign. This figure tends to be even higher in lower-level, less-salient races (Erikson and Tedin, 2007). The proportion of "undecideds" drops substantially in the final days of the election, although 5% of voters in the 2004 exit poll claimed they made up their minds on Election Day.

Some polling organizations treat undecided voters as a separate category and include that percentage in the final preelection estimate; others remove them and recalculate the percentage for each candidate or party, and others have attempted to allocate them. Allocation schemes have varied considerably. One approach has assumed that undecided voters who end up voting do so randomly, so truly undecided respondents should be allocated equally to the main candidates. This procedure can yield more accurate forecasts than eliminating undecided respondents altogether (Erikson and Sigelman, 1995; Visser et al., 2000). Another approach allocates undecided voters disproportionately to the challenger, as preelection surveys may systematically underestimate support for the challenger because of the "spiral of silence" (Noelle-Neumann, 1993 [1984]). In one study of a wide range of statewide, congressional and municipal primary and general election races, undecided respondents disproportionately voted for challengers in 82% of elections (Panagakis, 1989).

Changes in procedures also make comparisons difficult. In 1992, for example, the Gallup Organization changed its allocation method and, as a result, severely overestimated support for Bill Clinton and underestimated support for George W. Bush and independent challenger Ross Perot (Traugott and Lavrakas, 2004). In the weekend before the election, Gallup decided to assign all undecideds in its tracking poll to Clinton, citing the tendency of undecided voters to ultimately choose challengers. Clinton's lead grew from two points on Friday to 12 on Monday. But rather than moving to Clinton, many undecided voted for independent candidate Ross Perot (Erikson and Tedin, 2005).

### *3.2. Weighting and determining likely voters*

The British difficulties in 1992 and the reaction to them underscored a major methodological issue for preelection polls. Should one weight or adjust the results? The adjustments can range from insuring that the original sample reflects the appropriate population parameters and the probabilities of selection to weighting on past voting behaviors. Adjustments may also be required to ensure that the final published results reflect the opinions of actual voters, not all adults.

In countries like the United States, where voter registration is not automatic, a large portion of respondents may not vote. Candidate preferences of nonvoters may be different from those of actual voters.

In the United States, pollsters have almost always first asked respondents whether or not they were registered and then asked those registered a series of questions designed to separate voters from nonvoters, including whether the respondent had voted in the past and would vote in the current election, as well as a measure of political interest. Some screens have included whether respondents knew the location of their polling places. Based on their answers, registered respondents can be assigned a probability of voting which is then used as a weight when tallying the projected vote. A more common solution (used by Gallup and many other pollsters) has been to divide registered respondents into two groups. Respondents who scored beyond a specified cutoff have been designated as "likely" voters, whose choices are then counted in the tally. The choices of those scoring below the cutoff are excluded in the estimation (Asher, 2007; Crespi, 1988; Daves, 2000; Erikson et al., 2004).

Estimates of likely voters in the weeks and months prior to Election Day may reflect transient political interest on the day of the poll, and might have little relationship to behavior on the day of the election. Even though such likely voter samples might well represent the pool of potential voters sufficiently excited to vote if a snap election were to be called on the day of the poll, they may not be the same people voting on Election Day. An analysis of Gallup polls in the 2000 presidential election indicated that the sorting of likely and unlikely voters is volatile and that much of the change (although certainly not all) is an artifact of classification. Pollsters can mistake shifts in the excitement level of the two candidates' core supporters for real, lasting changes in preferences (Erikson et al., 2004).

### *3.3. Tracking polls*

Tracking polls, which monitor campaign dynamics on a daily basis, originated with political campaigns, which used them to evaluate campaign events and the impact of political advertizing. Journalists adopted them in the 1980s. Typically, tracking polls contact small samples of respondents (100–350) each day. To update results, a new day's sample of respondents is added to the total sample and the oldest day's sample of respondents is dropped. On their own, these samples are too small to provide precise estimates of preferences, but pollsters have used rolling averages of two or three consecutive days' worth of interviewing. Thus, estimates can be based on 500–600 interviews aggregated across all days (Traugott and Lavrakas, 2004).

Tracking polls may be useful to assess campaign dynamics, but there are shortcuts used in these surveys. Tracking polls have typically been one-night surveys that have not always employed the rigorous sampling and respondent selection procedures that many other polls do (Traugott and Lavrakas, 2004). Respondent call-back appointments have rarely been made and interviewers have not always selected respondents randomly within households. There have been disagreements over whether samples should be weighted each day or over several days. The *Washington Post* tracking poll in the 2004 presidential election, for example, adjusted each day's randomly selected samples of adults to match the voting-age population percentages by age, sex, race, and education, as reported by the Census Bureau's Current Population Survey. The *Post* also adjusted the percentages of self-identified Democrats and Republicans by partially weighting to bring the percentages of those groups to within three percentage points of their proportion of the electorate, as measured by national exit polls of voters in the last three presidential elections (*Washington Post*, 2004). Despite these challenges, the accuracy of tracking polls has been shown to be superior to other polls in some studies (Lau, 1994).

Several companies, in both the United States and Great Britain, have conducted election polls among a sample of individuals recruited to be part of web panels. The resulting interviews conducted online have been adjusted by demographics and politics to reflect a predetermined estimate of the electorate (Taylor, 2000).

For the most part, pollsters have been reluctant to weight by party identification. The main hesitation has been that party identification is not a fixed characteristic of the electorate in the United States, and there has been evidence of significant short-term fluctuation in party ID. But overall partisanship exhibits considerable stability over time. Pollsters can estimate the underlying proportions of Democrats and Republicans in the electorate based on moving averages of results from surveys conducted over several weeks, and these estimates might be used, with caution, to weight samples (see Abramowitz, 2006).

### *3.4. Preelection poll accuracy*

Although lopsided attention has been devoted to notable failures to predict election outcomes, the results of preelection polls conducted at the end of an election cycle have overall tended to come within a few percentage points of the actual outcome (Crespi, 1988). Analyzing presidential races between 1984 and 1992, one study reported an average error of 4.5 percentage points (Gelman and King, 1993). The National Council on Public Polls (NCPP) calculated that the average error of national polls in 2004 and in 2008 was just 0.9 percentage points (see Martin et al., 2005; Traugott, 2005; NCPP, 2004, 2008). In fact, the "trauma" that has often followed inaccurate poll-based predictions of election results has been a testament to the general reliability of polls (Mitofsky, 1998).

## 4. Exit polling

### *4.1. Uses*

Exit polls are polls of *voters*, interviewed after they have left their places of voting and no later than Election Day. They may include the interviewing before Election Day of postal, absentee and other early voters. Exit poll functions are not mutually exclusive: they can predict election results, describe the patterns of voter support for parties, candidates, and issues; and support extensive academic research efforts with which the results are formulated and disseminated.

Election projections can be made in ways other than by interviewing voters as they exit the polling place. Though most projections are based on exit polls, interviewing voters after having voted at a polling place, other forecasting models may include: CAPI, CATI or other interviews on Election Day with voters after *or before* having cast their votes and counts of official votes in a sample of precincts, often known as quick counts.

A standard use for exit polls in new democracies has been as a check on voting itself. In recent years, exit poll results in Venezuela, the Ukraine, Georgia, Peru, and Serbia have been hailed by some as better indicators of election outcomes than the vote count. Although a well-conducted exit poll can sometimes be a check on fraud, sampling error limits any poll's precision, and operational difficulties, including restrictions on carrying out exit polls, and possible bias due to interviewer–respondent interactions can call into question the accuracy of those, and other, exit poll results.

The first exit poll was perhaps conducted inadvertently in the United States by Ruth Clark in 1964. Clark, a well-known newspaper researcher, began her research career as an interviewer. In 1964 she worked for Louis Harris and was sent to conduct interviews in Maryland on its primary election day. Tired of door-to-door interviewing to look for voters, she decided to talk with them as they left the polling place (Rosenthal, 1998, p. 41).

The exit poll did not become a staple of news election coverage until the 1970s and 1980s. CBS News, under the leadership of Warren Mitofsky, began exit polling in a

1967 gubernatorial race in Kentucky to collect voting data in precincts that did not make their vote available at poll closing. It later expanded its questionnaire to include questions about voter demographics and issue positions. The process was adopted by other American news organizations and then quickly spread to other countries (Frankovic, 2007; Mitofsky, 1999).

The first exit poll in Great Britain was conducted in 1974 for ITN by Humphrey Taylor's United Kingdom company, part of Louis Harris and Associates, followed soon after by exit polling in other Western European countries. Other non-European democracies adopted exit polling soon after – for example, Social Weather Stations in the Philippines conducted its first Election Day poll in 1992, and Mexican researchers did so in 1994. Mitofsky himself did exit polling in the Russian elections starting in 1993, working with the Russian firm CESSI.

To provide a random sample of voters, the exit poll locations (precincts) must be selected using probability sampling, proportionate to precinct size, with some stratification by geographic location and past vote. Interviewers have to be hired and trained, and stationed at the selected poll locations. Voters at the polling locations have to be sampled, either by interviewing every voter or a probability sample of them (every *n*th, with *n* determined ahead of time depending on the expected size of the precinct). Records of nonresponse normally should be kept – indicating its size and composition. Results then must be transmitted to a central location for processing, either physically, by telephone, or electronically.

In the United States, estimates of election outcome are made on a state-by-state basis, because of the allocation of electoral votes by states to the presidential candidates. The precinct tallies are weighted by size and their probabilities of selection, a nonresponse adjustment is made following a quality control check, and the results are entered into several estimation models – stratified by geography or past vote, including simple estimates and ratio estimates using the past vote. The models include tests for significance.

There have been different types of exit poll questionnaires. Some, as in Britain, have simply asked which candidate the respondent voted for. In contrast, a typical United States exit poll may contains 25 questions on both sides of a single sheet of paper including the importance of issues and demographic characteristics.

### *4.2. Problems for exit polls*

The most serious methodological issue for exit polls has the level and distribution of nonresponse, as this may result in bias due to differences between those voters willing and those unwilling to respond. In addition, interviewer effects can be great because exit polls are conducted in person, although paper and pencil questionnaires preserve confidentiality and can reduce the impact of this concern on respondents (Bishop and Fisher, 1995). Examples of differential nonresponse have been documented in response rates of voters to interviewers of different races in elections with a racial component (Traugott and Price, 1992), and in other highly intense elections where interviewers may be perceived (correctly or incorrectly) as favoring one or another candidate or party. In the 2004 U.S. presidential election, exit poll overestimates of the vote for Democrat John Kerry were frequently cited as evidence of fraud by some activists; but all analysis indicated the difference was more likely caused by a differential response rate due to the interviewer–respondent interaction (Edison Media Research and Mitofsky International, 2005; Traugott et al., 2005). Younger interviewers achieved lower response rates than older interviewers. Younger voters in general were more likely than average to be Kerry voters, and that perception (frequently reported before the election from preelection polls) may have influenced potential respondents.

In the United States, some states have passed legislation requiring exit poll interviewers to stand as far as 300 feet (nearly 100 m) from the polling location, effectively making a good sampling of voters impossible. In some countries, the difficulties of interviewing at the polling place (either through legal restrictions or fear of violence) have forced researchers to use different methodologies, such as in-person interviewing at home after people have voted. In the Philippines, day of election surveys at voters' homes are substituted for an exit poll at the polling place (voters there can be identified by an indelible mark on their hands).

In the 2000 United States election, mistaken projections in the state of Florida were attributed to exit polls, although when news organizations first projected that Al Gore would win Florida's electoral votes (7:50 p.m. ET), more than just exit poll results had been received. Twelve of the 120 sample precincts had reported *actual* tabulated results, and six of those precincts were part of the exit poll sample. Four percent of *all* precincts statewide had reported their votes. At 7:50 p.m., all of the estimation models indicated a Gore victory, and the estimates met the tests of significance.

There were several data problems. Precincts selected for the exit poll were not a true reflection of the state results. The difference between the actual precinct vote and the state totals was at the outer edge of sampling error. The ratio estimation model used only one past race for comparison, and that was the 1998 gubernatorial election, which had a 0.91 correlation with the vote for the 2000 Republican candidate, Bush. But using this race, the size of the absentee vote was underestimated – at only 8% of the total. As it turned out, the correlation of the 2000 Bush vote and the 1996 vote for Bob Dole was nearly as high (0.88). In addition, the correlation of the Democratic vote in those races was significantly higher than for the 1998–2000 comparison (0.81 vs. 0.71). Had that race been chosen for use by the ratio estimate, the absentee vote would have not been so grossly underestimated. Accurately estimating the size of the absentee vote is extremely important in states like Florida, where absentee votes historically were more than 20 points more Republican than the in-polling place day of election votes.

There was also differential nonresponse. In comparing exit poll results by precinct with the actual vote in that precinct, one can compute the average *Within Precinct Error*. This differential nonresponse has been attributed to many things, including variations in levels of enthusiasm for each candidate.[3] Early on election night 2000, it appeared the exit poll was understating the vote for Gore, and overstating the vote for Bush. That had been the pattern in Kentucky, the only other state where a *WPE* calculation could be made at the time. However, though the overestimate of the Bush vote in the exit poll remained true for Kentucky at the end of the night, it did not remain true in Florida. (The later projection of Bush as the victor in Florida, which was also withdrawn, was made without any use of exit poll results, only tabulated vote counts).

[3] The 1992 Republican Presidential primary in New Hampshire provided an instructive example of this. The exit poll indicated that Pat Buchanan might receive as much as 40% of the total vote against then President George H.W. Bush. He did not. According to the exit poll, Buchanan voters were more enthusiastic about their candidate than Bush voters were.

WPE has been rarely calculated elsewhere in the world, because the results in each precinct are not available in many places. In the United Kingdom, votes are aggregated and released publicly only at the constituency, not the polling place level. And in countries where election day polls have not been conducted at the polling place, comparisons can only be made to larger geographic units.

After the 2000 election, the U.S. exit poll operation was reviewed by RTI-Research Triangle Institute, which suggested a number of improvements that could be made to the methods of the Voter News Service (VNS), the organization that conducted the exit polls in that year. The main suggestions were as follows: improving the methodology for estimating the impact of absentee voters, improving the methodology for estimating outstanding votes in close races, improving the measures of uncertainty for election estimates, improving quality control, developing better decision rules, and exploring new approaches (RTI, 2001).

In the United States, the average within precinct error has been consistently in favor of the Democratic candidate. In three recent elections, 1988, 1996, and 2000, the average error on the difference between the candidates has been about 2 points (2.2 in 1988 and 1996, 1.8 in 2000). But in 1992 and 2000 the errors were 5.0 and 6.6 points, respectively. (This WPE calculation does not include polling places where there are many different precincts voting.) Turnout was higher in 1992 and 2004 than in the other elections, and the level of interest in the campaign was also high. In 1992 and 2004, two-thirds of voters reported paying a lot of attention to the campaign; fewer than half did in 1988, 1996, and 2000.

WPE was higher in larger precincts, in urban precincts, and in precincts where the respondent selection rate was high. It was higher in more competitive states. It was also greater in precincts with more Bush voters. WPE was correlated with interviewer reports of legal or other difficulties with election officials, with the distance an interviewer was forced to stand from the polling place, and with bad weather. But it was also correlated with interviewer characteristics: younger interviewers had higher WPE than older interviewers. The adjustments made in 2006 were the recruitment of a greater number of older interviewers and active attempts to encourage good relations with polling place officials. There was still some evidence for similar problems in the 2006 midterm election exit poll; the early afternoon tabulations compared to official vote returns showed that the Democratic candidates had a margin of vote in the exit poll that was about 4 percentage points too large (Lindeman, 2007).

Exit poll accuracy has also come under scrutiny in other countries. Investigation by a blue ribbon panel of an over-report of the vote in the exit poll for Gloria Macapagal–Arroyo was traced to exceptionally high nonresponse in metropolitan Manila, where many respondents were simply not available during the interview period. The Philippine exit poll was conducted away from the polling place, at respondents' residences.

Changes in the ways elections are conducted will affect exit polls. Absentee voting, vote by mail, and other forms of early voting have been increasingly permitted in the United States, so interviews conducted only at polling places will not include many voters. In two U.S. states (Oregon and Washington), nearly all votes are, at this writing, cast by mail; in more than half, the absentee/early vote has become a quarter or more of the total. Consequently, in the U.S. exit polls must be combined with telephone surveys conducted in the days before the election to see a full portrait of the electorate.

## 5. Postelection and between-election polls

Postelection surveys are often conducted by academic organizations for scholarly purposes and facilitate the analysis of public attitudes and behavior during election cycles (see discussion of NES above).[4] Media organizations conduct extensive polling to track and monitor public preferences in the postelection period as well as reactions to election outcomes. Perhaps, the most extensive use of postelection polling is by elected officials and party organizations who find it useful to continuously monitor public opinion on issues of public policy. At the presidential level, the polling apparatus has essentially become institutionalized to provide private data about the state of public opinion to the chief executive and his key advisers (Jacobs and Shapiro, 1995; Murray, 2006). There is considerable debate about how this opinion data is used by politicians (cf. Eisinger, 2003; Heith, 2004; Jacobs and Shapiro, 2000), but a general consensus about politicians' growing reliance on postelection private polls in the era of what Sidney Blumenthal has named "the permanent campaign" (Blumenthal, 1982).

## 6. Other opinion measurements: focus groups, deliberative polls, and the effect of political events

### 6.1. *Focus groups*

Most polls through the early 1970s relied on face-to-face interviewing. Ancillary research may have included longer intensive interviews and the use of carefully planned and moderated small group discussions to learn about perceptions and attitudes. These were originally called "focused interviews," but are now widely known as the "focus group" (e.g., Delli Carpini and Williams, 1994; Morgan, 1996). Technically, focus groups are not polls but in-depth interviews with a small number of people (6–12) often selected to represent broad demographic groups (Asher, 2007). Focus groups became widely used in market research and later in political campaigns as ways to learn people's opinions about products and candidates and their perceived sources for these opinions. Focus group participants do not constitute random or purportedly representative samples, because they are rarely, or ever, selected through random sampling, but they can provide useful information and pretest questions being developed for a larger scale survey. Effective ways of framing issues and messages can be explored in focus groups, and conversations among their dozen or fewer participants can provide insights into how individuals' opinions are shaped and change in response to new information, such as candidates' statements, news media reports, and advertizements. Focus group

[4] One methodological consideration relevant to postelection studies is inaccurate respondent recall of past behavior. Studies reveal that nontrivial numbers of respondents misreport vote choice. Moreover, there is evidence that retrospective reports of vote choice systematically magnify the support actually received by winning candidates, Using NES data, Wright (1993) reported that the prowinner bias tends to be relatively modest for presidential contests (about 1.5 percentage points) but over-reported for winners of congressional or gubernatorial races average between 4 and 7 percentage points, differences that far exceed amounts we could expect from sampling error. Similar evidence of such "bandwagon effects" has been detected using other data sources. Lindeman (2006) also shows that "false recall" favoring winners in presidential elections often grows over time.

leaders are able to probe participants to react to various stimuli while simultaneously observing other participants' reactions and redirecting the discussion to keep it relevant as necessary.

Generalizations from focus group participants do not apply to broader populations. Still, it is an appropriate methodology to illuminate the process and complexities of preference formation and attitude change. Mainly, focus groups highlight that people regularly and continuously construct views on complex issues through cognitive processes rather than retrieve those views (Glynn et al., 2004).

### *6.2. Deliberative polling*

Deliberative polls have attempted to combine the virtues of focus group studies with those of standard public opinion surveys and to study and improve the quality of survey data by affecting the dynamics and quality of public opinion itself. Their conductors have first drawn a random sample of the public through probability sampling, interviewed respondents, and then brought the sample together to meet and learn about issues and problems through briefing materials, meetings with experts and political leaders, and small-group discussions. This survey method has provided a way of observing how public opinion is transformed, by the time of a later follow-up survey, through information and debate (Fishkin, 1997). The first national deliberative poll "sample" was convened in the United States in Austin, Texas, in 1996, in the context of the upcoming presidential election. Several other deliberative polls have been conducted to date, including British deliberative polls on "Europe 1995" and the "Monarchy;" an "Electric Utility" deliberative poll in Texas; "Australia Deliberates" in 1999 and one on "Aboriginal Reconciliation" there in 2001; one in Denmark on the Euro in 2000; and even a first-of-its-kind deliberative poll sponsored by the local government in Zeguo, China, on local infrastructure projects (Fishkin et al., 2006; see also Luskin et al., 2002). Other survey researchers have attempted to study deliberation through the context of a single survey itself as respondents are asked to react to new information provided in batteries of survey questions (see Kay, 1998). Another variant of the deliberative poll samples participants using a probability sampling method and has them interact in online groups to see whether and how opinions change when participants are interviewed again (see Lindeman, 2002; Price, 2006; Price and Neijens, 1998; on deliberation more generally, see Mendelberg, 2002).

Critics contend that the conclusions of deliberative polls cannot be generalized to the population at large, despite their randomly selected samples, because the public is unlikely to be exposed to information or experiences in the way participants in deliberative polls have been. Moreover, participants' attitudes may be influenced by the heightened sensitivity associated with participation.

### *6.3. The effect of events, political debates, and changing conditions*

One important academic and journalistic use of polling is estimating the effect of events, including crises, political debates, election campaigns, and other changing circumstances and conditions, on short-term changes in public opinion. This has been attempted when surveys have been conducted frequently over short periods of time, or by tracking polls. Impact can also be measured by *panel surveys*, where the same respondents are interviewed again, and individual change (and the reasons for it) can be specified, although the "panel" may be subject to attrition (see Chapters 5, 33, and 34).

The effects of presidential debates in the United States have been studied extensively through experiments or small group studies, as well as larger scale surveys conducted before and after those debates. Although there is not a consensus among researchers, there is evidence that debates (and political campaigns generally) increase public knowledge about the candidates and salience of the issues raised. Some studies have reported evidence of effects on candidate support and hence election outcomes (cf. Benoit et al., 2003; Geer, 1988). One found that presidential candidates perceived as victorious in debates against their opponents typically experience a surge in support following the debate: in 1984, Mondale was perceived as the winner of the first debate against Reagan, and experienced a bump of 3 to 4 percentage points (Holbrook, 1996).

U.S. political conventions have offered parties an opportunity to present their candidates and image to voters in a positive and relatively uncontested format. The resulting spike in support for the party's nominee can be substantial and have lasting implication that can carry through to Election Day. On average, presidential contenders between 1964 and 2004 received a 12-point boost in two-party support following their convention (Panagopoulos, 2007).

The impact of events on political attitudes and preferences can also be detected in performance evaluations of incumbent presidents. Major foreign policy actions, scandals, and other events can influence how respondents perceive the president. Short-term surges in presidential support have often followed momentous foreign policy events, for example, such as the attacks of 9/11 or the response to the invasion of Kuwait in 1991. Such rally-around-the-flag effects resulted in a net-positive shift of nearly 30 points in approval for President George W. Bush in 2001 following the 9/11 attacks (Mueller, 1973; see also Erikson et al., 2002). Generally, the impact of events dissipates over time, although traces of event-related effects on opinion have often lingered (Campbell, 2000; cf. more generally Page and Shapiro, 1992, on changes in the public's policy preferences).

## 7. Present and future challenges in polling

### 7.1. *Response rates and nonresponse bias*

The proliferation of telephone surveying, along with the growth of market research, telemarketing, and telephone solicitations in the United States since the 1970s created new challenges for pollsters and for the study of public opinion and voting. These challenges also have begun to occur in other countries as well. Telephone calling became increasingly disruptive, and household members became less willing to participate in such conversations. New technology also gave potential respondents the ability to screen calls through answering machines and caller identification devices. Decreased response rates increased the potential for "nonresponse" or selection bias in polls (See Chapter 11). In-person surveys are still conducted in some academic studies, such as the General Social Survey and the American National Election Studies, and government surveys including the U.S. Census. These surveys have had larger budgets and could maintain high-response rates through greater public relations and spending efforts. The trend in nonresponse in telephone surveys due to both noncontact and refusals was steeper from 1996 to 2003 than from 1979 to 1996 (Curtin et al., 2005; see also Zukin, 2006).

Despite this increase in nonresponse in the United States, there has been, surprisingly, little significant bias found *thus far* in comparisons of surveys with low (less than 30%) and substantially higher response rates (60% or more). What is still not known is whether there is still bias related to the sizeable hard core portion of the public that never responds (cf. Groves, 2006; Groves and Couper, 1998; Singer, 2006; Weisberg, 2005; Keeter et al., 2006). Exit polls have also faced increasing nonresponse.

## *7.2. Technological issues*

### *7.2.1. Polls using interactive voice response*

Rising costs in telephone polling and the increasing demand for polls has spurred not only the development of on-line surveys, but also the use of interactive voice response (IVR) technology. This methodology is an offshoot of audio Computer Assisted Self Interviewing (audio-CASI or ACASI). IVR polls are also referred to as automated polls or "robo-polls," which use a computer assisted polling method that replaces human interviewers with a prerecorded voice asking a short set of survey questions. Depending on the technology, respondents provide their answers verbally or key in responses on their touch-tone phones.

One advantage that IVR pollsters emphasize is that IVR controls and makes uniform how questions are asked, and how responses are received and data entered (though there can be respondent errors). A major disadvantage is that these surveys work best if limited to no more than 5 minutes of questions, which means less data can be collected. In addition, they are likely to have more break-offs because respondents are not hanging up rudely on a person who has attempted to build rapport, and respondents might have no hesitation to offer flippant or false responses. Consequently, the main use for these polls is to collect specific opinions and the most relevant background characteristics. Unless the initial introduction and screening of respondents is done by a human interviewer, these surveys may interview individuals who are not members of the sample of appropriate age, voter eligibility, or whatever required characteristic (see Couper et al., 2004; Li, 2006).

At this writing, IVR surveys have performed well in a number of election contests in many states and localities, as well as nationally, but there are also examples of preelection difficulties (see Blumenthal, 2005).

### *7.2.2. Cellular phones*

For telephone surveys, coverage issues in the past were limited to noncoverage of households without telephones and over-coverage of households with multiple phone lines. The latter can be dealt with by statistical weighting, the former by demographic adjustments. There are concerns about the increasing use of cellular phones not only to supplement regular "land-line" phones but also to replace such phones (see Lavrakas, 2007). According to a 2006 study conducted by the Pew Research Center for the People and the Press (2006), an estimated 7–9% of the American public was "cell phone only" in 2006, 53% of the public had access to both a landline and a cell phone, and 37% had a landline only. The remainder had no telephone access. Subsequently, the January–June 2008 National Health Interview Survey estimated cell phone-only households at 17.5%. Approximately 2% of households had no phone (Blumberg and Luke, 2008).

Thus far, surveys have dealt with this successfully through weighting the data to adjust for the typically young age and other characteristics of individuals who have no telephone or only cell phones at home. These phone issues were not a factor in 2004 preelection poll estimates, based on self-reports of household phone coverage in the 2004 U.S. national election exit poll (Keeter, 2006). The Pew study (2006) discussed earlier also found only minimal differences between cell-only respondents and those reachable by landline on key political questions, once appropriate weighting procedures were implemented. However, in a subsequent study, Keeter et al. (2007) did find that by including a cell phone-only sample with a standard RDD they could produce population estimates that were nearly the same as those from a landline-only sample, they also found evidence that the noncoverage of young adults (fully 25% of whom had only cell phones) in RDD surveying created biased estimates on certain survey measures.

As the proportion of cell phone-only households increases in ways that might lead to greater biases, it is likely to affect further survey response rates and costs. It is already clear that respondents aged 18–34 have become much harder to reach and that the "portability" of phone numbers in the United States has made it increasingly difficult for sampling purposes to identify the geographic residence of cell phone users (see Zukin, 2006). Further studies of cell phone users are under way to determine the feasibility and effectiveness of interviewing respondents on their cell phones, and compensating respondents for any costs incurred in receiving survey calls (Brick et al., 2007). Response rates for a cell phone sample frame are typically lower than a landline sample. The contact rate for the cell phone sample may be higher, although greater accessibility has not lead to a higher rate of cooperation; in one study half of the people reached in the landline sample (50%) cooperated, when compared with 28% of those reached in the cell phone sample. However, interviewers working on the survey reported that cell phone respondents were as focused and cooperative as those reached on a landline telephone (Pew Research Center, 2006; see also Lavrakas, 2007).

Telephone surveying has greater problems in other countries that do not have the extensive availability of land lines. In some places cell phone penetration is very great (Zukin, 2006), and in others land-line expansion has been by-passed by the large-scale introduction of cell phones. Response rates are a major issue everywhere. In some countries it is still necessary to do in-person interviewing, and in others researchers are turning to the Internet and Internet panels, which have become increasingly appealing (especially to interview young adults).

### *7.3. Threat of government regulation*

In contemporary politics, attacking or fending off negative polls are a normal part of campaigns. In 1992, when George H. W. Bush was trailing Bill Clinton, Bush attacked polls in more than 30 speeches: the equivalent of once in every four times he spoke publicly. In 1996, Bob Dole talked about the polls in one-third of all his speeches (Frankovic, 1998). In recent elections, campaigns and news stories frequently describe differences in preelection polls, and raise questions about methods, including queries about how likely voters are defined, question order, weighting, and assumptions about partisanship (Frankovic, 2005).

Most recently, pollsters themselves have come under direct attack – almost literally – for their election polling. There has been continuing debate and sensitivity to what might be call "polling politics" in which local news media and other public pollsters have been accused of partisan biases in their polls, in which normal variations and the occasional outlier in poll results that can occur due to chance are attributed to manipulative polling practices (Daves and Newport, 2005; Jacobs and Shapiro, 2005).

Some governments have attempted to limit the publication of poll results – both during the preelection period and on Election Day. As late as 2002, at least 30 countries had legal restrictions on the publication of preelection poll results. There had been little change in that absolute number since 1996, when at least 31 countries had embargos on the publication of political poll results on or prior to Election Day. Nine of these embargos applied to Election Day only; 46 countries (61%) had no embargo. Nine countries had increased the time restrictions between 1996 and 2002, while 15 others had decreased it, or eliminated it entirely. Countries with limits on the publication of preelection polls in 2002 included Western European countries like Portugal, Spain, and Switzerland, and countries in Asia and Latin America. In Italy, publication was allowed, but required a poll report to be accompanied by an "information note" with several specifications related to the poll, which must be published together with the results of the poll in the media and recorded on a dedicated website (Spangenberg, 2003).

Several nations ban the publication of preelection poll results at certain stages of the campaign. In Canada, for example, poll results cannot be published during the final 3 days of the campaign. Greece, Italy, and Ukraine are even more restrictive, prohibiting poll publication for the final 15 days of the electoral campaign (Plasser and Plasser, 2002). South Africa bans poll publication for the last 6 weeks of a campaign. In Lithuania, poll publication is prohibited for the entire duration of the official campaign period.

The regulatory framework as it applies to public and media preelection polls varies in other meaningful ways cross-nationally. According to data provided by the ACE Electoral Knowledge Network, 16 countries required the sponsor of a poll to be indicated. Disclosure of the sample characteristics is required by law in 17 countries, and the margin of error is legally required to be disclosed in 13 countries including Albania, Portugal, and Russia (ACE Electoral Knowledge Network, 2006).

The U.S. Congress held hearings after the early projection of a Ronald Reagan victory in the 1980 election (as it had after a previous electoral landslide in 1964), and some claimed that projections of a Ronald Reagan victory before all polls had closed affected turnout in Western states (Jackson, 1983). Some states, including Oregon, passed laws restricting those polls. As of 2002, 41 countries restricted publication or broadcast of poll results until after the polling places have closed. In addition, in both the United States and in Hong Kong, there are no government regulations about the release of exit poll information, but pollsters and news organizations have agreed not to report exit poll results until after the polls close (Spangenberg, 2003).

Bans on reporting preelection polls have been circumvented by posting results on the Internet, and restrictions (whether government-imposed or self-imposed) on reporting exit polls before polls close have also been circumvented, as leaks of exit polling results and their reporting on the Internet have become routine. In 2004, early leaks of partial exit poll results (with the overestimate of the Kerry vote) fueled speculation of voter fraud that continued even after the election.

Despite near-universal belief that poll information affects voting, there is minimal supporting evidence. According to one review of studies about the impact of election polling. “The conclusion is that any effects are difficult to prove and in any case are minimal. Opinion polls do provide a form of ‘interpretative assistance’ which helps undecided voters make up their mind. But the media are full of such interpretative aids, including interviews and commentaries, and in this perspective, election polls are a relatively neutral and rational interpretative aid” (Donsbach, 2001, p. 12; see also, Adams, 2005).

### *7.4. Issues in reporting polls accurately*

Beyond the technical details in conducting and analyzing polling data, how poll results are reported and interpreted is itself an issue. News stories about polls are increasingly common. One recent estimate is that the number of stories reporting on poll results has nearly doubled from the 1992 and 1996 U.S. Presidential election years to 2000 and 2004 (Frankovic, 2005, p. 684–685).

In the United States, problems in reporting have been affected by changes in journalism – cutbacks and 24 hour reporting leading to more reliance on poll results – and the repackaging of releases of poll results – as news. Journalists too often do not have the time or skills to evaluate fully the quality of the polls they report on (Rosenstiel, 2005).

## 8. Continued interest in public opinion and polling

Public opinion and election polling has been one of the constant features of late 20th century and early 21st century social and political life in the United States. They have been a persistent source of discussion and debate in the press, and the latest opinions of the public toward political issues and candidates are persistent topics of political contentions, and academics and commentators continue to debate the positive role for American democracy that George Gallup saw for public opinion through polling. Many critics doubt that the public is sufficiently knowledgeable, attentive to politics, skilled in interpretation and analysis, and wise enough overall to deserve attention in governing beyond casting votes on Election Day. Rather, the public should defer to political leaders and experts. The defenders of the “rationality” of public opinion argue that the public – as individuals and especially as a *collective* – was sufficiently capable of taking cues or learning from political leaders and other sources. Indeed, the public had defensible reasons for its opinions to warrant ongoing consideration in the political process.

This has raised classic questions: To what extent do political leaders follow or lead public opinion? What are the implications for this for democracy? Some critics have also argued that the existence of polling gives the public the false sense that its voice is amply represented in the political process (see Ginsberg, 1986; Herbst, 1993). Although the common wisdom, beginning with George Gallup, was that polls enabled political leaders to learn about public opinion and, under electoral pressure, to follow the public’s wishes, political leaders are not always under such immediate pressure; they have room to maneuver and attempt to lead – or even manipulate – public opinion, using polling information to learn how best to “craft” their messages for this purpose (see Glynn et al., 2004, Chapter 9; Jacobs and Shapiro, 2000).

Polls continue to vary in their type, scope, and quality; innovations may create new problems for pollsters to wrestle with. As survey researchers try to get good estimates of the public's opinions and behavior, the problems cited in the early days of polling remain the same: sampling coverage; identifying and representing an identifiable population and obtaining a sufficiently high response rate; statistical sampling error, assuming there is acceptable sampling coverage; the effects of questions wording; the use of fixed choice versus open-ended questions; the treatment of "don't know," "no opinion," "undecided" and similar types of responses; effects of question order ("context effects"); lack of clarity in the research questions being studied and assumptions about respondents' familiarity with a particular issue or topic; whether reported behavior validly represents actual behavior (in the present, past, or future); interviewer effects; and the effects of the type of survey method used ("mode effects").

Given such attention to public opinion, the accuracy of poll results and how these results are reported have become increasingly important academic and political issues. Journalists have been widely criticized for their shortcomings in reporting about poll results, often accepting them uncritically without researching the quality of polls and the questions asked in them. The opportunities first offered by polling have led to challenges for the pollsters and democratic politics on a number of fronts in the United States and (increasingly) worldwide, as the reach of democracy and survey research has expanded to more and more countries.

## Acknowledgments

We thank Kirk Wolter, one anonymous reviewer and the editor for excellent advice and helpful comments. Robert Shapiro is grateful to the Russell Sage Foundation where he was a 2006–2007 Visiting Scholar and to Columbia University's Institute for Social and Economic Research and Policy (ISERP).

# References

AAPOR (2006). *Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for RDD Telephone Surveys and In-person Household Surveys,* 4th ed. American Association for Public Opinion Research, Lenexa, KS.

Abbott, O., Marques dos Santos, M. (2007). The design of the 2011 Census Coverage Survey. *Proceedings of ISI Satellite Conference on Innovative Methodologies for Censuses in the New Millennium*, Southampton, UK. Available at: http://www.s3ri.soton.ac.uk/isi2007/papers/Paper25.pdf.

Abowd, J.M., Lane, J. (2004). New approaches to confidentiality protection: synthetic data, remote access and research data centers. In: Domingo-Ferrer, J., Torra, V. (Eds.), *Privacy in Statistical Databases*. *Lecture Notes in Computer Science*, 3050. Springer, Berlin, pp. 290–297.

Abraham, K.G., Maitland, A., Bianchi, S.M. (2006). Nonresponse in the American Time Use Survey: who is missing from the data and how much does it matter? *Public Opinion Quarterly* **70**, 676–703.

Abramowitz, A. (2006). Just weight! The case for dynamic party identification weighting. *PS: Political Science & Politics* **39**(3), 473–475.

Abramson, P.R., Aldrich, J.H., Rohde, D.W. (2007). *Change and Continuity in the 2004 and 2006 Elections*. CQ Press, Washington, DC.

ACE Electoral Knowledge Network. (2006). Comparative Data Summary. Available at: http://www.aceproject.org/epic-en. Accessed March 20, 2007.

Adams, W.C. (2005). *Election Night News and Voter Turnout: Solving the Projection Puzzle*. Lynne Rienner Publishers, Boulder, CO.

Agresti, A. (2002). *Categorical Data Analysis*, 2nd ed. John Wiley & Sons, New York.

Aires, N. (1999). Algorithms to find exact inclusion probabilities for conditional Poisson sampling and Pareto $\pi$ps sampling designs. *Methodology and Computing in Applied Probability* **4**, 457–469.

Aires, N. (2000). Comparisons between conditional Poisson sampling and Pareto $\pi$ps sampling designs. *Journal of Statistical Planning and Inference* **82**, 1–15.

Alexander, C.H. (2002). Still rolling: Leslie Kish's "rolling samples" and the American Community Survey. *Survey Methodology* **28**, 35–41. Available at: http://www.statcan.ca/english/ads/12-001-XIE/12-001-XIE20020016413.pdf.

Al-Hamad, A., Lewis, D., Silva, P.L.N. (2008). *Assessing the Performance of the Thousand Pounds Automatic Editing Procedure at the ONS and the Need for an Alternative Approach*. UN/ECE Work Session on Statistical Data Editing, Vienna, Austria.

Alho, J.M. (1990). Logistic regression in capture-recapture models. *Biometrics* **46**, 623–635.

Allen, R., Hanuschak, G., Craig, M. (2002). History of remote sensing for crop acreage in USDA's National Agricultural Statistics Service. Available at: www.nass.usda.gov/surveys/remotely_sensed_data_crop_acreage/index.asp./Accessed on 7 April 2009.

Althaus, S.L. (2003). *Collective Preferences in Democratic Politics: Opinion Surveys and the Will of the People*. Cambridge University Press, New York.

Alwin, D. (2007). *Margins of Error*. John Wiley & Sons, Hoboken, NJ.

Alzola, C., Harrell, F. (2006). *An Introduction to S and the Hmisc and Design Libraries. Report.* Vanderbilt University School of Medicine, Nashville, TN.

Amahia, G.N., Chaubey, Y.P., Rao, T.J. (1989). Efficiency of a new estimator in PPS sampling for multiple characteristics. *Journal of Statistical Planning and Inference* **21**, 74–85.

American Association of Public Opinion Research (2006). *Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys.* AAPOR, Lenexa, KS.

American National Election Study (2007). *The National Election Study 2000–2002–2004*. Available at: http://www.electionstudies.org/studypages/2004_panel/2004_panel.htm.

Anderson, B., Silver, B., Abramson, P.R. (1988). The effects of the race of the interviewer on race-related attitudes of black respondents on in SRC/CPS National Election Studies. *Public Opinion Quarterly* **52**(3), 289–324.

Andersson, C., Nordberg, L. (1998). *CLAN97–A SAS-program for Computation of Point- and Standard Error Estimates in Sample Surveys*. Statistics Sweden, Örebro, Sweden.

Angradi, T.R. (Ed.). (2006). *Environmental Monitoring and Assessment Program: Great River Ecosystems, Field Operations Manual*. EPA/620/R-06/002. US Environmental Protection Agency, Washington, DC.

Ardilly, P. (1991). Echantillonnage représentatif optimum á probabilités inégales. *Annales d'Economie et de Statistique* **23**, 91–113.

Asher, H. (2007). *Polling and the Public: What Every Citizen Should Know*, 7th ed. CQ Press, Washington, DC.

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.

Atrostic, B.K., Bates, N., Burt, G., Silberstein, A. (2001). Nonresponse in U.S. government household surveys: consistent measures, recent trends, and new insights. *Journal of Official Statistics* **17**(2), 209–226.

Australian Bureau of Statistics. (2007). *Information Paper: Measuring Net Undercount in the 2006 Population Census* (ABS Catalogue No. 2940.0.55.001). Available at: http://www.ausstats.abs.gov.au/Ausstats/subscriber.nsf/0/2D54EC8EAE7C5045CA2572D10020C8BB/$File/2940055001_2007.pdf.

Ayhan, H.O., Ekni, S. (2003). Coverage error in population censuses: the case of Turkey. *Survey Methodology* **29**, 155–165. Available at: http://www.statcan.ca/english/ads/12-001-XIE/12-001-XIE20030026780.pdf.

Bailar, B.A. (1968). Recent research on reinterview procedures. *Journal of the American Statistical Association* **63**, 41–63.

Bailar, B.A. (1975). The effects of rotation group bias on estimation from panel surveys. *Journal of the American Statistical Association* **70**, 23–30.

Bailar, B.A. (1989). Information needs, surveys, and measurement errors. In: Kasprzyk, D., Duncan, G., Kalton, G., Singh, M.P. (Eds.), *Panel Surveys.* John Wiley & Sons, New York, pp. 1–24.

Bailey, J.T., Kott, P.S. (1997). An application of multiple list frame sampling for multi-purpose surveys. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, Washington, DC, 496–500.

Baker, L.C., Bundorf, M.K., Singer, S., Wagner, T.D. (2003). *Validity of the Survey of Health and Internet and Knowledge Network's Panel and Sampling.* Unpublished report, Department of Veteran's Affairs & Stanford University.

Bankier, M.D. (1988). Power allocations: determining sample sizes for subnational areas. *The American Statistician* **42**, 174–177.

Bankier, M., Poirier, P., Lachance, M., Mason, P. (2000). A generic implementation of the nearest-neighbour imputation methodology (NIM). *Proceedings of the Second International Conference on Establishment Surveys*, Buffalo, NY, 571–578.

Bankier, M.D. (1986). Estimators based on several stratified samples with applications to multiple frame surveys. *Journal of the American Statistical Association* **81**, 1074–1079.

Bardsley, P., Chambers, R.L. (1984). Multipurpose estimation from unbalanced samples. *Journal of the Royal Statistical Society, Series C (Applied Statistics)* **33**, 290–299.

Barnard, J., Rubin, D.B. (1999). Small sample degrees of freedom with multiple imputation. *Biometrika* **86**, 949–955.

Barnett, V., Lewis, T. (1994). Outliers in Statistical Data, 3rd ed. John Wiley and Sons Inc., New-York.

Bartholomew, D.J. (1961). A method of allowing for "Not-at-Home" bias in sample surveys. *Applied Statistics* **10**, 52–59.

Bartolucci, F., Montanari, G.E. (2006). A new class of unbiased estimators of the variance of the systematic sample mean. *Journal of Statistical Planning and Inference* **136**, 1512–1525.

Basu, D. (1958). On sampling with and without replacement. *Sankhyā* **20**, 287–294.

Basu, D. (1969). Role of the sufficiency and likelihood principles in sample survey theory. *Sankhyā* **A31**, 441–454.

Basu, D. (1971). An essay on the logical foundations of survey sampling. In: Godambe, V.P., Sprott, D.A. (Eds.), *Foundations of Statistical Inference*. Holt, Rinehart and Winston, Toronto, Canada, pp. 203–233.

Basu, D., Ghosh, J.K. (1967). Sufficient statistics in sampling from a finite universe. *Proceedings of the 36th Session of International Statistical Institute*, 850–859.

Beaton, A.E., Tukey, J.W. (1974). The fitting of power series, meaning polynomials, illustrated on band-spectroscopic data. *Technometrics* **16**, 147–185.

Beatty, P.C., Willis, G.B. (2007). Research Synthesis: The practice of cogntive interviewing. *Public Opinion Quarterly* **71**, 287–311.

Beaumont, J.-F. (2000). An estimation method for nonignorable nonresponse. *Survey Methodology* **26**, 131–136.

Beaumont, J.-F. (2004). Robust estimation of a finite population total in the presence of influential units. Report for the Office for National Statistics, July 23, 2004, Newport.

Beaumont, J.-F. (2005). Calibrated imputation in surveys under a quasi-model-assisted approach. *Journal of the Royal Statistical Society 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.

Beaumont, J.-F., Haziza, D., Bocci, C. (2007). On variance estimation under auxiliary value imputation in sample surveys. Technical Report, Statistics Canada, Ottawa, Canada.

Bechtold, W.A., Patterson, P.L. (2005). The Enhanced Forest Inventory and Analysis Program—National Sampling Design and Estimation Procedures. General technical report SR-80, U.S. Department of Agriculture Forest Service Southern Research Station, Asheville, NC. Available at: http://www.srs.fs.usda.gov/pubs/gtr/gtr_srs080/gtr_srs080.pdf. Accessed on 4/7/09.

Béguin, C., Hulliger, B. (2004). Multivariate outlier detection in incomplete survey data: the epidemic algorithm and transformed rank correlations. *Journal of the Royal Statistical Society, Series A* **167**, 275–294.

Béguin, C., Hulliger, B. (2008). The BACON-EEM algorithm for multivariate outlier detection in incomplete survey data. *Survey Methodology* **34**, 91–103.

Béland, Y. (1999). Release of public use microdata files for NPHS? Mission partially accomplished. *Proceedings of the Section on Survey Research Method*s. American Statistical Association, 404–409.

Béland, Y., Dale, V., Dufour, J., Hamel, M. (2005). The Canadian Community Health Survey: building on the success from the past. *Proceedings of the Survey Research Methods Section, Joint Statistical Meetings*, American Statistical Association, pp. 2738–2746.

Belcher, R. (2003). Application of the Hidiroglou-Berthelot method of outlier detection for periodic business surveys. *Proceeding of the Survey Methods Section*, Statistical Society of Canada, June 8–11, 2003, Halifax, 25–30.

Belin, T.R., Rubin, D.B. (1995). A method for calibrating false-match rates in record linkage. *Journal of the American Statistical Association* **90**, 694–707.

Bell, P.A. (1998). Using state space models and composite estimation to measure the effects of telephone interviewing on labour force estimates. Working Paper No. 98/2, Australian Bureau of Statistics; Canberra, Australia. Available at: http://www.abs.gov.au/Websitedbs/D3110122.NSF/.

Bell, P.A., Clarke, C.F., Whiting, J.P. (2007). *An Estimating Equation Approach to Census Coverage Adjustment* (Australian Bureau of Statistics Research Paper 1351.0.55.019). Available at: http://www.ausstats.abs.gov.au/Ausstats/subscriber.nsf/0/46A26CC908FA49E3CA2572D10020C786/$File/1351055019_may%202007.pdf.

Bell, W.R. (1993). Using information from demographic analysis in post enumeration survey estimation. *Journal of the American Statistical Association* **88**, 1106–1118.

Bell, W.R. (2001). *ESCAP II: Estimation of Correlation Bias in 2000 A.C.E. Using Revised Demographic Analysis Results.* Executive Steering Committee for A.C.E. Policy II, Report No. 10, October 13, U.S. Census Bureau. Available at: http://www.census.gov/dmd/www/pdf/Report10.PDF.

Bell, W.R. (2007). A review of some recent work at the U.S. Census Bureau on dual system estimation of census coverage. Presentation given at the *ISI Satelite Conference on Innovative Methodologies for Censuses in the New Millennium*, Southampton, UK. Available at: http://www.s3ri.soton.ac.uk/isi2007/slides/Slides23.pdf.

Bellhouse, D.R. (1988). Systematic Sampling. In: Krishnaiah, P.R., Rao, C.R. (Eds.), *Handbook of Statistics Volume 6: Sampling*. Elsevier/North-Holland, Amsterdam, pp. 125–145.

Bellhouse, D.R., Rao, J.N.K. (1975). Systematic sampling in the presence of a trend. *Biometrika* **62**, 694–697.

Belli, R.F., Traugott, M.W., Beckmann, M.N. (2001). What leads to voting overreports? Contrasts of overreporters to validated voters and admitted nonvoters in the American National Election Studies. *Journal of Official Statistics* **17**(4), 479–498.

Benjamini, Y., Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B* **57**, 289–300.

Benjamini, Y., Yekutieli, Y. (2001). The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics* **29**, 1165–1188.

Benoit, W.L., Hansen, G.J., Verser, R.M. (2003). A meta-analysis of the effects of viewing U.S. presidential debates. *Communication Monographs* **70**, 335–350.

Berelson, B.R., Lazarsfeld, P.F., McPhee, W.N. (1954). *Voting: A Study of Opinion Formation in a Presidential Campaign*. University of Chicago Press, Chicago, IL.

Berger, J.O. (2003). Could Fisher, Jeffreys, and Neyman have agreed on testing? *Statistical Science* **18**, 1–12.

Berger, Y.G. (1998). Variance estimation using list sequential scheme for unequal probability sampling. *Journal of Official Statistics* **14**, 315–323.

Berger, Y.G. (2003). A modified Hájek variance estimator for systematic sampling. *Statistics in Transition* **6**, 5–21.

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 and 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., El Haj Tirari, M., Tillé, Y. (2003). Toward optimal regression estimation in sample surveys. *Australian and New-Zealand Journal of Statistics* **45**, 319–329.

Berger, Y.G., Rao, J.N.K. (2006). Adjusted jackknife for imputation under unequal probability sampling without replacement. *Journal of the Royal Statistical Society* **B68**, 531–547.

Berger, Y.G., Skinner, C.J. (2005). A jackknife variance estimator for unequal probability sampling. *Journal of the Royal Statistical Society* **B67**, 79–89.

Bernier, N., Lavallée, P. (1994). *La macro SAS CALJACK*. Statistique Canada, Division des méthodes d'enquêtes sociales, Ottawa, Canada.

Bertrand, P., Christian, B., Chauvet, G., Grosbras, J.-M. (2004). Plans de sondage pour le recensement rénové de la population. In: *Séries INSEE Méthodes: Actes des Journées de Méthodologie Statistique*. INSEE, Paris. Available at: http://jms.insee.fr/files/documents/2005/335_1-JMS2002_SESSION3_GROSBRAS_PLANS_SONDAGE_NOUVEAU_RECENSEMENT_ACTES.pdf.

Bethlehem, J.G. (1988). Reduction of nonresponse bias through regression estimation. *Journal of Official Statistics* **4**, 251–260.

Bethlehem, J.G. (2007). *Reducing the Bias of Web Survey Based Estimates*. Discussion paper 07001, Statistics Netherlands, Voorburg.

Bethlehem, J.G., Keller, W.J. (1987). Linear weighting of sample survey data. *Journal of Official Statistics* **3**, 141–153.

Bethlehem, J.G. (1987). The editing research project of the Netherlands Central Bureau of Statistics. *Proceedings of the Third Annual Research Conference of the Census Bureau*, Baltimore, MD, 194–203.

Bethlehem, J.G. (1988). Reduction of nonresponse bias through regression estimation. *Journal of Official Statistics* **4**, 251–260.

Bethlehem, J.G. (1998). *Bascula 4.0 for Adjustment Weighting*. Technical report. Statistics Netherlands, Department of Statistical Methods, Voorburg, The Netherlands.

Bethlehem, J.G. (2002). Weighting nonresponse adjustments based on auxiliary information. In: Groves, R.M., Dillman, D.A., Eltinge, J.L., Little, R.J.A. (Eds.), *Survey Nonresponse*. Wiley, New York, pp. 275–287.

Bethlehem, J.G., Hundepool, A.J. (2000). Analysing and documenting electronic questionnaires. *Research in Official Statistics* **2**, 7–32.

Bethlehem, J.G., Keller, W.J., Pannekoek, J. (1990). Disclosure control for microdata. *Journal of the American Statistical Association* **85**, 38–45.

Bickford, C.A., Mayer, C.E., Ware, K.D. (1963). An efficient sampling design for forest inventory: The Northeast Forest Resurvey. *J. For.* **61**, 826–833.

Biemer, P.P. (2001). Nonresponse bias and measurement bias in a comparison of face to face and telephone interviewing. *Journal of Official Statistics* **17**(2), 295–320.

Biemer, P.P. (2004a). An analysis of classification error for the revised current population survey employment questions. *Survey Methodology* **30**(2), 127–140.

Biemer, P.P. (2004b). Simple response variance: then and now. *Journal of Official Statistics* **20**, 417–439.

Biemer, P.P., Bushery, J. (2001). On the validity of Markov latent class analysis for estimating classification error in labor force data. *Survey Methodology* **26**(2), 136–152.

Biemer, P.P., Christ, S., Wiesen, C. (2006). Scale score reliability in the National Survey of Child and Adolescent Well-being. Internal RTI Project Report.

Biemer, P.P., Herget, D., Morton, J., Willis, G. (2000). The feasibility of monitoring field interview performance using computer audio recorded interviewing (CARI). *Proceedings of the Survey Research Methods Section, Joint Statistical Meetings*, American Statistical Association, pp. 1068–1073.

Biemer, P.P., Wiesen, C. (2002). Measurement Error Evaluation of Self-Reported Drug use: A latent class Analysis of the US National Household Survey on Drug Abuse. *Journal of the Royal Statistical Society Series A* **165**(1), 97–119.

Biemer, P.P. (1984). Methodology for optimal dual frame sample design. U.S. Bureau of the Census Statistical Research Division Report RR-84/07. Available at: http://www.census.gov/srd/papers/pdf/rr84-07.pdf.

Biemer, P.P. (1988). Measuring data quality. In: Groves, R., Biemer, P.P., Lyberg, L., Massey, J., Nicholls, W., Waksberg, J. (Eds.), *Telephone Survey Methodology.* John Wiley & Sons, New York, pp. 273–282.

Biemer, P.P., Forsman, G. (1992). On the quality of reinterview data with applications to the current population survey. *Journal of the American Statistical Association* **87**(420), 915–923.

Biemer, P.P., Groves, R.M., Lyberg, L., Mathiowetz, N.A., Sudman, S. (1991). *Measurement Errors in Surveys*. John Wiley & Sons, New York.

Biemer, P.P., Lyberg, L.E. (2003). *Introduction to Survey Quality*. John Wiley & Sons, Hoboken, NJ.

Biemer, P.P., Stokes, S.L. (1991). Approaches to modeling measurement error. In: Biemer, P.P., Tucker, C. (2001). Estimation and correction for underreporting errors in expenditure data: a Markov latent class modeling approach. *Proceedings of the International Statistical Institute*, Seoul, Korea.

Biemer, P.P., Tucker, C. (2001). Estimation and correction for underreporting errors in expenditure data: a Markov latent class modeling approach. *Proceedings of the International Statistical Institute*, Seoul, Korea, pp. 285–293.

Biemer, P.P., Woltman, H., Raglin, D., Hill, J. (2001). Enumeration accuracy in a population census: an evaluation using latent class analysis. *Journal of Official Statistics* **17**(1), 129–149.

Bienas, J.L., Lassman, D.M., Scheleur, S.A., Hogan, H. (1997). Improving outlier detection in two establishment surveys. *Statistical Data Editing (Volume 2); Methods and Techniques*, United Nations, Geneva, Switzerland.

Bilmes, J.A. (1998). *A Gentle Tutorial of the EM Algorithm and its Application for Parameter Estimation for Gaussian Mixture and Hidden Markov Models*. International Computer Science Institute, Berkeley, CA. Available at: http://ssli.ee.washington.edu/people/bilmes/mypapers/em.pdf.

Binder, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review* **51**, 279–292.

Binder, D.A., Babyak, C., Brodeur, M., Hidiroglou, M.A., Jocelyn, W. (2000). Variance estimation for two phase stratified sampling. *The Canadian Journal of Statistics* **28**(4), 751–764.

Binder, D.A., Hidiroglou, M.A. (1988). Sampling in time. In: Krishnaiah, P.R., Rao, C.R. (Eds.), *Handbook of Statistics*, vol. 6. North-Holland, Amsterdam, pp. 187–211.

Binder, D.A., Sun, W. (1996). Frequency valid multiple imputation for surveys with a complex design. *ASA Proceedings of the Section on Survey Research Methods*, American Statistical Association, pp. 281–286.

Bischoping, K., Schuman, H. (1992). Pens and polls in Nicaragua: An analysis of the 1990 pre-election surveys. *American Journal of Political Science* **36**, 331–350.

Bishop, G.F. (2004). *The Illusion of Public Opinion: Fact and Artifact in American Public Opinion Polls*. Rowman & Littlefield, Lanham, MD.

Bishop, G.F., Fisher, B.S. (1995). “Secret ballots” and self-reports in an exit-poll experiment. *Public Opinion Quarterly* **59**(4), 568–588.

Bishop, Y.M.M., Fienberg, S.E., Holland, P.W. (1975). *Discrete Multivariate Analysis*. MIT Press, Cambridge, MA.

Bjørnstad, J.F. (2007). Non-Bayesian multiple imputation. *Journal of Official Statistics* **33**, 433–452.

Blackwell, L., Akinwale, B., Antonatos, A., Haskey, J. (2005). Opportunities for new research using the post-2001 ONS Longitudinal Study. *Population Trends* **121**, 8–16.

Blair, E. (1983). Sample issues in trade area maps drawn from shopper surveys. *Journal of Marketing* **47**, 98–106.

Blair, E., Blair, J. (2006). Dual-frame web-telephone sampling for rare groups. *Journal of Official Statistics* **22**, 211–220.

Blair, J., Czaja, R. (1982). Locating a special population using random digit dialing. *Public Opinion Quarterly* **46**, 585–590.

Blankenship, A.B. (1977). *Professional Telephone Surveys*. McGraw-Hill, New York.

Blattberg, R.C., Sen, S.K. (1976). Market segments and stochastic brand choice models. *Journal of Marketing Research* **13**, 34–45.

Blien, U., Wirth, H., Muller, M. (1992). Disclosure risk for microdata stemming from official statistics. *Statistica Neerlandica* **46**, 69–82.

Blumberg, S.J., Luke, J.V. (2007). Wireless substitution: Early release of Estimates based on data from the National Health Interview Survey, July–December 2006. National Center for Health Statistics.

Blumberg, S.J., Luke, J.V. (2008). Wireless substitution: Early Release of Estimates from the national health interview survey, January–June 2008, at http://www.cdc.gov/nchs/data/nhis/earlyrelease/wireless200812.pdf. Last accessed March 2, 2009.

Blumberg, S.J., Luke, J.V., Cynamon, M.L. (2005). Telephone coverage and health survey estimates: evaluating the need for concern about wireless substitution. *American Journal of Public Health* **96**(5), 926–931.

Blumberg, S.J., Luke, J.V., Cynamon, M.L., Frankel, M.R. (2006). Recent trends in household telephone coverage in the United States. In: Lepkowski, J.M., Tucker, C., Brick, J.M., De Leeaw, E.D., Japec, L., Lavrakas, P.J., Link, M.W., Sangster, R.L. (Eds.), *Advances in Telephone Survey Methodology.* Wiley, New York.

Blumenthal, M.M. (2005). Toward an open-source methodology: What we can learn from the blogosphere. *Special Issue: Polling Politics, Media, and Election Campaigns. Public Opinion Quarterly* **69**(5), 655–669.

Blumenthal, S. (1982). *The Permanent Campaign*. Simon and Schuster, New York.

Blumer, H. (1948). Public opinion and public opinion polling. *American Sociological Review* **13**, 542–549.

Bocci, C., Beaumont, J.-F. (2006). Dealing with the problem of combined reports at the sampling design stage for the Workplace and Employee Survey. *Proceedings of the Survey Methods Section*, Statistical Society of Canada. Ottawa.

Bohn, T.W. (1980). Broadcasting national election returns, 1952–1976. *Journal of Communication* **30**(3), 140–153.

Bolfarine, H., Zacks, S. (1992). *Prediction Theory for Finite Populations*. Springer, New York.

Bollen, K.A. (1989). *Structural Equations with Latent Variables*. Wiley-Interscience, New York.

Bol'shev, L.N. (1965). On a characterization of the Poisson distribution. *Teoriya Veroyatnostei i ee Primeneniya* **10**, 64–71.

Bondesson, L., Traat, I., Lundqvist, A. (2004). *Pareto sampling versus Sampford and conditional Poisson sampling*. Technical Report 6, Umeå University, Sweden.

Bondesson, L., Traat, I., Lundqvist, A. (2006). Pareto sampling versus Sampford and conditional Poisson sampling. *Scandinavian Journal of Statistics* **33**, 699–720.

Booth, J.G., Butler, R.W., Hall, P. (1994). Bootstrap methods for finite population. *Journal of the American Statistical Association* **89**, 1282–1289.

Borkar, V., Deshmukh, K., Sarawagi, S. (2001). Automatic segmentation of text into structured records. *Electronic Proceedings of Association of Computing Machinery SIGMOD 2001*, Santa Barbara, California, 175–186.

Børke, S. (2008). *Using "Traditional" Control (Editing) Systems to Reveal Changes when Introducing New Data Collection Instruments*. UN/ECE Work Session on Statistical Data Editing, Vienna, Austria.

Borkowski, J.J. (1999). Network inclusion probabilities and Horvitz-Thompson estimation for adaptive simple Latin square sampling. *Environmental and Ecological Statistics* **6**, 291–311.

Boskovitz, A. (2008). *Data Editing and Logic: the Covering Set Method from the Perspective of Logic*. Ph.D. Thesis, Australian National University, Canberra, Australia.

Botman, S.L., Moore, T.F., Moriarity, C.L., Parsons, V.L. (2000). Design and estimation for the National Health Interview Survey, 1995–2004. National Center for Health Statistics. *Vital and Health Statistics* **2**(130). Available at: http://www.cdc.gov/nchs/data/series/sr_02/sr02_130.pdf.

Boucher, L. (1991). Micro-editing for the Annual Survey of Manufactures: what is the value added? *Proceedings of the Bureau of the Census Annual Research Conference*, Washington, DC, USA, 765–781.

Bowley, A.L. (1906). Address to the economic science and statistics section of the British Association for the Advancement of Science. *Journal of the Royal Statistical Society* **68**, 540–588.

Bowley, A.L. (1912). Working class households in Reading. *Journal of the Royal Statistical Society* **76**, 672–701.

Bowley, A.L. (1926). Measurement of the precision obtained in sampling. *Bulletin of the International Statistical Institute* **22**, 11–62 (supplement).

Box, G.E.P. (1979). Section title from "Robustness in the strategy of scientific model building. In: Wilkinson, G.N., Launer, R.L. (Eds.), *Robustness in Statistics*. Academic Press, New York, p. 202.

Box, G.E.P., Jenkins, G.M. (1976). *Time Series Analysis, Forecasting, and Control*, revised ed. Holden-Day, Oakland, CA.

Brackstone, G.J. (1987). Issues in the use of administrative records for statistical purposes. *Survey Methodology* **13**(1), 29–43.

Brackstone, G.J., Rao, J.N.K. (1979). An investigation of raking ratio estimators. *Sankhyā*, *Series C* **41**, 97–114.

Brady, H.E., Orren, G.R. (1992). Polling pitfalls: sources of error in public opinion surveys. In: Mann, T.E., Orren, G.R. (Eds.), *Media Polls in American Politics*, The Brookings Institution, Washington, DC, pp. 55–94.

Breidt, F.J. (1995). Markov chain designs for one-per-stratum sampling. *Survey Methodology* **21**, 63–70.

Breidt, F.J. (2007). Alternatives to the multiyear period estimation strategy for the American Community Survey. In: Citro, C.F., Kalton, G. (Eds.), Appendix C in *Using the American Community Survey: Benefits and Challenges*. National Research Council, Panel on the Functionality and Usability of Data from the American Community Survey, Committee on National Statistics, Division of Behavioral and Social Sciences and Education, The National Academies Press, Washington, DC, pp. 290–312. Available at: http://books.nap.edu/openbook.php?record_id=11901&page=290.

Breidt, F.J., Fuller, W.A. (1993). Regression weighting for multiphase samples. *Sankhyā Series B* **55**, 297–309.

Breidt, F.J., Fuller, W.A. (1999). Design of supplemented panel surveys with application to the National Resources Inventory. *Journal of Agricultural, Biological and Environmental Statistics* **4**(4), 391–403.

Breidt, F.J., McVey, A.M., Fuller, W.A. (1996). Two-phase estimation by imputation. *Journal of Agricultural Statistics (Golden Jubilee Issue)* **49**, 79–90.

Breiman, L., Friedman, J.H., Olsen, R.A., Stone, C.J. (1984). *Classification and Regression Trees*. Wadsworth International Group, Belmont, CA.

Brennan, M., Hoek, J. (1992). The behavior of respondents, nonrespondents, and refusers across mail surveys. *Public Opinion Quarterly* **56**, 530–535.

Brewer, K.R.W. (1963). A model of systematic sampling with unequal probabilities. *Australian Journal of Statistics* **5**, 5–13.

Brewer, K.R.W. (1975). A simple procedure for $\pi$pswor. *Australian Journal of Statistics* **17**, 166–172.

Brewer, K.R.W. (1994). Survey sampling inference, some past perspectives and present prospects. *Pakistan Journal of Statistics* **10**, 213–233.

Brewer, K.R.W. (1995). Combining design-based and model-based inference. In: Cox, B.G., Binder, D.A., Chinnappa, B.N., Christianson, A., Colledge, M.J., Kott, P.S. (Eds.), *Business Survey Methods.* Wiley, New York, pp. 589–606.

Brewer, K.R.W. (1999a). Design-based or model-based inference? Stratified random vs stratified balanced sampling. *International Statistical Review* **67**, 35–47.

Brewer, K.R.W. (1999b). Cosmetic calibration with unequal probability sampling. *Survey Methodology* **25**, 205–212.

Brewer, K.R.W. (2002). *Combined Survey Sampling Inference, Weighing Basu's Elephants*. Arnold, London.

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, Early, L.J., Hanif, M. (1984). Poisson, modified Poisson and collocated sampling. *Journal of Statistical Planning and Inference* **10**, 15–30.

Brewer, K.R.W, Early, L.J., Joyce, S.F. (1972). Selecting several samples from a single population. *Australian Journal of Statistics* **3**, 231–239.

Brewer, K.R.W., Hanif, M. (1983). *Sampling with Unequal Probabilities*. Springer-Verlag, New York.

Brick, J.M., Brick, P.D., Dipko, S., Presser, S., Tucker, C., Yuan, Y. (2007). Cell phone survey feasibility in the U.S.: sampling and calling cell numbers versus landline numbers. *Public Opinion Quarterly* **71**(1), 23–39.

Brick, J.M., Dipko, S., Presser, S., Tucker, C., Yuan, Y. (2006). Nonresponse bias in a dual frame sample of cell and landline numbers. *Public Opinion Quarterly* **70**, 780–793.

Brick, J.M., Ferraro, D., Strickler, T., Liu, B. (2002). *2002 NSAF Sample Design: Report No. 2.* Urban Institute: Methodology Reports. Available at: http://www.urban.org/UploadedPDF/900690 2002 Methodology 2.pdf.

Brick, J.M., Jones, M. (2008). Propensity to respond and nonresponse bias. *Metron*, LXVI, 51–73.

Brick, J.M., Kalton, G. (1996). Handling missing data in survey research. *Statistical Methods in Medical Research* **5**, 215–238.

Brick, J.M., Kalton, G., Kim, J.K. (2004). Variance estimation with hot-deck imputation using a model. *Survey Methodology* **30**, 57–66.

Brick, J.M., Montaquila, J., Roth, S. (2003). Identifying problems with raking estimators. *Proceedings of the Survey Research Methods Section of the American Statistical Association* (CD-ROM).

Brick, J.M., Waksberg, J. (1991). Avoiding sequential sampling with random digital dialing. *Survey Methodology* **17**(1), 27–41.

Brick, J.M., Waksberg, J., Keeter, S. (1996). Using data on interruptions in telephone service as coverage adjustments. *Survey Methodology* **22**, 185–197.

Brick, J.M., Waksberg, J., Kulp, D.W., Starer, A. (1995). Bias in list-assisted telephone samples. *Public Opinion Quarterly* **59**(10), 218–235.

Brooks, R.T., Wardrop, D.H., Perot, J.K. (1999). *Development and Application of Assessment Protocols for Determining the Ecological Condition of Wetlands in the Juniata River Watershed.* EPA/600/R-98/181. US Environmental Protection Agency, National Health and Enviromental Effects Laboratory, Western Ecology Division, Corvallis, OR.

Brown, G., Biemer, P. (2004). Estimating erroneous enumerations in the decennial census using four lists. *Proceedings of the ASA Survey Research Methods Section*, Joint Meetings of the American Statistical Association, Toronto, CN, pp. 3325–3332.

Brown, J., Abbott, O., Diamond, I. (2006). Dependence in the 2001 one-number census project. *Journal of the Royal Statistical Society, Series A* **169**, 883–902.

Brown, J., Buckner, L., Diamond, I.D., Chambers, R., Teague, A. (1999). A methodological strategy for a one number census in the UK. *Journal of the Royal Statistical Society, Series A* **162**, 247–267. Available at: http://www.blackwell-synergy.com/doi/pdf/10.1111/1467-985X.00133.

Brown, J.A. (1999). A comparison of two adaptive sampling designs. *Australian & New Zealand Journal of Statistics* **41**, 395–403.

Brown, J.A. (2003). Designing an efficient adaptive cluster sample. *Environmental and Ecological Statistics* **10**, 95–105.

Brown, J.A., Manly, B.J.F. (1998). Restricted adaptive cluster sampling. *Environmental and Ecological Statistics* **5**, 49–63.

Brown, M.B., Bromberg, J. (1984). An efficient two-stage procedure for generating random variates from the multinomial distribution. *The American Statistician* **38**, 216–219.

Bru, B. (1988). Estimation Laplaciennes. Un exemple: la recherche de la population d'un grand Empire 1785–1812. In: Mairesse, J. (Ed.), *Estimation et sondages. Cinq contributions à lhistoire de la statistique*. Economica, Paris, pp. 7–46.

Bruni, R., Reale, A., Torelli, R. (2001). Optimization techniques for edit validation and data imputation. *Proceedings of Statistics Canada Symposium 2001 "Achieving Data Quality in a Statistical Agency: a Methodological Perspective" XVIII-th International Symposium on Methodological Issues*.

Bruni, R., Sassano, A. (2001). *Logic and Optimization Techniques for an Error Free Data Collecting*. Report, University of Rome "La Sapienza", Rome, Italy.

Brunner, J.A., Brunner, G.A. (1971). Are voluntary unlisted telephone subscribers really different? *Journal of Marketing Research* **8**, 121–124.

Bryant, B.E. (1975). Respondent selection in a time of changing household composition. *Journal of Marketing Research* **12**, 129–135.

Bucklin, L.B., Carman, J. (1967). *The Design of Consumer Research Panels: Conception and Administration of the Berkeley Food Panel*. Institute of Business and Economic Research, University of California, Berkeley, CA.

Bucks, B.K., Kennickell, A.B., Moore, K.B. (2006). Recent changes in U.S. family finances: evidence from the 2001 and 2004 survey of consumer finances. *Federal Reserve Bulletin*. Available at: http://www.federalreserve.gov/pubs/bulletin/2006/financesurvey.pdf.

Budzinsky, C.D. (1991). Automated Spelling Correction. Statistics Canada Technical Report, Ottawa.

Bunge, J., Fitzpatrick, M. (1993). Estimating the number of species: a review. *Journal of the American Statistical Association* **88**, 364–373.

Burkhauser, R., Lillard, D.R. (2007). The expanded Cross-National Equivalent File: HILDA joins its international peers. *Australian Economic Review* **40**, 208–215.

Bush, A.J., Hair, J.F. Jr. (1985). An assessment of the mall intercept as a data collection method. *Journal of Marketing Research* **22**, 158–167.

Bynner, J. (2004). Longitudinal cohort designs. In: Kempf-Leonard, K. (Ed.), *Encyclopedia of Social Measurement*, vol. 2. Elsevier/Academic Press, Boston, pp. 591–599.

Calderwood, L., Lessof, C. (2009). Enhancing longitudinal surveys by linking to administrative data. In: Lynn, P. (Ed.), *Methodology of Longitudinal Surveys*. John Wiley & Sons, Chichester, U.K., pp. 55–72.

Campbell, A., Converse, P.E., Miller, W.E., Stokes, D.E. (1960). *The American Voter*. University of Chicago Press, Chicago, IL.

Campbell, A., Gurin, G., Miller, W.E. (1954). *The Voter Decides*. Row, Peterson and Company, White Plains, NY.

Campbell, C. (1980). A different view of finite population estimation. *Proceedings of the Survey Research Methods Section of the American Statistical Accociation*, Baltimore, MD, 319–324.

Campbell, D., Fiske, D. (1959). Convergent and discriminant validation by the multi-trait-multi-method matrix. *Psychological Bulletin* **6**, 81–105.

Campbell, J. (2000). *The American Campaign.* Texas A&M University Press, College Station, TX.

Cantor, D. (1989). Substantive implications of longitudinal design features: The National Crime Survey as a case study. In: Kasprzyk, D., Duncan, G., Kalton, G. Singh, M.P. (Eds.), *Panel Surveys*. John Wiley & Sons, New York, pp. 25–51.

Cantor, D. (2007). A review and summary of studies on panel conditioning. In: Menard, S. (Ed.), *Handbook of Longitudinal Research: Design, Measurement and Analysis.* Elsevier, San Diego, CA, pp. 123–139.

Cantor, D., O'Hare, B.C., O'Connor, K.S. (2008). The use of monetary incentives to reduce non-response in random digit dial telephone surveys. In: Lepkowski, J.M., Tucker, C., Brick, J.M. et al. (Eds.), *Advances in Telephone Survey Methodology*, Wiley, New York.

Cantril, H., Strunk, M. (1951). *Public Opinion, 1935–1946*. Princeton University Press, Princeton, NJ.

Cantwell, P.J., Ernst, L.R. (1992). New developments in composite estimation for the Current Population Survey. *Proceedings of the Symposium on the Design and Analysis of Longitudinal Surveys*, Statistics Canada, Ottawa, Canada, pp. 121–130.

Carefoot, J. (1982). Copy testing with scanners. *Journal of Advertising Research* **1**(22), 25–27.

Carfagna, E., Gallego, F.J. (2005). Using remote sensing for agricultural statistics. *International Statistical Review* **73**(3), 389–404.

Caron, N. (1998). Le logiciel POULPE: aspects méthodlogiques.*Actes des Journees de Méthodologie*, INSEE, Paris.

Carroll, J. (1970). *Allocation of a Sample Between States*. Unpublished memorandum, Australian Bureau of Census and Statistics.

Carroll, R., Ruppert, D., Stephanski, L., Crainiceanu, C. (2006). *Measurement Error in Nonlinear Models*, 2nd ed. CRC Press, Boca Raton, Florida.

Casado Valero, C., Del Castillo Cuervo-Arango, F., Mateo Ayerra, J., De Santos Ballesteros, A. (1996). Quantitative data editing: quadratic programming method. Presented at the *COMPSTAT 1996 Conference*, Barcelona, Spain.

Casady, R.J., Lepkowski, J.M. (1993). Stratified telephone survey designs. *Survey Methodology* **19**, 103–113.

Casady, R.J., Snowden, C.B., Sirken, M.G. (1981). A study of dual frame estimators for the national health interview survey. *Proceedings of the Survey Research Section, American Statistical Association* 444–447.

Cassel, C.-M., Särndal, C.-E., Wretman, J.H. (1976). Some results on generalized difference estimation and generalized regression estimation for finite populations. *Biometrika* **63**, 615–620.

Cassel, C.-M., Särndal, C.-E., Wretman, J.H. (1993a). *Foundations of Inference in Survey Sampling*. Wiley, New York.

Cassel, C.-M., Särndal, C.-E., Wretman, J.H. (1993b). *Foundations of Inference in Survey Sampling.* Wiley, New York.

Casselton, W.F., Zidek, J.V. (1984). Optimal monitoring network designs. *Statistics and Probability Letters* **2**, 223–227.

Centre for Longitudinal Studies (2007). Cohort studies. Institute of Education, University of London. Available at: http://www.cls.ioe.ac.uk/.

Cervantes, I.F., Kalton, G. (2007). Methods for sampling rare populations in telephone surveys. In: Lepkowski, J.M., Tucker, C., Brick, J.M., de Leeuw, E., Japec, L., Lavrakas, P.J., Link, M.W., Sangster, R.L. (Eds.), *Advances in Telephone Survey Methodology*. John Wiley & Sons, New York, pp. 113–132.

Chambers, J.M., Cleveland, W.S., Kleiner, B., Tukey, P.A. (1983). *Graphical Methods for Data Analysis.* Duxbury Press, Boston, MA.

Chambers, R., Hentges, A., Zhao, X. (2004). Robust automatic methods for outlier and error detection. *Journal of the Royal Statistical Society, Series A* **167**, 323–339.

Chambers, R.L. (1986). Outlier robust finite population estimation. *Journal of the American Statistical Association* **81**, 1063–1069.

Chambers, R.L. (1996). Robust case-weighting for multipurpose establishment surveys. *Journal of Official Statistics* **12**, 3–32.

Chambers, R.L. (1997). Weighting and calibration in sample survey estimation. *Proceedings of the Conference on Statistical Science Honouring the Bicentennial of Stefano Franscini's Birth*, Birkhäuser Verlag Basel, Monte Verità, Switzerland, 125–147.

Chambers, R.L., Dorfman, A.H., Hall, P. (1992). Properties of estimators of the finite distribution function. *Biometrika* **79**, 577–582.

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**, 268–277.

Chambers, R.L., Dunstan, R. (1986). Estimating distribution functions from survey data. *Biometrika* **73**, 597–604.

Chambers, R.L., Kokic, P. (1993). Outlier robust sample survey inference. Proceedings of the 49th Session of the International Statistical Institute, Firenze, Italy, 55–72.

Chaudhury, A., Vos, J.W.E. (1988). *Unified Theory and Strategies of Survey Sampling*. North-Holland, Amsterdam.

Chambers, R.L., Kokic, P., Smith, P., Cruddas, M. (2000). Winsorization for identifying and treating outliers in business surveys. *Proceedings of the Second International Conference on Establishment Surveys*, American Statistical Association, Alexandria, Virginia, 717–726.

Chao, A. (2001). An overview of closed capture–recapture models. *Journal of Agricultural, Biological, and Environmental Statistics* **6**, 158–175.

Chao, C.-T., Thompson, S.K. (2001). Optimal adaptive selection of sampling sites. *Environmetrics* **12**, 517–538.

Chao, A., Tsay, P.K. (1998). A sample coverage approach to multiple-system estimation with application to census undercount. *Journal of the American Statistical Association* **93**, 283–293.

Chao, M.T. (1982). A general purpose unequal probability sampling plan. *Biometrika* **69**, 653–656.

Chapman, D.G. (1951). Some properties of hypergeometric distribution with applications to zoological sample censuses. *University of California Publications in Statistics* **1**, 131–159, Berkely.

Charlton, P., Ehrenberg, A. (1976). An experiment in brand choice. *Journal of Marketing Research* **13**, 152–160.

Chaudhuri, S., Gamjam, K., Ganti, V., Motwani, R. (2003). Robust and efficient match for on-line data cleaning. *Proceedings of the ACM SIGMOD 2003*, San Diego, California, 313–324.

Chaudhuri, S., Ganti, V., Motwani, R. (2005). Robust identification of fuzzy duplicates. *IEEE International Conference on Data Engineering*, 865–876.

Chauvet, G., Tillé, Y. (2005). New SAS macros for balanced sampling. In: *Actes des Journées de Méthodologie Statistique*. INSEE, Paris. Available at: http://jms.insee.fr/files/documents/2006/404_1-JMS2005_SESSION03_CHAUVET-TILLE_ACTES.pdf.

Chauvet, G., Tillé, Y. (2006). A fast algorithm of balanced sampling. *Journal of Computational Statistics* **21**, 9–31.

Chen, G., Keller-McNulty, S. (1998). Estimation of identification disclosure risk in microdata. *Journal of Official Statistics* **14**, 79–95.

Chen, H.L., Rao, J.N.K., Sitter, R.R. (2000). Efficient random imputation for missing survey data in complex survey. *Statistica Sinica* **10**, 1153–1169.

Chen, H.L., Shao, J. (2000). Nearest-neighbour imputation for survey data. *Journal of Official Statistics* **16**, 583–599.

Chen, H.L., Shao, J. (2001). Jackknife variance estimation for nearest-neighbour imputation. *Journal of the American Statistical Association* **96**, 260–269.

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**, 53–67.

Chen, J., Rao, J.N.K. (2006). Asymptotic normality under two-phase sampling designs. *Statistica Sinica*: **27**, 1047–1064.

Chen, S.X. (1998). Weighted polynomial models and weighted sampling schemes for finite population. *The Annals of Statistics* **26**, 1894–1915.

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, S.X., Liu, J.S. (1997). Statistical applications of the Poisson-binomial and conditional Bernoulli distributions. *Statistica Sinica* **7**, 875–892.

Christen, P., Churches, T., Zhu, J.X. (2002). Probabilistic name and address cleaning and standardization. *The Australian Data Mining Workshop*. Available at: http://datamining.anu.edu.au/projects/linkage.html.

Christine, M. (2006). Use of balanced sampling in the framework of the master sample for French household surveys. *Joint Statistical Meeting of the American Statistical Association*, Seattle, WA. Available at: http://www.amstat.org/ASAStore/2006_JSM_Proceedings_CD_P202C4.cfm.

Christine, M., Wilms, L. (2003). Theoretical and practical problems in constructing the MSX: how can the precision of regional extensions of national surveys be improved through additional sampling? *Proceedings of Statistics Canada Symposium 2003 Challenges in Survey Taking for the Next Decade*, Ottawa, Canada. Available at: http://www.statcan-gc.ca/pub/11-522-x/2003001/session18/7730_eng.pdf.

Christman, M.C. (1997). Efficiency of adaptive sampling designs for spatially clustered populations. *Environmetrics* **8**, 145–166.

Christman, M.C. (2000). A review of quadrat-based sampling of rare, geographically clustered populations. *Journal of Agricultural, Biological & Environmental Statistics* **5**, 168–201.

Christman, M.C. (2003). Adaptive two-stage one-per-stratum sampling. *Environmental and Ecological Statistics* **10**, 43–60.

Christman, M.C., Lan, F. (2001). Inverse adaptive cluster sampling. *Biometrics* **57**, 1096–1105.

Christman, M.C., Pontius, J.S. (2000). Bootstrap confidence intervals for adaptive cluster sampling. *Biometrics* **56**, 503–510.

Chua, T.C., Fuller, W.A. (1987). A model for multinomial response error applied to labor flows. *Journal of the American Statistical Association* **82**, 46–51.

Chung, Y.S., Kim, C. (2004). Measuring robustness for weighted distributions: Bayesian perspectives. *Statistical Papers* **45**, 15–31.

Churches, T., Christen, P., Lu, J., Zhu, J.X. (2002). Preparation of name and address data for record linkage using hidden Markov models. *BioMed Central Medical Informatics and Decision Making* **2**(9). Available at: http://www.biomedcentral.com/1472-6947/2/9/.

Citro, C.F., Kalton, G. (1993). *The Future of the Survey of Income and Program Participation*. National Academies Press, Washington, DC.

Citro, C.F., Kalton, G. (Eds.). (2007). *Using the American Community Survey: Benefits and Challenges*. National Academies Press, Washington, DC.

Clark, R.G., Steel, D. (2007). Sampling within households in household surveys. *Journal of the Royal Statistical Society, Series A* **170**, 63–82.

Clarke, R.G. (1995). Winsorization methods in sample surveys. Masters thesis, Department of Statistics, Australian National University. Canberra.

Clickner, R.P., Marker, D.A., Viet, S.M., Rogers, J., Broene, P. (2002). National survey of lead and allergens in housing volume 1: analysis of lead hazards. Final report to the Office of Healthy Homes and Lead Hazard Control of the US Department of Housing and Urban Development.

Clogg, C., Eliason, S. (1985). Some common problems in log-linear analysis. *Sociological Methods and Research* **16**, 8–14.

Cocchi, D., Fabrizi, E., Trivisano, V. (2003). A hierarchical model for the analysis of local census undercount in Italy. *Survey Methodology* **29**, 167–175. Available at: http://www.statcan.ca/english/ads/12-001-XIE/12-001-XIE20030026781.pdf.

Cochran, W.G. (1977). *Sampling Techniques*, 3rd ed. John Wiley & Sons, New York.

Cochran, W.G. (1946). Relative accuracy of systematic and stratified random samples for a certain class of populations. *The Annals of Mathematical Statistics* **17**, 164–177.

Cochran, W.G. (1953). *Sampling Techniques* (1st ed.). Wiley, New York.

Cochran, W.G. (1963). *Sampling Techniques* (2nd ed.). Wiley, New York.

Cochran, W.G. (1977). *Sampling Techniques* (3rd ed.). Wiley, New York.

Cochran, W.G. (1978). Laplace's ratio estimator. In: David, H.A. (Ed.), *Contributions to Survey Sampling and Applied Statistics: Papers in Honor of H.O. Hartley*. Academic Press, New York, pp. 3–10.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurements* **20**, 37–46.

Cohen, W.W., Ravikumar, P., Fienberg, S.E. (2003a). A comparison of string metrics for matching names and addresses. *International Joint Conference on Artificial Intelligence*, *Proceedings of the Workshop on Information Integration on the Web*, Acapulco, Mexico.

Cohen, W.W., Ravikumar, P., Fienberg, S.E. (2003b). A comparison of string distance metrics for name-matching tasks. *Proceedings of the ACM Workshop on Data Cleaning, Record Linkage and Object Identification,* Washington, DC.

Colledge, M.J. (1995). Frame and business register: an overview. In: Cox, B.G., Binder, D.A., Chinnappa, B.N., Christianson, A., Colledge, M.J., Kott, P.S. (Eds.), *Business Survey Methods.* John-Wiley and Sons, New York, pp. 21–48.

Converse, J.M. (1987). *Survey Research in the United States: Roots and Emergence, 1890–1960.* University of California Press, Berkely, California.

Converse, P.E. (1964). The nature of belief systems in mass publics. In: Apter, D. (Ed.), *Ideology and Discontent*. The Free Press, New York, pp. 206–261.

Cook, J.R., Stefanski, L.A. (1994). Simulation-extrapolation estimation in parametric measurement error models. *Journal of the American Statistical Association* **89**, 1314–1328.

Cooper, W.S., Maron, M.E. (1978). Foundations of probabilistic and utility-theoretic indexing. *Journal of the Association of Computing Machinery* **25**, 67–80.

Cordell, W., Rahmel, H. (1962). Are Nielsen ratings affected by noncooperation, conditioning, or response error? *Journal of Marketing Research* **2**, 45–49.

Cordy, C. (1993). An extension of the Horvitz-Thompson theorem to point sampling from a continuous universe. *Probability and Statistics Letters* **18**, 353–362.

Cormack, R.M., Jupp, P.E. (1991). Inference for Poisson and multinomial models for capture-recapture experiments. *Biometrika* **78**, 911–916.

Cormen, T.H., Leiserson, C.E., Rivest, R.L. (1990). *Introduction to Algorithms*. The MIT Press/McGraw-Hill Book Company, Cambridge, MA.

Cotter, J., Nealon, J. (1987). *Area Frame Design for Agricultural Surveys*. US Department of Agriculture, National Agricultural Statistics Service, Research and Applications Division, Area Frame Section.

Cotton, C., Giles, P. (1998). *The seam effect in the Survey of Labour and Income Dynamics*. SLID Research Paper Series, Catalogue No. 98-18. Statistics Canada, Ottawa, Canada.

Council of the European Union. (1998). On the organisation of a labour force sample survey in the community. T. C. o. t. E. Union (Ed.), L 77/73–77: European Communities.

Couper, M.P. (2000). Web surveys: a review of issues and approaches. *Public Opinion Quarterly* **64**, 464–494.

Couper, M.P., Baker, R.P., Bethlehem, J., Clark, C.Z.F., Martin, J., Nichols II, W.L., O'Reilly, J.M. (Eds.). (1998). *Computer Assisted Survey Information Collection*. John Wiley & Sons, New York.

Couper, M.P., Blair, J., Triplett, T. (1999). A comparison of mail and e-mail for a survey of employees in federal statistical agencies. *Journal of Official Statistics* **15**, 39–56.

Couper, M.P., Nichols II, W.L. (1998). The history and development of computer assisted survey information collection methods. In: Couper, M.P., Baker, R.P., Bethlehem, J., Clark, C.Z.F., Martin, J., Nichols II, W.L., O'Reilly, J.M. (Eds.), *Computer Assisted Survey Information Collection*. John Wiley & Sons, New York, pp. 1–21.

Couper, M.P., Ofstedal, M.B. (2009). Keeping in contact with mobile sample members. In: Lynn, P. (Ed.), *Methodology of Longitudinal Surveys*. John Wiley & Sons, Chichester, U.K., pp. 183–203.

Couper, M.P., Singer, E., Tourangeau, R. (2004). Does voice matter: An interactive voice response (IVR) experiment. *Journal of Statistics* **20**, 551–570.

Cox, B.G., Binder, D.A., Chinnappa, B.N., Christianson, A., Colledge, M.J., Kott, P.S. (Eds.). (1995). *Business Survey Methods.* John Wiley and Sons, New York.

Cox, D.R. (1969). Some sampling problems in technology. In: Johnson, O.L., Smith, H. (Eds.), *New Developments in Survey Sampling*. Wiley, New York.

Cox, L.H. (2001). Disclosure risk for tabular economic data. In: Doyle, P., Lane, J.I., Theeuwes, J.J.M., Zayatz, L.V. (Eds.), *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies*. Elsevier, Amsterdam, pp. 167–183.

Cox, L.H., Kelly, J.P., Patil, R. (2004). Balancing quality and confidentiality for multivariate tabular data. In: Domingo-Ferrer, J., Torra, V. (Eds.), *Privacy in Statistical Databases*. *Lecture Notes in Computer Science*, 3050. Springer, Berlin, pp. 87–98.

Crespi, I. (1988). *Pre-election Polling: Sources of Accuracy and Error*. Russell Sage Foundation, New York.

Crespi, L.P. (1999). Some reflections on a near half-century of U.S. government survey research abroad. *International Journal of Public Opinion Research* **11**, 361–367.

Cressie, N. (1993). *Statistics for Spatial Data*, revised ed. John Wiley & Sons, New York.

Cressie, N. (1996). Change of support & the modifiable area unit problem. *Geographical Systems* **3**, 159–180.

Cressie, N., Gotway, C.A., Grondona, M.O. (1990). Spatial prediction from networks. *Chemometrics and Intelligent Laboratory Systems I* 251–271.

Cumberland, W.G., Royall, R.M. (1981). Prediction models and unequal probability sampling. *Journal of the Royal Statistical Society, Series B* **43**, 353–367.

Curtin, R., Presser, S., Singer, E. (2005). Changes in telephone survey nonresponse error over the past quarter century. *Public Opinion Quarterly* **69**, 87–98.

Czaja, R., Blair, J., Sebestik, J.P. (1982). Respondent selection in a telephone survey: a comparison of three techniques. *Journal of Marketing Research* **19**, 381–385.

Czaja, R.F., Snowden, C.B., Casady, R.J. (1986). Reporting bias and sampling errors in a survey of a rare population using multiplicity counting rules. *Journal of the American Statistical Association* **81**, 411–419.

Dacey, M.F. (1964). A note on some number properties of a hexagonal hierarchical plane lattice. *Journal of Regional Science* **5**, 63–67.

Dagpunar, J. (1988). *Principles of Random Numbers Generation*. Clarendon, Oxford, England.

Dagum, E.B., Cholette, P., Chen, Z.G. (1998). A unified view of signal extraction, benchmarking, interpolation and extrapolation of time series. *International Statistical Review* **66**, 245–269.

Dalenius, T. (1950). The problem of optimum stratification. *Skandinavisk Aktuarietidskrift* **33**, 203–211.

Dalenius, T. (1957). *Sampling in Sweden: Contributions to the Methods and Theories of Sample Survey Practice*. Almquist and Wicksell, Stockholm, Sweden.

Dalenius, T., Hájek, J., Zubrzycki, S. (1961). On plane sampling and related geometrical problems. *Proceedings of the 4th Berkeley Symposium on Probability and Mathematical Statistics* **1**, 125–150.

Dalenius, T., Reiss, S.P. (1982). Data-swapping: a technique for disclosure control. *Journal of Statistical Planning and Inference* **6**, 73–85.

Dalton, J., Gangi, M.E. (Eds.) (2007). *Special Studies in Federal Tax Statistics, 2006.* Methodology Report Series, No. 6. Statistics of Income Division, Internal Revenue Service, Washington, DC. Available at: http://www.irs.gov/taxstats/bustaxstats/article/0,,id=168008,00.html.

Darroch, J.N., Fienberg, S.E., Glonek, G.F.V., Junker, B.W. (1993). A three-sample multiple-recapture approach to census population estimation with heterogeneous catchability. *Journal of the American Statistical Association* **88**, 1137–1148.

Das, A.C. (1950). Two-dimensional systematic sampling and the associated stratified and random sampling. *Sankhyana* **10**, 95–108.

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 of adjusting for unit non-response in sample surveys. *The Canadian Journal of Statistics* **34**, 563–579.

Daves, R.P. (2000). Who will vote? Ascertaining likelihood to vote and modeling a probable electorate in pre-election polls. In: Lavrakas, P., Traugott, M. (Eds.), *Election Polls, the News Media and Democracy*. Chatham House, New York, pp. 205–223.

Daves, R.P., Newport, F. (2005). Pollsters under attack: 2004 election incivility and its consequences. *Special Issue: Polling Politics, Media, and Election Campaigns. Public Opinion Quarterly* **69**(5), 670–681.

Davis, C.S. (1993). The computer generation of multinomial random variables. *Computational Statistics and Data Analysis* **16**, 205–217.

Davis, D., Silver, B. (2003). Stereotype threat and race of interviewer effects in a survey on political knowledge. *American Journal of Political Science* **47**(1), 33–45.

Davis, J.A., Smith, T.W., Marsden, P.V. (2005). *General Social Surveys, 1972–2004, Cumulative Codebook*. National Opinion Research Center, University of Chicago, Chicago, IL.

Davison, A.C., Sardy, S. (2007). Resampling variance estimation in surveys with missing data. *Journal of Official Statistics* **23**, 371–386.

Defays, D., Anwar, M.N. (1998). Masking microdata using micro-aggregation. *Journal of Official Statistics* **14**, 449–461.

de Heer, W. (1999). International response trends: results of an international survey. *Journal of Official Statistics* **15**(2), 129–142.

de Jong, A. (2002). *Uni-Edit: Standardized Processing of Structural Business Statistics in the Netherlands*. UN/ECE Work Session on Statistical Data Editing, Helsinki, Finland.

de Leeuw, E.D. (2005). To mix or not to mix data collection modes in surveys. *Journal of Official Statistics* **21**, 233–255.

de Leeuw, E., Callegaro, M., Hox, J., Korendijk, E., Lensvelt-Mulders, G. (2007). The influence of advance letters on response in telephone surveys: a meta-analysis. *Public Opinion Quarterly* **71**(3), 413–443.

de Leeuw, E., de Heer, W. (2002). Trends in household survey nonresponse: a longitudinal and international comparison. In: Groves, R.M., Dillman, D.A., Eltinge, J.L., Little, R.J.A. (Eds.), *Survey Nonresponse*. Wiley, New York, pp. 41–54.

Delli Carpini, M.X., Williams, B. (1994). The method is the message: focus groups as a means of social, psychological, and political inquiry. In: Delli Carpini, M.X., Huddy, L., Shapiro, R.Y. (Eds.), *Research in Micropolitics. Volume 4. New Directions in Political Psychology*. Jai Press, Greenwich, Connecticut. 57–85.

Deming, W.E. (1953). On a probability mechanism to attain an economic balance between resultant error of response and the bias of nonresponse. *Journal of the American Statistical Association* **48**, 743–772.

Deming, W.E. (1960). *Sample Design in Business Research*. John Wiley and Sons, New York.

Deming, W.E., Gleser, G.J. (1959). On the problem of matching lists by samples. *Journal of the American Statistical Association* **54**, 403–415.

Deming, W.E., Stephan, F.F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal tables 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., Laird, N.M., Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B* **39**, 1–38.

Denton, F.T. (1971). Adjustment of monthly or quarterly series to annual totals: an approach based on quadratic minimization. *Journal of the American Statistical Association* **82**, 99–102.

De Ree, S.J.M. (1999). Co-ordination of business samples using measured response burden. Paper presented at the *Conference of the International Statistical Institute in Finland*, Helsinki.

Desario, J., Langton, S., (Eds.). (1984). Symposium on citizen participation and public policy. *Policy Studies Review* **3**, 207–322.

DesJardins, D. (1997). *Experiences with Introducing New Graphical Techniques for the Analysis of Census Data*. UN/ECE Work Session on Statistical Data Editing, Prague, Czech Republic.

Deville, J.-C. (1992). Constrained samples, conditional inference, weighting: three aspects of the utilisation of auxiliary information. *Proceedings of the Workshop on the Uses of Auxiliary Information in Surveys*, Örebro, Sweden.

Deville, J.-C. (1993). Estimation de la variance pour les enquêtes en deux phases. Internal manuscript note. INSEE, France.

Deville, J.-C. (1998). La correction de la non-réponse par calage ou par échantillonnage équilibré. *Actes du colloque de la Société Statistique du Canada*, Sherbrooke, Canada, pp. 103–110.

Deville, J.-C. (1999). Variance estimation for complex statistics and estimators: linearization and residual techniques. *Survey Methodology* **25**, 193–204.

Deville, J.-C. (2000). Note sur l'algorithme de Chen, Dempster et Liu. Technical Report, CREST-ENSAI, France.

Deville, J.-C. (2006). Stochastic imputation using balanced sampling. *Joint Statistical Meeting of the American Statistical Association*, Seattle, WA. Available at: http://www.amstat.org/ASAStore/2006_JSM_Proceedings_CD_P202C4.cfm.

Deville, J.-C., Grosbras, J.-M., Roth, N. (1988). Efficient sampling algorithms and balanced sample. *COMPSTAT, Proceedings in Computational Statistics*. Physica Verlag, Heidelberg, 255–266.

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. (1994). Variance estimation for the regression imputed Horvitz-Thompson estimator. *Journal of Official Statistics* **23**, 33–40.

Deville, J.-C., Särndal, C.-E., Sautory, O. (1993). Generalized raking procedures in survey sampling. *Journal of the American Statistical Association* **88**, 1013–1020.

Deville, J.-C., Tillé, Y. (1998). Unequal probability sampling without replacement through a splitting method. *Biometrika* **85**, 89–101.

Deville, J.-C., Tillé, Y. (2004). Efficient balanced sampling, the cube method. *Biometrika* **91**, 893–912.

Deville, J.-C., Tillé, Y. (2005). Variance approximation under balanced sampling. *Journal of Statistical Planning and Inference* **128**, 411–425.

Devroye, L. (1986). *Non-Uniform Random Variate Generation*. Spinger-Verlag, New York.

De Waal, A.G., Willenborg, L.C.R.J. (1997). Statistical disclosure control and sampling weights. *Journal of Official Statistics* **13**, 417–434.

De Waal, T. (2003a). *Processing of Erroneous and Unsafe Data*. Ph.D. Thesis, Erasmus University, Rotterdam, The Netherlands.

De Waal, T. (2003b). Solving the error localization problem by means of vertex generation. *Survey Methodology* **29**, 71–79.

De Waal, T. (2005). Automatic error localisation for categorical, continuous and integer data. *Statistics and Operations Research Transactions* **29**, 57–99.

De Waal, T., Coutinho, W. (2005). Automatic editing for business surveys: an assessment of selected algorithms. *International Statistical Review* **73**, 73–102.

De Waal, T., Quere, R. (2003). A fast and simple algorithm for automatic editing of mixed data. *Journal of Official Statistics* **19**, 383–402.

De Waal, T., Renssen, R., Van de Pol, F. (2000). Graphical macro-editing: possibilities and pitfalls. *Proceedings of the Second International Conference on Establishment Surveys*, Buffalo, NY, 579–588.

Di Battista, T. (2003). Resampling methods for estimating dispersion indices in random and adaptive designs. *Environmental and Ecological Statistics* **10**, 83–93.

Dick, P., You, Y. (2003). Methods used for small domain estimation of census net undercoverage in the 2001 Canadian census. *Proceedings of the Federal Committee on Statistical Methodology (FCSM) Research Conference*, Arlington, Virginia, pp. 43–48. Available at: http://www.fcsm.gov/03papers/DickYou.pdf.

Dickson, J.P., Maclachlan, D.L. (1996). Fax surveys: return patterns and comparison with mail surveys. *Journal of Marketing Research* **33**, 108–113.

Dielman, L., Couper, M. (1995). Data quality in a CAPI survey: keying errors. *Journal of Official Statistics* **11**(2), 141–146.

Diggle, P.J., Lophaven, S. (2006). Bayesian geostatistical design. *Scandanavian Journal of Statistics* **33**, 53–64.

Dijkstra, E.W. (1968). Go to considered harmful. *Letter to Cummunications of the ACM* **11**(3), 147–148.

Dillman, D.A. (1978). *Mail and Telephone Surveys*. Wiley, New York.

Dillman, D.A. (2007). *Mail and Internet Surveys: The Tailored Design Method*, 2nd ed. Wiley, New York.

Dillman, D.A., Christian, L.M. (2005). Survey mode as a source of instability in responses across surveys. *Field Methods* **17**, 30–52.

Di Zio, M., Guarnera, U., Luzi, O. (2005a). *Improving the Effectiveness of a Probabilistic Editing Strategy for Business Data*. ISTAT, Rome, Italy.

Di Zio, M., Guarnera, U., Luzi, O. (2005b). Editing systematic unity measure errors through mixture modelling. *Survey Methodology* **31**, 53–63.

Di Zio, M., Guarnera, U., Luzi, O. (2008). *Contamination Models for the Detection of Outliers and Inuential Errors in Continuous Multivariate Data*. UN/ECE Work Session on Statistical Data Editing, Vienna, Austria.

Di Zio, S., Fontanella, L., Ippoliti, L. (2004). Optimal spatial sampling schemes for environmental surveys. *Environmental and Ecological Statistics* **11**, 397–414.

Dobra A., Karr A.F., Sanil A.P. (2003). Preserving confidentiality of high-dimensional tabulated data: statistical and computational issues. *Statistics and Computing* **13**, 363–370.

Dohrmann, S., Curtin, L.R., Mohadjer, L., Montaquila, J., Le, T. (2002). National Health and Nutrition Examination Survey: limiting the risk of data disclosure using replication techniques in variance estimation. *Proceedings of the Survey Research Methods Section*, American Statistical Association, Washington, DC, 807–812.

Doll, R., Hill, A.B. (1964). Mortality in relation to smoking: ten years' observation of British doctors. *British Medical Journal* **1**, 1399–1410, 1460–1467.

Donsbach, W. (2001). *Who's Afraid of Election Polls? Normative and Empirical Arguments for Freedom of Pre-Election Surveys*. Amsterdam: ESOMAR/WAPOR, Amsterdam.

D'Orazio, M. (2003). Estimating the variance of the sample mean in two-dimensional systematic sampling. *Journal of Agricultural, Biological, and Environmental* Statistics **8**, 280–295.

Drechsler, J., Raghunathan, T.E. (2008). *Evaluating Different Approaches for Multiple Imputation under Linear Constraints*. UN/ECE Work Session on Statistical Data Editing, Vienna, Austria.

Duchesne, P. (1999). Robust calibration estimators. *Survey Methodology* **25**, 43–56.

Dumais, J., Eghbal, S., Isnard, M., Jacod, M., Vinot, F. (1999). An alternative to traditional census taking: plans for France. *Proceedings of the Federal Committee on Statistical Methodology (FCSM) Research Conference*, Arlington, Virginia. Available at: http://www.fcsm.gov/99papers/dumais.pdf.

Dumais, J., Isnard, M. (2000). Le sondage de logements dans les grandes communes dans le cadre du recensement rénové de la population. In: *Séries INSEE Méthodes: Actes des Journées de Méthodologie Statistique*, vol. 100. INSEE, Paris, pp. 37–76.

Duncan, G.J., Hill, M.S. (1985). Conceptions of longitudinal households. Fertile or futile? *Journal of Economic and Social Measurement* **13**, 361–375.

Duncan, G.J., Kalton, G. (1987). Issues of design and analysis of surveys across time. *International Statistical Review* **55**, 97–117.

Duncan, G.T., Fienberg, S.E., Krishnan, R., Padman, R., Roehrig, S.F. (2001). Disclosure limitation methods and information loss for tabular data. In: Doyle, P., Lane, J.I., Theeuwes, J.J.M., Zayatz, L.V. (Eds.), *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies*. Elsevier, Amsterdam, pp. 135–166.

Duncan, G.T., Lambert, D. (1986). Disclosure-limited data dissemination. *Journal of the American Statistical Association* **81**, 10–28.

Duncan, G.T., Lambert, D. (1989). The risk of disclosure for microdata. *Journal of Business and Economic Statistics* **7**, 207–217.

Dunkelburg, W., Day, G. (1973). Nonresponse bias and callbacks in sample surveys. *Journal of Marketing Research* **10**, 160–168.

Durand, C., Blais, A., Larochelle, M. (2004). The polls in the 2002 French presidential election: An autopsy. *Public Opinion Quarterly* **68**, 602–622.

Durr, J.-M. (2004). *The New French Rolling Census*. Working Paper No. 2, UNECE Conference of European Statisticians, Geneva. Available at: http://www.unece.org/stats/documents/2004/11/censussem/wp.2.e.pdf.

Durr, J.-M., Dumais, J. (2002). Redesign of the French census of population. *Survey Methodology* **28**, 43–49. Available at: http://www.statcan.ca/english/ads/12-001-XIE/12-001-XIE20020016414.pdf.

ECHP (2007). European Community Household Panel (2007). Available at: http://circa.europa.eu/irc/dsis/echpanel/info/data/information.html.

EDIMBUS. (2007). *Recommended Practices for Editing and Imputation in Cross-Sectional Business Surveys*. Available at: http://edimbus.istat.it/dokeos/document/document.php.

Edison Media Research and Mitofsky International. (2005). *Evaluation of Edison/Mitofsky Election System 2004.* Prepared for the National Election Pool (NEP). Edison Media Research and Mitofsky International. January 19.

Edwards, B., Cantor, D., Moses, L. (2006). Survey response rate trends, with a focus on minority populations. Presented at the *Joint Statistical Meetings*, Seattle, WA.

Edwards, T.C. Jr., Cutler, D.R., Zimmerman, N.E., Geiser, L., Alegria, J. (2005). Model-based stratifications for enhancing the detection of rare ecological events. *Ecology* **86**, 1081–1090.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *The Annals of Statistics* **7**, 1–26.

Efron, B., Tibshirani, R., Storey, J.D., Tusher, V. (2001). Empirical Bayes analysis of a microarray experiment. *Journal of the American Statistical Association* **96**, 1151–1160.

Efron, B., Tibshirani, R.J. (1993). *An Introduction to the Bootstrap*. Chapman & Hall, New York.

Ehrenberg, A. (1960). A study of some potential biases in the operation of a consumer panel. *Applied Statistics* **9**, 20–27.

Eisinger, R.M. (2003). *The Evolution of Presidential Polling*. Cambridge University Press, New York.

Eisinger, R.M. (2005). Polling and research. In: Herrnson, P.S., Campbell, C., Ezra, M., Medvic, S.K. (Eds.), *Guide to Political Campaigns in America*. CQ Press, Washington, DC. pp. 256–269.

Elamir, E.A.H., Skinner, C.J. (2006). Record level measures of disclosure risk for survey microdata. *Journal of Official Statistics* **22**, 525–539.

Elliott, M.R., Davis, W.W. (2005). Obtaining cancer risk factor prevalence estimates in small areas: combining data from two surveys. *Applied Statistics* **54**, 595–609.

Elliott, M.R., Little, R.J.A. (2000). A Bayesian approach to combining information from a census, a coverage measurement survey, and demographic analysis. *Journal of the American Statistical Association* **95**, 351–362.

Elliott, M.R., Stettler, N. (2007). Using a mixture model for multiple imputation in the presence of outliers: the 'healthy for life' project. *Applied Statistics* **56**, 63–78.

Ellis, M. (1996). The postdiagnosis mobility of people with AIDS. *Environment & Planning A* **28**, 999–1017.

ELSA (2007). *The English Longitudinal Study of Ageing*. Available at: http://www.ifs.org.uk/elsa/

Eltinge, J.L., Yansaneh, I.S. (1997). Diagnostics for formation of nonresponse adjustment cells, with an application to income nonresponse in the U.S. Consumer Expenditure Survey. *Survey Methodology* **23**, 33–40.

Engström, P., Ängsved, C. (1994). A description of a geographical macro-editing application. Paper presented at the *Work Session on Statistical Data Editing*, Cork, Ireland.

Ericksen, E.P. (1976). Sampling a rare population: a case study. *Journal of the American Statistical Association* **71**, 816–822.

Erickson, B.H. (1979). Some problems of inference from chain data. *Sociological Methodology* **10**, 276–302.

Erikson, J., Nordberg, L. (2001). Use of administrative data as substitutes for survey data for small enterprises in the Swedish Annual Structural Business Statistics. *Proceedings of ICES II-The Second International Conference on Establishment Surveys*. 813–820.

Erikson, R.S., MacKuen, M.B., Stimson, J. (2002). *The Macro Polity*. Cambridge University Press, New York.

Erikson, R.S., Panagopoulos, C., Wlezien, C. (2004). Likely (and unlikely) voters and the assessment of campaign dynamics. *Public Opinion Quarterly* **68**(4), 588–601.

Erikson, R.S., Sigelman, L. (1995). Poll-based forecasts of midterm congressional elections: do the pollsters get it right? *Public Opinion Quarterly* **59**, 589–605.

Erikson, R.S., Tedin, K.L. (2005). *American Public Opinion: Its Origins, Content, and Impact*, 6th ed. Pearson Education, New York.

Erikson, R.S., Tedin, K.L. (2007). *American Public Opinion: Its Origins, Content, and Impact.* Updated 7th ed. Pearson Longman, New York.

Ernst, L.R. (1989). Weighting issues for longitudinal household and family estimates. In: Kasprzyk, D., Duncan, G., Kalton, G., Singh, M.P. (Eds.), *Panel Surveys*. John Wiley & Sons, New York, pp. 135–159.

Ernst, L.R. (1999). The maximization and minimization of sample overlap: A half century of results. *Proceedings of the International Association of Survey Statisticians* (Invited papers from 52$^{nd}$ session of the International Statistical Institute), Helsinki, pp. 168–182.

Ernst, L.R., Valliant, R., Casady, R.J. (2000). Permanent and collocated random number sampling and the coverage of births and deaths. *Journal of Official Statistics* **16**, 211–228.

Eskin, G. (1973). Dynamic forecasts of new product demand using a depth of repeat model. *Journal of Marketing Research* **10**, 115–129.

Esposito, R., Fox, J.K., Lin, D., Tidemann, K. (1994). ARIES: a visual path in the investigation of statistical data. *Journal of Computational and Graphical Statistics* **3**, 113–125.

Esposito, R., Lin, D., Tidemann, K. (1993). The ARIES system in the BLS current employment statistics program. *Proceedings of the International Conference of Establishment Survey*, Buffalo, New York.

Estevao, V., Hidiroglou, M.A., Särndal, C.-E. (1995). Methodological principles for a generalized estimation system at Statistics Canada. *Journal of Official Statistics* **11**, 181–204.

EUREDIT Project. (2004a). *Towards Effective Statistical Editing and Imputation Strategies—Findings of the Euredit Project, Volume 1*. Available at: http://www.cs.york.ac.uk/euredit/results/results.html.

EUREDIT Project. (2004b). *Methods and Experimental Results from the Euredit Project, Volume 2*. Available at: http://www.cs.york.ac.uk/euredit/results/results.html.

European Parliament and Council of the European Union. (2002). Amending Council Regulation (EC) No 577/98 on the organisation of a labour force sample survey in the community. T. C. o. t. E. Union (Ed.), L 308/301–302: European Communities.

Eurostat (2005). *The Continuity of Indicators during the Transition between ECHP and EU-SILC*. Office for Official Publications of the European Communities, Luxembourg.

Evans, T., Zayatz, L., Slanta, J. (1998). Using noise for disclosure limitation of establishment tabular data. *Journal of Official Statistics* **14**, 537–551.

Evfimievski, A. (2004). Privacy preserving information sharing. Ph.D. Dissertation, Cornell University, New York. Available at: http://www.cs.cornell.edu/aevf/.

Ezzati-Rice, T.M., Cohen, S.B. (2004). Design and estimation strategies in the Medical Expenditure Panel Survey for investigation of trends on health care expenditures. In: Cohen, S.B., Lepkowski, J.M. (Eds.), *Eighth Conference on Health Survey Research Methods.* U. S. National Center for Health Statistics, Hyattsville, MD, pp. 23–28.

Ezzati-Rice, T.M., Frankel, M.R., Hoaglin, D.C., Loft, D., Coronado, V.G., Wright, R.A. (2000). An alternative measure of response rate in random-digit-dialing surveys that screen for eligible subpopulations. *Journal of Economic and Social Measurement* **26**, 99–109.

Fairfield Smith, H. (1938). An empirical law describing heterogeneity in the yields of agricultural crops. *Journal of Agricultural Science* **28**, 1–23.

Farley, J., Howard, J., Lehman, D. (1976). A working system model for car buyer behavior. *Management Science* **23**, 235–247.

Farwell, K., Raine, M. (2000). Some current approaches to editing in the ABS. *Proceedings of the Second International Conference on Establishment Surveys*, Buffalo, NY, 529–538.

Fay, R.E., Train, G. (1995). Aspects of survey and model-based postcensal estimation of income and poverty characteristics for states and counties. *Proceedings of the Section on Government Statistics*, American Statistical Association, Alexandria, VA, 154–159.

Fay, R.E. (1991). A design-based perspective on missing data variance. *Proceedings of the 1991 Annual Research Conference*, Washington, DC: US Bureau of the Census, 429–440.

Fay, R.E. (1992). When are inferences from multiple imputation valid? *Proceedings of the Section on Survey Research Methods*, American Statistical Association, pp. 227–232.

Fay, R.E. (1996). Alternative paradigms for the analysis of imputed survey data. *Journal of the American Statistical Association* **91**, 490–498.

Fecso, R.S., Baskin, R., Chu, A., Gray, C., Kalton, G., Phelps, R. (2007). Design options for SESTAT for the current decade. Working Paper SRS 07-021, Division of Science Resource Statistics, U. S. National Science Foundation, Washington, DC. Available at: http://www.nsf.gov/statistics/srs07201/pdf/srs07201.pdf.

Federal Committee on Statistical Methodology. (1990). *Data Editing in Federal Statistical Agencies*. Statistical Policy Working Paper 18, U.S. Office of Management and Budget, Washington, DC.

Federal Committee on Statistical Methodology. (2005). *Report on Statistical Disclosure Limitation Methodology* (Statistical Policy Working Paper 22, 2nd Version). U.S. Office for Management and Budget, Washington, DC.

Fellegi, I.P., Holt, D. (1976). A systematic approach to automatic edit and imputation. *Journal of the American Statistical Association* **71**, 17–35.

Fellegi, I.P., Sunter, A.B. (1969). A theory for record linkage. *Journal of the American Statistical Association* **64**, 1183–1210.

Ferber, R., Sudman, S. (1974). Effects of compensation in consumer expenditure studies. *Annals of Economic and Social Measurement* **3**, 319–331.

Ferguson, D.P. (1994). An introduction to the data editing process. *Statistical Data Editing (Volume 1); Methods and Techniques*, United Nations, Geneva, Switzerland.

Fienberg, S.E. (1992). Bibliography on capture-recapture modeling with application to census undercount adjustment. *Survey Methodology* **18**, 143–154.

Fienberg, S.E., Makov, U.E. (1998). Confidentiality, uniqueness and disclosure limitation for categorical data. *Journal of Official Statistics* **14**, 385–397.

Fink, P., Ombol, E., Steen Hansen, M., Sondergaard, L., De Jonge, P. (2004). Detecting mental disorders in general hospitals by the SCL-8 scale. *Journal of Psychosomatic Research* **56**, 371–375.

Finkel, S., Guterbock, T., Borg, M. (1999). Race-of-interviewer effects in a pre-election poll: Virginia 1989. *Public Opinion Quarterly* **55**(3), 313–330.

Fisher, R.A. (1934). The effect of methods of ascertainment upon the estimation of frequencies. *Annals of Eugenics* **6**, 13–25.

Fishkin, J.S. (1997). *The Voice of the People*. Yale University Press, New Haven, CT.

Fishkin, J.S., He, B., Luskin, R.C., Siu, A. (2006). Deliberative Democracy in an Unlikely Place: Deliberative Polling in China. cdd.stanford.edu/research/papers/2006/china-unlikely.pdf.

Flanagan, P.E., Lewis, J.M. (2006). Recommendation for Unduplication in the 2010 Sample Redesign. Demographic Surveys Sample Redesign Document #2010-1.0-R-1, U.S. Census Bureau, December 12, 2006.

Fletcher, J., Thompson, H. (1974). Telephone directory samples and random telephone number generation. *Journal of Broadcasting* **18**(2), 187–191.

Forsman, G., Schreiner, I. (1991). The design and analysis of reinterview: an overview. In: Biemer, P., Groves, R., Lyberg, L., Mathiowetz, N., Sudman, S. (Eds.), *Measurement Errors in Surveys*. John Wiley & Sons, New York, pp. 279–302.

Fortune (1935). A New Technique in Journalism, 65–68, 111–124. July, 1935. vol. **XII**, 1.

Foster, H.S. (1983). *Activism Replaces Isolationism: U.S. Public Attitudes 1940–1975.* Foxhall, Washington, DC.

Francis, R.I.C. (1984). An adaptive strategy for stratified random trawl surveys. *New Zealand Journal of Marine and Freshwater Research* **18**, 59–71.

Frankel, M.R., Frankel, L.R. (1977). Some recent developments in sample survey design. *Journal of Marketing Research* **14**, 280–293.

Frankel, M.R., Srinath, K.P., Hoaglin, D.C., Battaglia, M.P., Smith, P.J., Wright, R.A., Khare, M. (2003). Adjustments for non-telephone Bias in random-digit-dialling surveys. *Statistics in Medicine* **22**, 1611–1626.

Franklin, S., Thomas, S., Brodeur, M. (2000). Robust multivariate outlier detection using Mahalanobis' distance and a modified Stahel-Donoho estimator. *Proceedings of the Second International Conference on Establishment Surveys*, June 2000, American Statistical Association, Buffalo, NY, 697–706.

Frankovic, K.A. (1998). Public opinion and polling. In: Graber, D., McQuail, D., Norris, P. (Eds.), *The Politics of News, The News of Politics*. CQ Press, Washington, DC, pp. 150–170.

Frankovic, K.A. (2005). Reporting "The Polls" in 2004. *Special Issue: Polling Politics, Media, and Election Campaigns. Public Opinion Quarterly* **69**(5), 682–697.

Frankovic, K.A. (2007). Exit polls and pre-election polls. In: Donsbach, W., Traugott, M. (Eds.), *Handbook of Public Opinion Research.* Sage, Thousand Oaks, CA. 570–579.

Freedman, D.A., Wachter, K.W. (2003). On the likelihood of improving the accuracy of the census through statistical adjustment. In: Goldstein, D.R. (Ed.), Science and Statistics: A Festschrift for Terry Speed. Institute of Mathematical Statistics Monograph 40, pp. 197–230. Available at: http://www.stat.berkeley.edu/~census/612.pdf.

Freedman, D.A., Wachter, K.W. (2007). Methods for Census 2000 and statistical adjustments. In: Turner, S., Outhwaite, W. (Eds.), to appear in the Handbook of Social Science Methodology. Sage, pp. 232–245.

Freund, R.J., Hartley, H.O. (1967). A procedure for automatic data editing. *Journal of the American Statistical Association* **62**, 341–352.

Frey, J.H. (1983). *Survey Research by Telephone*. Sage Publications, Beverly Hills, CA.

Fricker, S., Galesic, M., Tourangeau, R., Yan, T. (2005). An experimental comparison of web and telephone surveys. *Public Opinion Quarterly* **69**(3), 370–392.

Fuller, W.A., Chua, T.C. (1985). Gross change estimation in the presence of response error. *Proceedings of the Conference on Gross Flows in Labor Force Statistics*, U.S. Census Bureau and U.S. Bureau of Labor Statistics, Washington, DC, 65–77.

Fuller, W.A. (1975). Regression analysis for sample survey. *Sankhyā Series C* **37**, 117–132.

Fuller, W.A. (1987). *Measurement Error Models*. John Wiley & Sons, New York.

Fuller, W.A. (1990). Analysis of repeated surveys. *Survey Methodology* **16**, 167–180.

Fuller, W.A. (1991). Simple estimators for the mean of skewed populations. *Statistica Sinica* **1**, 137–158.

Fuller, W.A. (1993). Masking procedures for microdata disclosure limitation. *Journal of Official Statistics* **9**, 383–406.

Fuller, W.A. (1998). Replication variance estimation for two-phase samples. *Statistica Sinica* **8**, 1153–1164.

Fuller, W.A. (1999). Environmental surveys over time. *Journal of Agricultural, Biological, and Environmental Statistics* **4**(4), 331–345.

Fuller, W.A., Burmeister, L.F. (1972). Estimators for samples selected from two overlapping frames. *ASA Proceedings of the Social Statistics Section*, Alexandria, VA, pp. 245–249.

Fuller, W.A., Kim, J.K. (2005). Hot-deck imputation for the response model. *Survey Methodology* **31**, 139–149.

Fuller, W.A., Rao, J.N.K. (2001). A regression composite estimator with application to the Canadian Labour Force Survey. *Survey Methodology* **27**, 45–51.

Gabler, S. (1981). A comparison of Sampford's sampling procedure versus unequal probability sampling with replacement. *Biometrika* **68**, 725–727.

Gabler, S. (1984). On unequal probability sampling: sufficient conditions for the superiority of sampling without replacement. *Biometrika* **71**, 171–175.

Gallup, G.H., Rae, S.F. (1940). *The Pulse of Democracy*. Simon and Schuster, New York.

Gallup, G.H. (1972). *The Gallup Poll: Public Opinion 1935–1971*. Random House, New York.

Gambino, J. (1999). Discussion of "Issues in weighting household and business surveys." *52nd Session of the International Statistical Institute*, International Statistical Institute, Helsinki, Finland, 187–188.

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.

Garcia, M., Thompson, K.J. (2000). Applying the generalized edit/imputation system AGGIES to the Annual Capital Expenditures Survey. *Proceedings of the International Conference on Establishment Surveys, II*, Buffalo, New York, USA, pp. 777–789.

Garfinkel, R.S., Kunnathur, A.S., Liepins, G.E. (1986). Optimal imputation of erroneous data: categorical data, general edits. *Operations Research* **34**, 744–751.

Gates, R., Solomon, P. (1982). Research using the mall intercept: State of the art. *Journal of Advertising Research* **22**(4), 43–49.

Gbur, P.M., Fairchild, L.D. (2002). Overview of the U.S. Census 2000 long form direct variance estimation. *Proceedings of the Survey Research Methods Section*, American Statistical Association, New York City, New York, 1139–1144. Available at: http://www.amstat.org/sections/srms/Proceedings/y2002/Files /JSM2002-000567.pdf.

Geer, J.G. (1988). The effects of presidential debates on the electoral preferences for candidates. *American Politics Quarterly* **16**, 486–501.

Gelman, A., King, G. (1993). Why are American presidential election campaign polls so variable when votes are so predictable? *British Journal of Political Science* **23**, 409–451.

German Institute for Economic Research (2007). *The German Socio-Economic Panel*. Available at: http://www.diw.de/english/sop/uebersicht/index.html#1.2.

Gershunskaya, J., Huff, L. (2004). Outlier detection and treatment in the Current Employment Statistics Survey. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, Alexandria, Virginia.

Gfroerer, J., Eyerman, J., Chromy, J. (Eds.). (2002). *Redesigning an Ongoing National Household Survey: Methodological Issues.* Substance Abuse and Mental Health Services Administration, Office of Applied Studies, Rockville, MD.

Ghangurde, P.D. (1982). Rotation group bias in the LFS estimates. *Survey Methodology* **8**, 86–101.

Ghosh-Dastidar, B., Schafer, J.L. (2006). Outlier detection and editing procedures for continuous multivariate data. *Journal of Official Statistics* **22**, 487–506.

Ghosh, D., Vogt, A. (2002). Sampling methods related to Bernolli and Poisson Sampling. *Proceedings of the Joint Statistical Meetings*, American Statistical Association. Alexandria, VA, 3569–3570.

Giesbrecht, L.H., Kulp, D.W., Starer, A.W. (1996). Estimating coverage bias in RDD samples with current population survey data. *Proceedings of the Section on Survey Research Methods.* American Statistical Association, Alexandria, VA, pp. 503–508.

Giessing, S. (2001). Nonperturbative disclosure control methods for tabular data. In: Doyle, P., Lane, J.I., Theeuwes, J.J.M., Zayatz, L.V. (Eds.), *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies*. Elsevier, Amsterdam, pp. 185–213.

Giessing, S., Dittrich, S. (2006). Harmonizing table protection: results of a study. In: Domingo-Ferrer, J., Franconi, L. (Eds.), *Privacy in Statistical Databases*. *Lecture Notes in Computer Science*, 4302. Springer, Berlin, pp. 35–47.

Gill, L. (1999). OX-LINK: the Oxford medical record linkage system. In: Alvey, W., and Kilss, B. (Eds.), *Record Linkage Techniques 1997*. National Academy Press, Washington, DC, pp. 15–33.

Gillespie, A.J.R. (1999). Rationale for a national annual forest inventory. *Journal of Forestry* **97**, 16–20.

Gini, C., Galvani, L. (1929). Di una applicazione del metodo rappresentativo all' ultimo censimento italiano della popolazione (1 dicembre 1921). *Annali di Statistica VI* **4**, 1–107.

Ginsberg, B. (1986). *The Captive Public; How Mass Opinion Promotes State Power*. Basic Books, New York.

Glasser, G.J. (1962). On the complete coverage of large units in a statistical study. *Review of the International Statistical Institute* **30**(1), 28–32.

Glasser, G.J., Metzger, G.D. (1972). Random digit dialing as a method of telephone sampling. *Journal of Marketing Research* **9**, 59–64.

Glasser, G.J., Metzger, G.D. (1975). National estimates of nonlisted telephone households and their characteristics. *Journal of Marketing Research* **12**, 359–361.

Glickman, H., Nirel, R., Ben-Hur, D. (2003). False captures in capture-recapture experiments with application to census adjustment. *Bulletin of the International Statistical Institute*, The 54th Session, Contributed Papers, vol. LX, 413–414. Available at: http://isi.cbs.nl/iamamember/cd3/abstracts/single/2273.html.

Glynn, C.J., Herbst, S., O'Keefe, G.J., Shapiro, R.Y., Lindeman, M. (2004). *Public Opinion*, 2nd ed. Westview, Boulder, CO.

Godambe, V.P. (1955). A unified theory of sampling from finite populations. *Journal of the Royal Statistical Society, Series B* **17**, 269–278.

Goebel, J.J. (1998). The national resources inventory and its role in US Agriculture, *Agricultural Statistics 2000: Proceedings of the conference on agricultural statistics* organized by the National Agricultural Statistics Service of the US Department of Agriculture, under the auspices of the International Statistical Institute.

Goebel, J.J., Schmude, K.O. (1982). Planning the SCS National Resources Inventory. Arid Land Resource Inventories' Workshop. USDA, Forest Service General Technical Report, WO-28, Washington, DC, 148–153.

Gollin, A.E. (Ed.). (1980a). Polls and the News Media: A Symposium. *Public Opinion Quarterly* **44**.

Gollin, A.E. (Ed.). (1980b). Exploiting the liaison between polling and the press. *Public Opinion Quarterly* **44**, 445–461.

Gollin, A.E. (Ed.). (1987). Polling and the news media. *Public Opinion Quarterly* **52**(Part 2: Supplement: 50th Anniversary Issue), S86–S94.

González-Villalobos, A., Wallace, M.A. (1996). *Multiple Frame Agriculture Surveys*, vols. 1 and 2. Food and Agriculture Organization of the United Nations, Rome.

Goodman, L.A. (1949). On the estimation of the number of classes in a population. *The Annals of Mathematical Statistics* **20**, 572–579.

Goodman, L.A. (1973). The analysis of multidimensional contingency tables when some variables are posterior to others: a modified path analysis approach. *Biometrika* **60**, 179.

Goodman, L.A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika* **61**, 215–231.

Goodman, R., Kish, L. (1950). Controlled selection—a technique in probability sampling. *Journal of the American Statistical Association* **45**, 350–372.

Gosnell, H.F., David, M.C. (1949). Instruction and research: Public opinion research in government. *American Political Science Review* **43**, 564–572.

Gotway Crawford, C.A., Young, L.J. (2006). The support of spatial data: what is it and why does it matter? *Newsletter of the American Statistical Association Section on Statistics and the Environment* 8:1.

Gouweleeuw, J.M., Kooiman, P., Willenborg, L.C.R.J., de Wolf, P.-P. (1998). Post randomisation for statistical disclosure control: theory and implementation. *Journal of Official Statistics* **14**, 463–478.

Goyder, J. (1985). Face-to-face interviews and mailed questionnaires: the net difference in response rates. *Public Opinion Quarterly* **49**(2), 234–252.

Goyder, J. (1987). *The Silent Minority: Nonrespondents on Sample Surveys*. Westview Press, Boulder, CO.

Granquist, L. (1984). Data editing and its impact on the further processing of statistical data. *Workshop on Statistical Computing*, Budapest, Hungary.

Granquist, L. (1987). Macro-editing: the top-down method. *Statistics Sweden Report*, Stockholm, Sweden.

Granquist, L. (1990). A review of some macro-editing methods for rationalizing the editing process. *Proceedings of the Statistics Canada Symposium*, Ottawa, Canada, pp. 225–234.

Granquist, L. (1995). Improving the traditional editing process. In: Cox, B.G., Binder, D.A., Chinnappa, N., Christianson, A., Colledge, M.J., Kott, P.S. (Eds.), *Business Survey Methods.* John Wiley & Sons, New York, pp. 385–401.

Granquist, L. (1997). The new view on editing. *International Statistical Review* **65**, 381–387.

Granquist, L., Kovar, J. (1997). Editing of survey data: how much is enough? In: Lyberg, L., Biemer, P., Collins, M., De Leeuw, E., Dippo, C., Schwartz, N., Trewin, D. (Eds.), *Survey Measurement and Process Quality.* John Wiley & Sons, New York, pp. 415–435.

Green, D.P., Gerber, A.S. (2006). Can registration-based sampling improve the accuracy of midterm election forecasts? *Public Opinion Quarterly* **70**, 197–223.

Green, R.H., Young, R.C. (1993). Sampling to detect rare species. *Ecological Applications* **3**, 351–356.

Greenberg, B.V., Zayatz, L.V. (1992). Strategies for measuring risk in public use microdata files. *Statistica Neerlandica* **46**, 33–48.

Greenlees, J.S., Reece, W.S., Zieschang, K.D. (1982). Imputation of missing values when the probability of response depends on the variable being imputed. *Journal of the American Statistical Association* **77**, 251–261.

Grootaert, C. (1986). The use of multiple diaries in a household expenditure survey in Hong Kong. *Journal of the American Statistical Association* **396**(81), 938–944.

Gross, W.F., Bode, G., Taylor, J.M., Lloyd-Smith, C.W. (1986). Some finite population estimators which reduce the contribution of outliers. In: Francis, I.S., Manly, B.F.J., Lam, F.C. (Eds.), *Proceedings of the Pacific Statistical Congress*. Elsevier Science Publishers B.V., Amsterdam, The Netherlands, pp. 386–390.

Groves, R.M., Couper, M. (1998). *Nonresponse in Household Surveys*. John Wiley & Sons, New York.

Groves, R.M., Presser, S., Dipko, S. (2004). The role of topic interest in survey participation decisions. *Public Opinion Quarterly* **68**, 2–31.

Groves, R.M., Singer, E., Corning, A. (2000). Leverage-salience theory of survey participation: description and an illustration. *Public Opinion Quarterly* **64**, 299–308.

Groves, R.M. (1978). An empirical comparison of two telephone sample designs. *Journal of Marketing Research* **15**, 622–631.

Groves, R.M. (2006). Non-response rates and non-response bias in household surveys. *Public Opinion Quarterly* **70(5)**, 646–675. Special Issue.

Groves, R.M., Couper, M.P., Presser, S., Singer, E., Tourangeau, R., Acosta, G.P., Nelson, L. (2006). Experiments in producing nonresponse bias. *Public Opinion Quarterly* **70**, 720–736.

Groves, R.M., Couper, M.P. (1998). *Non-response in Household Interview Surveys*. John Wiley & Sons, New York.

Groves, R.M., Heeringa, S.G. (2006). Responsive designs for household surveys: tools for actively controlling survey errors and costs. *Journal of the Royal Statistical Society*, *Series A* **169**, 439–458.

Groves, R.M., Kahn, R.L. (1979). *Surveys by Telephone*. Academic Press, New York.

Groves, R.M., Lepkowski, J.M. (1986). An experimental implementation of a dual frame telephone sample design. *Proceedings of the Survey Research Section, American Statistical Association*.

Guggenmoos-Holzmann, I. (1996). The meaning of kappa: probabilistic concepts of reliability and validity revisited. *Journal of Clinical Epidemiology* **49**(7), 775–782.

Guggenmoos-Holzmann, I., Vonk, R. (1998). Kappa-like indices of observer agreement viewed from a latent class perspective. *Statistics in Medicine* **17**, 797–812.

Gunning, P., Horgan, J. (2004). A new algorithm for the construction of stratum boundaries in skewed populations. *Survey Methodology* **30**(2), 159–165.

Gwet, J.-P. (1998). Influential observations: identification and treatment by M-estimators. *Proceedings of the Survey Methods Section*, Statistical Society of Canada, Ottawa, 231–235.

Gwet, J.-P., Lee, H. (2000). An evaluation of outlier-resistant procedures in establishment surveys. *Proceedings of the Second International Conference on Establishment Surveys*, American Statistical Association, Alexandria, Virginia, 707–716.

Gwet, J.-P., Rivest, L.-P. (1992). Outlier resistant alternatives to the ratio estimator. *Journal of the American Statistical Association* **87**, 1174–1182.

Haberman, L. (1979). *Analysis of Qualitative Data: New Developments* (vol. 2). Academic Press, New York.

Hagan, D.E., Collier, C.M. (1983). Must respondent selection procedures for telephone surveys be invasive? *Public Opinion Quarterly* **47**, 547–556.

Hagenaars, J.A. (1988). Latent structure models with direct effects between indicators: local dependence models. *Sociological Methods and Research* **16**, 379–405.

Haines, D.E., Pollock, K.H. (1998). Combining multiple frames to estimate population size and totals. *Survey Methodology* **24**, 79–88.

Haines, D.E., Pollock, K.H., Pantula, S.G. (2000). Population size and total estimation when sampling from incomplete list frames with heterogeneous inclusion probabilities. *Survey Methodology* **26**, 121–129.

Hájek, J. (1960). Limiting distributions in simple random sampling from a finite population. *Publications of the Mathematical Institute of the Hungarian Academy* **5**, 361–374.

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). *Foundations of Statistical Inference*. Holt, Rinehart, Winston, Toronto, Canada. Chap. Discussion of an essay on the logical foudations of survey sampling, part one by D. Basu.

Hájek, J. (1981). *Sampling from a Finite Population*. Marcel Dekker, New York.

Hald, A. (1998). *A History of Mathematical Statistics from 1750 to 1930.* Wiley, New York.

Hall, P.A.V., Dowling, G.R. (1980). Approximate string comparison. *Association of Computing Machinery, Computing Surveys* **12**, 381–402.

Hamburger, T., Wallsten, W. (2006). *One Party Country: The Republican Quest for Dominance in the 21st Century*. St. Martin's Press, New York.

Hamel, N., Martineau, P. (2007). *Évaluation de la qualité des données T2 produites par la Division des données fiscales*. Internal document from Statistics Canada, Ottawa.

Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J., Stahel, W.A. (1986). *Robust Statistics: the Approach Based on Influence Functions*. John Wiley and Sons Inc., New-York.

Hanselman, D.H., Quinn, T.J., Lunsford, C., Heifetz, J., Clausen, D. (2003). Applications in adaptive cluster sampling of Gulf of Alaska rockfish. *Fishery Bulletin* **101**, 501–513.

Hansen, M.H. (1981). Discussion. *Proceedings of the American Statistical Association Survey Research Methods Section*, 53–54.

Hansen, M.H., Hurwitz, W.N., Bershad, M. (1961). Measurement errors in censuses and surveys. *Bulletin of the International Statistical Institute* **38**(2), 359–374.

Hansen, M.H., Hurwitz, W.N., Pritzker, L. (1964). The estimation and interpretation of gross differences and the simple response variance. In: Rao, C.R. (Ed.), *Contributions to Statistics (Presented to P.C. Mahalanobis on the Occasion of His 70th Birthday).* Statistical Publishing Society, Calcutta, pp. 111–136.

Hansen, M.H., Hurwitz, W.N. (1943). On the theory of sampling from finite populations. *The Annals of Mathematical Statistics* **14**, 333–362.

Hansen, M.H., Hurwitz, W.N. (1946). The problem of non-response in sample surveys. *Journal of the American Statistical Association* **41**, 517–529.

Hansen, M.H., Hurwitz, W.N., Madow, W.G. (1953). *Sample Survey Methods and Theory* (2 Volumes). Wiley, New York (republished 1993).

Hansen, M.H., Madow, W.G., Tepping, B.J. (1983). An evaluation of model dependent and probability sampling inferences in sample surveys. *Journal of the American Statistical Association* **78**, 776–793.

Hardin, D., Johnson, R. (1971). Patters of use of consumer purchase panels. *Journal of Marketing Research* **8**, 364–367.

Harris-Kojetin, B., Tucker, C. (1999). Exploring the relation of economic and political conditions with refusal rates to a government survey. *Journal of Official Statistics* **15**(2), 167–184.

Harrison, L. (1997). The validity of self-reported drug use in survey research: an overview and critique of research methods. In: Harrison, L., Hughes, A. (Eds.). The validity of Self-Reported Drug use: Improving the Accuracy of Self-Reported Drug use, National Institute on Drug Abuse, NIH Publication No. 97–4147, Washington, DC, pp. 17–36.

Hartley, H.O. (1958). Maximum likelihood estimation from incomplete data. *Biometrics* **14**, 174–194.

Hartley, H.O. (1962). Multiple frame surveys. *Proceedings of the Social Statistics Section, American Statistical Association*, Alexandria, VA, pp. 203–206.

Hartley, H.O. (1974). Multiple frame methodology and selected application. *Sankhyā* **36**, 99–118.

Hartley, H.O., Rao, J.N.K. (1962). Sampling with unequal probabilities and without replacement. *The Annals of Mathematical Statistics* **33**, 350–374.

Haslinger, A. (2004). Data matching for the maintenance of the business register of statistics Austria. *Austrian Journal of Statistics* **33**(1&2), 55–67.

Hauck, M., Cox, M. (1974). Locating a sample by random digit dialing. *Public Opinion Quarterly* **38**(2), 253–260.

Hausman, J., Wise, D. (1979). Attrition bias in experimental and panel data: the Gary income maintenance experiment. *Econometrica* **2**(47), 455–473.

Hazard, J.W., Law, B.E. (1989). Forest survey methods used in the USDA Forest Service. EPA/600/3-89/065. NTIS PB89 220 594/AS. US EPA Environmental Research Laboratory, Corvallis, OR.

Haziza, D. (2007). Inference for a ratio under imputation for missing data. *Survey Methodology* **33**, 159–166.

Haziza, D., Beaumont, J.-F. (2007). On the construction of imputation classes in surveys. *International Statistical Review* **75**, 25–43.

Haziza, D., Mecatti, F., Rao, J.N.K. (2004). Comparison of variance estimators under Rao-Sampford method: a simulation study. *Proceedings of the Survey Methods Section*, American Statistical Association, CD-Rom, Toronto, Canada.

Haziza, D., Rao, J.N.K. (2003). Inference for totals in cluster sampling under mean imputation for missing data. *Proceedings of the Statistics Canada Symposium*, Ottawa, Canada, 1–11.

Haziza, D., Rao, J.N.K. (2005). Inference for domain means and totals under imputation for missing data. *The Canadian Journal of Statistics* **33**, 149–161.

Haziza, D., Rao, J.N.K. (2006). A nonresponse model approach to inference under imputation for missing survey data. *Survey Methodology* **32**, 53–64.

Heberlein, T.A., Baumgertner, R. (1978). Factors affecting response rates to mailed questionnaires: a quantitative analysis of the published literature. *American Sociological Review* **43**, 447–462.

Heckathorn, D.D. (1997). Respondent driven sampling: a new approach to the study of hidden populations. *Social Problems* **44**, 174–199.

Heckathorn, D.D. (2002). Respondent-driven sampling II: Deriving valid population estimates from chain-referral samples of hidden populations. *Social problems*, **49**, 11–34.

Hedayat, A.S., Majumdar, D. (1995). Generating desirable sampling plans by the technique of trade-off in experimental design. *Journal of Statistical Planning and Inference* **44**, 237–247.

Hedayat, A.S., Sinha, B.K. (2003). On a sampling design for estimation of negligible accident rates involving electronic toys. *The American Statistical Association* **57**, 249–252.

Hedlin, D. (2003). Score functions to reduce business survey editing at the UK Office for National Statistics. *Journal of Official Statistics* **19**(2), 177–199.

Hedlin, D. (2008). *Local and Global Score Functions in Selective Editing*. UN/ECE Work Session on Statistical Data Editing, Vienna, Austria.

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.

Hedlin, D., Fenton, T., McDonald, J.W., Pont, M., Wang, S. (2006). Estimating the undercoverage of a sampling frame due to reporting delays. *Journal of Official Statistics* **22**(1), 53–70.

Hefter, S.P., Gbur, P.M. (2002). Overview of the U.S. Census 2000 long form weighting. *Proceedings of the Survey Research Methods Section*, American Statistical Association. New York city, New York, 1418–1423. Available at: http://www.amstat.org/sections/srms/Proceedings/y2002/Files/JSM2002-000555.pdf.

Heise, D.R. (1970). Comment on 'the estimation of measurement error in panel data'. *American Sociological Review* **35**, p. 117.

Heise, D.R. (1969). Separating reliability and stability in test-retest correlation. *American Sociological Review* **34**, 93–101.

Heith, D.J. (2004). *Polling to Govern: Public Opinion and Presidential Leadership*. Stanford University Press, Stanford, CA.

Henderson, T. (2006). *Estimating the variance of the Horvitz-Thompson estimator*. M.Phil. thesis, School of Finance and Applied Statistics, The Australian National University.

Henderson, T.S. (2006). *Estimating the variance of the Horvitz-Thompson estimator.* Honours Year Thesis. Australian National University, Canberra, Australia.

Herbst, S. (1993). *Numbered Voices: How Opinion Polling Has Shaped American Politics*. University of Chicago Press, Chicago, IL.

Herrnson, P. (2004). *Congressional Elections: Campaigning at Home and in Washington*, 4th ed. CQ Press, Washington, DC.

Herzog, T.N., Scheuren, F., Winkler, W.E. (2007). *Data Quality and Record Linkage*. Springer, New York.

Hess, J., Singer, E., Bushery, J. (1999). Predicting test-retest reliability from behavior coding. *International Journal of Public Opinion Research* **11**(4), 346–360.

Hidiroglou, M.A., Srinath, K.P. (1981). Some estimators for a population total from simple random samples containing large units. *Journal of the American Statistical Association* **76**, 690–695.

Hidiroglou, M.A. (2001). Double sampling. *Survey Methodology* **27**, 143–154.

Hidiroglou, M.A., Choudhry, H., Lavallée, P. (1991)**.** A sampling and estimation methodology for sub-annual business surveys. *Survey Methodology* **17**(2), 195–210.

Hidiroglou, M.A. (1986). The construction of a self-representing stratum of large units in survey design, *The American Statistician*, **40**, 27–31.

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**(1), 71–80.

Hidiroglou, M.A., Särndal, C.-E. (1998). Use of auxiliary information for two-phase sampling. *Survey Methodology* **24**, 11–20.

Hidiroglou, M.A., Srinath, K.P. (1993). Problems associated with designing sub-annual business surveys. *Journal of Business and Economic Statistics* **11**, 397–406.

Hill, D.H. (1994). The relative empirical validity of dependent and independent data collection in a panel survey. *Journal of Official Statistics* **10**, 359–380.

Hill, M.S. (1992). *The Panel Study of Income Dynamics. A User's Guide*. Sage Publications, Newbury Park, CA.

Hinde, R., Young, D. (1984). *Synchronised Sampling and Overlap Control Manual.* Unpublished report of the Australian Bureau of Statistics, Canberra.

Ho, F.C.M., Gentle, J.E., Kennedy, W.J. (1979). Generation of random variables from the multinomial distribution. *Proceedings of the American Statistical Association, Statistical Computing Section*, 336–339.

Hogan, H. (1993). The 1990 post-enumeration survey: operations and results. *Journal of the American Statistical Association* **88**, 1047–1060.

Hogan, H. (2003). The accuracy and coverage evaluation: theory and design. *Survey Methodology* **29**, 129–138.

Hogan, H., Wolter, K. (1988). Measuring accuracy in a post-enumeration survey. *Survey Methodology* **14**, 99–116.

Holbrook, A., Krosnick, J., Pfent, A. (2008). The causes and consequences of response rates in surveys by the news media and government contractor survey research firms. In: Lepkowski, J. M., Tucker, C., Brick, J.M. et al. (Eds.), *Advances in Telephone Survey Methodology.* Wiley, New York.

Holbrook, A., Pfent, A., Krosnick, J. (2003). Response rates in surveys by the news media and government contractor firms. Presented at the *Annual Meeting of the American Association for Public Opinion Research*, Nashville, TN.

Holbrook, T. (1996). *Do Campaigns Matter?* Sage Publications, Thousand Oaks, CA.

Holt, D., Elliot, D. (1991). Methods of weighting for unit non-response. *The Statistician* **40**, 333–342.

Holt, D., Smith, T.M.F. (1979). Post-stratification. *Journal of the Royal Statistical Society A* **142**, 33–46.

Hoogendoorn, A.W. (2004). A questionnaire design for dependent interviewing that addresses the problem of cognitive satisficing. *Journal of Official Statistics* **20**, 219–232.

Hoogland, J. (2002). *Selective Editing by Means of Plausibility Indicators*. UN/ECE Work Session on Statistical Data Editing, Helsinki, Finland.

Hoogland, J., Smit, R. (2008). *Selective Automatic Editing of Mixed Mode Questionnaires for Structural Business Statistics*. UN/ECE Work Session on Statistical Data Editing, Vienna, Austria.

Hoogland, J., Van der Pijll, E. (2003). *Summary of the Evaluation of Automatic versus Manual Editing of the Production Statistics 2000 Trade and Transport*. UN/ECE Work Session on Statistical Data Editing, Madrid, Spain.

Hopkins, D.J. (2008). No more wilder effect: when and why polls mislead about black and female candidates. Available at: http://people.iq.harvard.edu/~dhopkins/wilder13.pdf. Last accessed February 25, 2009.

Horton, J.H., Lipsitz, R. (2001). Multiple imputation in practice: comparison of software for regression models with missing variables. *The American Statistician* **55**, 244–254.

Horvitz, D.G., Thompson, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* **47**, 662–685.

Houbiers, M., Knottnerus, P., Kroese, A.H., Renssen, R.H., Snijders, V. (2003). Estimating consistent table sets: position paper on repeated weighting. *Discussion Paper*, 70. Statistics Netherlands, Voorburg, The Netherlands.

Houston, G., Bruce, A.G. (1993). Geographical editing for business and economic surveys. *Journal of Official Statistics* **9**, 81–90.

Huang, E.T., Fuller, W.A. (1978). Nonnegative regression estimation for survey data. *Proceedings of the Social Statistics Section, American Statistical Association*, Alexandria, 300–303.

Hubbard, R., Bayarri, M.J. (2003). Confusion over measures of evidence ($p$'s) versus errors ($\alpha$'s) in classical statistical testing. *The American Statistician* **57**, 171–178.

Hubble, N. (2005). *Mass Observation and Everyday Life: Culture, History, Theory*. Palgrave Macmillan, Basingstoke.

Hudson, J.C. (1967). An algebraic relation between the Losch and Christaller central plane networks. *The Professional Geographer* **19**, 133–135.

Huggins, R.M. (1989). On the statistical analysis of capture experiments. *Biometrika* **76**, 133–140.

Hughes, S., Hinkins, S. (1995). Creation of panel data from cross-sectional surveys. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, pp. 408–413.

Hui, S.L., Walter, S.D. (1980). Estimating the error rates of diagnostic tests. *Biometrics* **36**, 167–171.

Hulliger, B. (1995). Outlier robust Horvitz-Thompson estimators. *Survey Methodology* **21**, 79–87.

Hundepool, A., Van de Wetering, A., De Wolf, P.P., Giessing, S., Fischetti, M., Salazar, J.J., Caprara, A. (2005). *τ-ARGUS User's Manual - Version 3.1.* Statistics Netherlands, Voorburg, The Netherlands.

Hundepool, A., Van de Wetering, A., Ramaswamy, R., Franconi, L., Capobianchi, A., De Wolf, P.P., Domingo, J., Torra, V., Brand, R., Giessing, R. (2004). *μ-ARGUS User's Manual - Version 4.0.* Statistics Netherlands, Voorburg, The Netherlands.

Hunter, L., Carbonneau, J.-F. (2005). An active management approach to survey collection. *Proceedings of Symposium 2005: Methodological Challenges for Future Information Needs*, Statistics Canada, Ottawa, Canada.

Hurtado, A. (1994). Does similarity breed respect? Evidence from interviewer evaluations of Mexican-decent respondents in a bilingual survey. *Public Opinion Quarterly* **58**, 77–95.

Iachan, R. (1982). Systematic sampling: a critical review. *International Statistical Review* **50**, 293–303.

Iachan, R. (1983). Asymptotic theory of systematic sampling. *The Annals of Statistics* **11**, 959–969.

Iachan, R. (1985). Plane sampling. *Statistics and Probability Letters* **3**, 151–159.

Iachan, R., Dennis, M.L. (1993). A multiple frame approach to sampling the homeless and transient population. *Journal of Official Statistics* **9**, 747–764.

ICES-II (2001). *Proceedings of the Second International Conference on Establishment Surveys.* American Statistical Association, Buffalo, NY, June 17–21, 2000.

Ilieva, J., Baron, S., Healey, N. (2002). Online surveys in marketing research: pros and cons. *International Journal of Market Research* **44**(3), 361–376.

International Monetary Fund. (2001). *Guide to the General Data Dissemination System (GDDS).* IMF Statistics Department.

International Statistical Institute. (1986). Declaration on professional ethics. *International Statistical Review* **54**, 227–242.

Isaki, C.T., Fuller, W.A. (1982). Survey design under the regression super-population model. *Journal of the American Statistical Association* **77**, 89–96.

Isnard, M. (1999). *Post 2000 Census in INSEE.* Working Paper No. 29, UNECE Conference of European Statisticians, Paris.

Jabine, T.P. (1985). Flow charts: a tool for developing and understanding survey questionnaires. *Journal of Official Statistics* **1**, 189–207.

Jäckle, A. (2009). Dependent interviewing: a framework and application to current research. In: Lynn, P. (Ed.), *Methodology of Longitudinal Surveys*. John Wiley & Sons, Chichester, UK, pp. 93–111.

Jackson, J.E. (1983). Election night reporting and Voter Turnout. *American Journal of Political Science* **27**, 615–635.

Jacobs, L.R., Shapiro, R.Y. (1995). The rise of presidential polling: The Nixon White House in historical perspective. *Public Opinion Quarterly* **59**, 163–195.

Jacobs, L.R., Shapiro, R.Y. (1995–1996). Presidential manipulation of polls and public opinion: The Nixon administration and the pollsters. *Political Science Quarterly* **110**, 519–538.

Jacobs, L.R., Shapiro, R.Y. (2000). *Politicians Don't Pander: Political Manipulation and the Loss of Democratic Responsiveness*. University of Chicago Press, Chicago, IL.

Jacobs, L.R., Shapiro, R.Y. (2005). Polling politics, media, and election campaigns. In: Jacobs, L.R., Shapiro, R.Y. (Eds.), *Special Issue: Polling Politics, Media, and Election Campaigns. Public Opinion Quarterly* **69**(5), 635–641.

Jaro, M.A. (1989). Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida. *Journal of the American Statistical Association* **89**, 414–420.

Jay, G., Belli, R., Lepkowski, J. (1994). Quality of last doctor visit reports: a comparison of medical records and survey data. *Proceedings of the ASA Section on Survey Research Methods*, 362–367.

Jenkins, J.G. (1938). Dependability of psychological brand barometers: I. The problem of reliability. *Journal of Applied Psychology* **22**, 1–7, Toronto, Canada.

Jensen, R.J. (1971). *The Winning of the Midwest: Social and Political Conflict, 1888–1896.* University of Chicago Press, Chicago, IL.

Jessen, R.J. (1942). Statistical investigation of a sample survey for obtaining farm facts. *Iowa Agricultural Experiment Statistical Research Bulletin* **304**, 54–59.

Johnson, N.L., Kotz, S., Balakrishnan, N. (1997). *Discrete Multivariate Distributions*. Wiley, New York.

Johnson, W.O., Su, C-L., Gardner, I.A., Christensen, R. (2003). Sample size calculations for surveys to substantiate freedom of populations from infectious agents. *Biometrics* **60**, 165–171.

Jolly, G.M. (1965). Explicit estimates from capture-recapture data with both death and immigration - Stochastic model. *Biometrika* **52**, 225–247.

Jonasson, J., Nerman, O. (1996). *On maximum entropy πps-sampling with fixed sample size*. Technical Report, Gäoteborg University, Sweden.

Jöreskog, K.G. (1979). Statistical models and methods for analysis of longitudinal data. In: Jöreskog, K.G., Sörbom, D. (Eds.), *Advances in Factor Analysis and Structural Equation Models*. Abt Books, Cambridge, MA, pp. 129–169.

Jöreskog, K.G., Sörbom, D. (1978). *LISREL IV: Analysis of Linear Structural Relationships by the Method of Maximum Likelihood.* National Educational Resources, Inc., Chicago, IL.

Jowell, R., Hedges, B., Lynn, P., Farrant, G., Heath, A. (1993). Review: The 1992 British Election: The Failure of the Polls. *Public Opinion Quarterly* **57**, 238–263.

Juster, F.T., Suzman, R. (1995). Overview. Special issue: the Health and Retirement Study. *Journal of Human Resources* **30**, S7–S56.

Kalton, G. (1983). *Compensating for missing survey data.* University of Michigan Press, Ann Arbor, MI.

Kalton, G. (1986). Handling wave nonresponse in panel surveys. *Journal of Official Statistics* **2**, 303–314.

Kalton, G. (1993). *Sampling rare and elusive populations, National Household Survey Capability Programme.* United Nations Department of Economic and Social Information and Policy Analysis Statistics Division, New York, Available at: http://unstats.un.org/unsd/publication/unint/UNFPA_UN_INT_92_P80_16E.pdf. Retrieved December 3, 2005.

Kalton, G. (2003). Practical methods for sampling rare and mobile populations. *Statistics in Transition* **6**, 491–501.

Kalton, G., Anderson, D.W. (1986). Sampling rare populations. *Journal of the Royal Statistical Society, Series A* **149**, 65–82.

Kalton, G., Brick, J.M. (1995). Weighting schemes for household panel surveys. *Survey Methodology* **21**, 33–44.

Kalton, G., Brick, J.M. (2000). Weighting in household panel surveys. In: Rose, D. (Ed.)*, Researching Social and Economic Change*. Routledge, London, pp. 96–112.

Kalton, G., Citro, C.F. (1993). Panel surveys: adding the fourth dimension. *Survey Methodology* **19**, 205–215.

Kalton, G., Flores-Cervantes, I. (2003). Weighting methods. *Journal of Official Statistics* **18**, 81–97.

Kalton, G., Kasprzyk, D. (1986). The treatment of missing survey data. *Survey Methodology* **12**, 1–16.

Kalton, G., Kasprzyk, D., McMillen, D.B. (1989). Nonsampling errors in panel surveys. In: Kasprzyk, D., Duncan, G., Kalton, G., Singh, M.P. (Eds.), *Panel Surveys*. John Wiley & Sons, New York, pp. 249–270.

Kalton, G., Kish, L. (1984). Some efficient random imputation methods. *Communications in Statistics, Part A-Theory and Methods* **13**, 1919–1939.

Kalton, G., Maligalig, D. (1991). A comparison of methods of weighting adjustment for nonresponse. *Proceedings of the U.S. Bureau of the Census 1991 Annual Research Conference*, 409–428.

Kalton, G., Miller, M.E. (1991). The seam effect with Social Security income in the Survey of Income and Program Participation. *Journal of Official Statistics* **7**, 235–245.

Kalton, G., Winglee, M., Rizzo, L., Jabine, T., Levine, D. (1998). *SIPP Quality Profile 1998*, 3rd ed. U.S. Census Bureau, Washington DC.

Kane, E.W., Macaulay, L.J. (1993). Interviewer gender and gender attitudes. *Public Opinion Quarterly* **57**, 1–28.

Kang, J.D.Y., Schafer, J.L. (2008). Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science* **22**, 523–539.

Kasprzyk, D., Duncan, G., Kalton, G., Singh, M.P. (Eds.) (1989). *Panel Surveys*. John Wiley & Sons, New York.

Katz, D. (1942). Do interviewers bias poll results? *Public Opinion Quarterly* **6**, 248–268.

Katz, I.R., Stinson, L.L., Conrad F.G. (1997). Questionnaire designer versus instrument authors: bottlenecks in the development of computer-administered questionnaires. *Fifty-second Annual Conference of the American Association for Public Opinion Research*, Norfolk, VA, 1029–1034.

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

Kawai, H., Garcia-Molina, H., Benjelloun, O., Menestrina, D., Whang, E., Gong, H. (2006). P-Swoosh: Parallel Algorithm for Generic Entity Resolution. Stanford University CS technical report.

Kay, A.F. (1998). *Locating Consensus for Democracy: A Ten-Year U.S. Experiment*. Americans Talk Issues Foundation, St. Augustine, FL.

Keeter, S. (1995). Estimating noncoverage bias from a phone survey. *Public Opinion Quarterly* **59**, 196–217.

Keeter, S. (2006). The impact of cell phone noncoverage bias on polling in the 2004 presidential election. *Public Opinion Quarterly* **70**, 88–98.

Keeter, S., Kennedy, C., Clark, A., Thompson, T., Mokrzycki, M. (2007). What's missing from national landline RDD surveys? The impact of the growing cell phone only population, Special Issue, Lavrakas, P.J. (Ed.), *Public Opinion Quarterly* **71**(5), 772–792.

Keeter, S., Kennedy, C., Dimock, M., Best, J., Craighill, P. (2006). Gauging the impact of growing non-response on estimates from a national RDD telephone survey. In: Singer, E. (Ed.), *Special Issue: Non-response Bias in Household Surveys. Public Opinion Quarterly* **70**(5), 759–779.

Kellner, P. (2004). Can online polls accurate findings. *International Journal of Market Research* **46**(1), 3–19.

Kemp, C.D., Kemp, A.W. (1987). Rapid generation of frequency tables. *Applied Statistics* **36**, 277–282.

Kennickell, A.B., McManus, D. (1993). Sampling for household financial characteristics using frame information on past income. *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 88–97.

Kent, J.P., Willenborg, L.C.R.J. (1997). *Documenting Questionnaires*. Research Paper no. 9708, Department of Statistical Methods, Statistics Netherlands, Voorburg, The Netherlands.

Keselman, H.J., Wilcox, R.R., Othmkan, A.R., Fradette, K. (2002). Trimming, transforming statistics, and bootstrapping: circumventing the biasing effects of heteroscedasticity and nonnormality. *Journal of Modern Applied Statistical Methods* **1**, 288–309.

Kessler, F. (1986). High-tech stocks in ad research. *Fortune* **7**, 58–60.

Keyfitz, N. (1951). Sampling with probabilities proportionate to size: adjustment for changes in probabilities. *Journal of the American Statistical Association* **46**, 105–109.

Khare, M., Chowdhury, S. (2006). An evaluation of methods to compensate for noncoverage of nontelephone households using information on interruptions in telephone service and presence of wireless phones. *Proceedings of the Section on Survey Research Methods* (CD-ROM), American Statistical Association, Alexandria VA, 3221–3228.

Kim, H.-S., Lee, D. (2007). Parallel linkage. *Conference on Information and Knowledge Management '07*, Lisbon, Portugal, pp. 283–292.

Kim, J.K., Brick, J.M., Fuller, W.A., Kalton, G. (2006). On the bias of the multiple-imputation variance estimator in survey sampling. *Journal of the Royal Statistical Society B* **68**, 509–521.

Kim, J.K., Fuller, W.A. (2004). Fractional hot deck imputation. *Biometrika* **91**, 559–578.

Kim, J.K., Navarro, A., Fuller, W.A. (2006). Replication variance estimation for two-phase stratified sampling. *Journal of the American Statistical Association* **101**, 312–320.

Kim, J.K., Park, H. (2006). Imputation using response probability. *The Canadian Journal of Statistics* **34**, 171–182.

Kimura, D.K., Somerton, D.A. (2006). Review of statistical aspects of survey sampling for marine fisheries. *Reviews in Fisheries Science* **14**, 245–283.

Kish, L. (1965). *Survey Sampling*. John Wiley and Sons, New York.

Kish, L. (1987). *Statistical Design for Research*. John Wiley & Sons, New York.

Kish, L. (1988). Multipurpose sample designs. *Survey Methodology* **14**, 19–32.

Kish, L. (1990). Rolling samples and censuses. *Survey Methodology* **16**, 63–79.

Kish, L. (1992). Weighting for unequal $P_i$. *Journal of Official Statistics* **8**, 183–200.

Kish, L. (1998). Space/time variations and rolling samples. *Journal of Official Statistics* **14**, 31–46. Available at: http://www.jos.nu/Articles/abstract.asp?article=14131.

Kish, L. (1999). Cumulating/combining population surveys. *Survey Methodology* **25**, 129–138.

Kish, L., Scott, A. (1971). Retaining units after changing strata and probabilities. *Journal of the American Statistical Association* **66**, 461–470.

Klapper, J.T. (1960). *The Effects of Mass Communication*. Free Press, Glencoe, IL.

Klovdahl, A. (1989). Urban social networks: some methodological problems and possibilities. In: Kochen, M. (Ed.), *The Small World*. Ablex Publishing, Norwood, NJ, pp. 176–210.

Kohut, A. (1981). A review of the gallup pre-election poll methodology in 1980. *Proceedings American Statistical Association Survey Research Methods Section*, pp. 41–46.

Kokic, P. (1998). On Winsorization in business surveys. *Proceedings of the Survey Methods Section*, Statistical Society of Canada, Ottawa, pp. 237–239.

Kokic, P.N., Bell, P.A. (1994). Optimal Winsorizing cutoffs for a stratified finite population estimator. *Journal of Official Statistics* **10**, 419–435.

Konschnik, C.A., King, C.S., Dahl, S.A. (1991). Reassessment of the use of an area sample for the monthly retail trade survey. *Proceedings of the Survey Research Methods Section, American Statistical Association*, Alexandria, 208–213.

Korn, E., Graubard, B. (1999). *Analysis of Health Surveys*. John Wiley & Sons, New York.

Kott, P.S. (1986). Some asymptotic results for the systematic and stratified sampling of a finite population. *Biometrika* **73**, 485–491.

Kott, P.S. (1990). Variance estimation when a first phase area sample is restratified. *Survey Methodology* **16**, 99–103.

Kott, P.S. (1994). A note on handling nonresponse in sample surveys. *Journal of American Statistical Association* **89**, 693–696.

Kott, P.S. (1995). A paradox of multiple imputation. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 380–383.

Kott, P.S. (2001). The delete-a-group jackknife. *Journal of Official Statistics* **17**, 521–526.

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. Available at: http://www.nass.usda.gov/research/reports/icespap2c8.pdf.

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., Vogel, F.A. (1995). Multiple-frame business surveys. In: Cox, B.G., Binder, D.A., Chinnappa, B.N., Christianson, A., Colledge, M.J., Kott, P.S. (Eds.), *Business Survey Methods*. John-Wiley and Sons, New York, pp. 185–203.

Kovar, J.G., Whitridge, P.G. (1995). Imputation for business survey data. In: Cox, B.G., Binder, D.A., Chinanappa, B.N., Christianson, A., Colledge, M.J., Kott, P.S. (Eds.), *Business Survey Methods*. John Wiley and Sons, New York, pp. 403–423.

Kroese, A.H., Renssen, R.H. (2001). New application of old weighted techniques: constructing a consistent set of estimates based on data from different sources. *Proceedings of ICES II–The Second International Conference on Establishment Surveys*, 831–840.

Krosnick, J. (1999). Survey research. *Annual Review of Psychology* **50**, 537–567.

Kuhn, T.S. (1996). *The Structure of Scientific Revolutions* (3rd ed.). University of Chicago Press, Chicago, IL.

Kulka, R.A. (1982). Monitoring social change via survey replication: prospects and pitfalls from a replication survey of social roles and mental health. *Journal of Social Issues* **38**, 17–38.

Kweit, M.G., Kweit, R.W. (1984). The politics of policy analysis: The role of citizen participation in analytic decision making. *Policy Studies Review* **3**, 234–245.

Laflamme, F., Barrett, C., Johnson, W., Ramsay, L. (1996). Experiences in re-engineering the approach to editing and imputing Canadian imports data. *Proceedings of the Bureau of the Census Annual Research Conference and Technology Interchange*, Washington, DC, pp. 1025–1037.

Lahiri, D.B. (1951). A method of sample selection providing unbiased ratio estimates. *Bulletin International Statistical Institute* **33**, 133–140.

Lahiri, P. (2003). On the impact of bootstrap in survey sampling and small-area estimation. *Statistical Science* **18**, 199–210.

Lahiri, P.A., Larsen, M.D. (2005). Regression analysis with linked data. *Journal of the American Statistical Association* **100**, 222–230.

Lambert, D. (1993). Measures of disclosure risk and harm. *Journal of Official Statistics* **9**, 313–331.

Lambert, D., Peterson, B., Terpenning, I. (1991). Nondetects, detection limits, and the probability of detection. *Journal of the American Statistical Association* **86**, 266–277.

Landers, D.H., Eilers, J.M., Brakke, D.F., Overton, W.S., Kellar, P.E., Silverstein, M.E., Schonbrod, R.D., Crowe, R.E., Linthurst, R.A., Omernik, J.M., Teague, S.A., Meier, E.P. (1987). Characteristics of lakes in the Western United States. Volume I: Population Descriptions and Physico-Chemical Relationships. EPA-600/3-86/054a. US Environmental Protection Agency, Washington, DC.

Langer, G., Cohen, J. (2005). Voters and values in the 2004 election. Special Issue: Polling Politics, Media, and Election Campaigns. *Public Opinion Quarterly* **69**(5), 744–759.

Laplace, P.S. (1783). Sur les naissances, les marriages et les morts à Paris, depuis 1771 jusqu'en 1784, et dans toute l'étendue de la France, pendant les années 1781 et 1782. *Mém. Acad. Roy. Sci. Paris* 693–702.

Laplace, P.S., (1814a). Essai philosophique sur les probabilités. Cutture et Civilization: Bruxelles. Mme Ve Courcier, Imprime ur-Libraire pour les Mathématiques, quai des Augustins, no. 57, Paris.

Laplace, P.S., (1814b). *Théorie analytique des probabilités* (2nd ed.). Courcier, Paris.

Lark, R.M. (2002). Optimized spatial sampling of soil for estimation of the variogram by maximum likelihood. Geoderma **105**, 49–80.

Larsen, M.D., Rubin, D.B. (2001). Alterative automated record linkage using mixture models. *Journal of the American Statistical Association* **79**, 32–41.

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.

Latouche, M., Berthelot, J.-M. (1992). Use of a score function to prioritize and limit recontacts in editing business surveys. *Journal of Official Statistics* **8**, 389–400.

Lau, R. (1994). An analysis of the accuracy of "trial heat" polls during the 1992 presidential election. *Public Opinion Quarterly* **58**, 2–20.

Lavallée, P. (1995). Cross-sectional weighting of longitudinal surveys of individuals and households using the weight share method. *Survey Methodology* **21**, 25–32.

Lavallée, P. (2007a). *Administrative Data Usage in the Framework of Social Statistics: Current and Future Picture*. Internal document from Statistics Canada, Ottawa.

Lavallée, P. (2007b). *Indirect Sampling*. Springer, New York.

Lavallée, P., Hidiroglou, M.A. (1988). On the stratification of skewed populations. *Survey Methodology* **14**, 33–43.

Lavrakas, P.J., (Ed.). (2007). Special issue: cell phone numbers and telephone surveys in the U.S. *Public Opinion Quarterly* **71**(5).

Lawrence, D., McDavitt, C. (1994). Significance editing in the Australian survey of average weekly earning. *Journal of Official Statistics* **10**, 437–447.

Lawrence, D., McKenzie, R. (2000). The general application of significance editing. *Journal of Official Statistics* **16**, 243–253.

Lazarsfeld, P., Fiske, M. (1938). The 'panel' as a new tool for measure opinion. *Public Opinion Quarterly* **2**, 596–612.

Lazarsfeld, P.F., Berelson, B., Gaudet, H. (1944). *The People's Choice: How the Voter Makes Up His Mind in a Presidential Campaign*. Columbia University Press, New York.

Le Guennec, J., Sautory, O. (2003). *La macro Calmar2*. Manuel d'utilisation, document interne, INSEE, Paris, France.

Lê, T.A., Verma, V.K. (1997). An analysis of sample designs and sampling errors of the Demographic and Health Surveys. *DHS Analytical Reports*, 284. Macro International Inc., Calverton, MD.

Lee, H. (1991). Model-based estimators that are robust to outliers. *Proceedings of the Annual Research Conference*, U.S. Bureau of the Census, Washington, DC, 178–202.

Lee, H. (1995). Outliers in business surveys. In: Cox, B.G., Binder, D.A., Chinnappa, B.N., Christianson, A., Colledge, M.J., Kott, P.S. (Eds.), *Business Survey Methods*, Chapter 26. John Wiley & Sons, Inc., New-York, pp. 503–526.

Lee, H., Patak, Z. (1998). Outlier robust generalized regression estimator. *Proceedings of the Survey Methods Section*, Statistical Society of Canada, Ottawa, pp. 231–235.

Lee, H, Rancourt, E., Särndal, C.-E. (2000). Variance estimation from survey data under single imputation. In: Groves, R.M., Dillman, D.A., Eltinge, J.L., Little, R.J.A. (Eds.), *Survey Nonresponse*. Wiley, New York, pp. 315–328.

Lee, H., Rancourt, E., Särndal, C.-E. (2002). Variance estimation from survey data under single imputation. In: Groves, R.M., Dillman, D.A., Eltinge, J.L., Little, R.J.A. (Eds.), *Survey Nonresponse.* John Wiley and Sons, New York, pp. 315–328.

Legg, J., Fuller, W.A., Nusser, S.M. (2005). Estimation for longitudinal surveys with repeated panels of observations. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, Alexandria, VA. [CD-ROM].

Legg, J., Fuller, W.A., Nusser, S.M. (2006). Estimation for two-phase panel surveys. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, Alexandria, VA. [CD-ROM].

Legg, J.C. (2006). Estimation for two-phase longitudinal surveys with application to the National Resources Inventory. Unpublished Ph.D. dissertation, Iowa State University, Ames, Iowa.

Lemaitre, G.E., Dufour, J. (1987). An integrated method for weighting persons and families. *Survey Methodology* **13**, 199–207.

Lent, J., Miller, S., Cantwell, P., Duff, M. (1999). Effects of composite weights on some estimates from the current population survey. *Journal of Official Statistics* **15**(3), 431–448.

Lepkowski, J.M., Sadosky, S.A., Weiss, P. (1998). Mode, behavior, and data recording error. In: Couper, M., Baker, R., Bethlehem, J., Clark, C., Martin, J., Nicholls, W. II, O'Reilly, J. (Eds.), *Computer Assisted Survey Information Collection*. John Wiley, New York, pp. 367–388.

Lepkowski, J.M. (1988). Telephone sampling methods in the United States. In: Groves, R.M., Biemer, P.P., Lyberg, L.E., Massey, J.T., Nicholls, W.L., Waksberg, J. (Eds.), *Telephone Survey Methodology*. John Wiley & Sons, New York, pp. 73–98.

Lepkowski, J.M. (1989). Treatment of wave nonresponse in panel surveys. In: Kasprzyk, D., Duncan, G., Kalton, G., Singh, M.P. (Eds.), *Panel Surveys*. John Wiley & Sons, New York, pp. 348–374.

Lepkowski, J.M., Groves, R.M. (1986). A mean squared error model for multiple frame, mixed mode survey design. *Journal of the American Statistical Association* **81**, 930–937.

Lessler, J.T., Kalsbeek, W.D. (1992). *Nonsampling error in surveys.* J.W. Wiley and Sons, Inc., New York.

Lessof, C. (2009). Ethical issues in longitudinal surveys. In: Lynn, P. (Ed.), *Methodology of Longitudinal Surveys*. John Wiley & Sons, Chichester, UK, pp. 35–54.

Leuthold, D.A., Scheele, R. (1971). Patterns of bias in samples based on telephone directories. *Public Opinion Quarterly* **42**, 104–114.

Levitt, J., Weiser, W.J., Muñoz, A. (2005). *Making the List: Database Matching and Verification Procedures for Voter Registration*. Brennan Center for Justice, New York University School of Law, New York.

Levy, P. (1977). Optimal allocation in stratified random network sampling for estimating the prevalence of attributes in rare populations. *Journal of the American Statistical Association* **72**, 758–763.

Lewis, D. (2007). Winsorization for estimates of change. *Proceedings of the Third International Conference on Establishment Surveys*, June 2007, American Statistical Association, Montréal, 1165–1172.

Li, N. (2006). *Interactive Voice Response.* Report. Public Policy Polling, University of North Carolina at Chapel Hill, December. Available at: http://www.publicpolicypolling.com/pdf/reports/IVR_Nebulahi.pdf. Last accessed March 10, 2009.

Lie, E. (2002). The rise and fall of sampling surveys in Norway, 1875–1906. *Science in Context* **15**, 385–409.

Lin, I., Schaeffer, N. (1995). Using survey participants to estimate the impact of nonparticipation. *Public Opinion Quarterly* **59**, 236–258.

Lindell, K. (1994). Evaluation of the editing process of the salary statistics for employees in country councils. Paper presented at the UN congress on data editing in Cork, Ireland. In: *Statistical Data Editing*, vol. 2. UN Statistical Commission and Economic Commission for Europe, Geneva, Switzerland, pp. 2–7.

Lindeman, M. (2002). Opinion quality and policy preferences in deliberative research. In: Delli Carpini, M.X., Huddy, L., Shapiro, R.Y. (Eds.), *Research in Micropolitics. Volume 6. Political Decision-Making, Deliberation and Participation.* Elsevier, New York. pp. 195–221.

Lindeman, M. (2006). Too many Bush voters? False recall and the 2004 exit poll. Paper Presented at the Annual Meeting of the Midwest Political Science Association, Chicago, Illinois, 20–23.

Lindeman, M. (2007). Condemned to repetition: The 2006 exit poll controversy. Public Opinion Pros. Available at: http://www.publicopinionpros.com/features/2007/jan/lindeman.asp, January. Accessed April 1, 2008.

Link, M.W., Mokdad, A.H., Kulp, D., Hyon, A. (2006). Has the National Do Not Call Registry helped or hurt state-level response rates? A time series analysis. *Public Opinion Quarterly* **70**(5), 794–809.

Link, M.W., Oldendick, R.W. (1999). Call screening: Is it really a problem for survey research? *Public Opinion Quarterly* **63**, 577–589.

Link, W.A., Barker, R.J. (2005). Modeling association among demographic parameters in analysis of open population capture-recapture data. *Biometrics* **61**, 46–54.

Linthurst, R.A., Landers, D.H., Eilers, J.M., Brakke, D.F., Overton, W.S., Meier, E.P., Crowe, R.E. (1986). Characteristics of lakes in the Eastern United States. Volume I: Population Descriptions and Physico-Chemical Relationships. EPA-600/4-86/007a. US Environmental Protection Agency, Washington, DC.

Little, R.J.A (1986). Survey nonresponse adjustments for estimates of means. *International Statistical Review* **54**, 139–157.

Little, R.J.A. (1988). Missing-data adjustments in large surveys. *Journal of Business and Economic Statistics* **6**, 287–296.

Little, R.J.A., An, H. (2004). Robust likelihood-based analysis for multivariate data with missing values. *Statistica Sinica* **14**, 949–968.

Little, R.J.A., Rubin, D. (2002). *Statistical Analysis with Missing Data*, 2nd ed. Wiley, New York.

Little, R.J.A., Smith, P.J. (1987). Editing and imputation for quantitative survey data. *Journal of the American Statistical Association* **82**, 58–68.

Little, R.J.A., Vartivarian, S. (2003). On weighting the rates in non-response weights. *Statistics in Medicine* **22**, 1589–1599.

Little, R.J.A., Wu, M.M. (1991). Models for contingency tables with known margins when target and sampled populations differ. *Journal of the American Statistical Association* **86**, 87–95.

Liu, B., Ferraro, D., Wilson, E., Brick, J.M. (2004). Trimming extreme weights in household surveys. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, Alexandria, Virginia.

Liu, K.J. (1999). Interval estimation of simple differences under independent negative binomial sampling. *Biometrical Journal* **41**, 83–92.

Lo, N.C.H., Griffith, D., Hunter, J.R. (1997). Using a restricted adaptive cluster sampling to estimate Pacific Hake larval abundance. *CalCOFI Report* **38**, 103–113.

Loeve, A. (2001). *Notes on Sensitivity Measures and Protection Levels*. Research paper 0129, Statistics Netherlands, Voorburg, The Netherlands.

Lohr, S. (1999). *Sampling: Design and Analysis.* Duxbury Press, Brooks/Cole Publishing, Pacific Grove, CA.

Lohr, S., Rao, J.N.K. (2000). Inference in dual frame surveys. *Journal of the American Statistical Association* **95**, 271–280.

Lohr, S., Rao, J.N.K. (2006). Estimation in multiple-frame surveys. *Journal of the American Statistical Association* **101**, 1019–1030.

Lohr, S. (1999). *Sampling: Design and Analysis*. Duxbury Press, Pacific Grove, CA.

Loukas, S., Kemp, C.D. (1983). On computer sampling from trivariate and multivariate discrete distribution. *Journal of Statistical Computation and Simulation* **17**, 113–123.

Lu, W.W., Brick, J.M., Sitter, R.R. (2006). Algorithms for constructing combined strata variance estimators. *Journal of the American Statistical Association* **101**, 1680–1692.

Lumley, T. (2004). *Analysis of Complex Survey Samples*. Department of Biostatistics, University of Washington, Seattle, WA.

Lundström, S., Särndal, C.-E. (1999). Calibration as a standard method for treatment of nonresponse. *Journal of Official Statistics* **15**, 305–327.

Luskin, R.C., Fishkin, J.S., Jowell, R. (2002). Considered opinions: Deliberative polling in Britain. *British Journal of Political Science* **32**, 455–487.

Lyberg, L., Biemer, P., Collins, M., De Leeuw, E., Dippo, C., Schwartz, N., Trewin, D. (1997). *Survey Measurement and Process Quality*. John Wiley & Sons, New York.

Lynn, P. (Ed.). (2009). *Methodology of Longitudinal Surveys*. John Wiley & Sons, Chichester, UK.

Lynn, P., Sala, E. (2004). The contact and response process in business surveys: Lessons from a multimode survey of employers in the UK. Working Paper No. 10, ESRC Research Methods Programme.

Lynn, P., Sala, E. (2006). Measuring change in employment characteristics: the effects of dependent interviewing. *International Journal of Public Opinion Research* **18**, 500–509.

Maciejewski, P.K., Prigerson, H.G., Mazure, C.M. (2000). Self-efficacy as a mediator between stressful life events and depressive symptoms. *British Journal of Psychiatry* **176**, 373–378.

Mack, S., Huggins, V., Keathley, D., Sundukchi, M. (1999). Do monetary incentives improve response rates in the Survey of Income and Program Participation? *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 529–534.

Macro International. (1996). *Sampling Manual, Demographic and Health Surveys, Phase III.* Macro International Inc., Calverton, MD.

Madow, L.H., Madow, W.G. (1944). On the theory of systematic sampling. *The Annals of Mathematical Statistics* **15**, 1–24.

Madow, W.G. (1949). On the theory of systematic sampling, II. *The Annals of Mathematical Statistics* **20**, 333–354.

Madsen, L., Ruppert, D., Altman, N.S. (2008). Regression with spatially misaligned data. *Environmetrics* **19**, 453–467.

Mahalanobis, P.C. (1946). Recent experiments in statistical sampling in the Indian Statistical Institute. *Journal of the Royal Statistical Society* **109**, 325–370.

Mahalanobis, P.C. (1965). Statistics as a key technology. *The American Statistician* **19**, 43–46.

Malec, D. (2005). Small area estimation from the American Community Survey using a hierarchical logistic model of persons and housing units. *Journal of Official Statistics* **21**, 411–432. Available at: http://www.jos.nu/Articles/abstract.asp?article=213411.

Mantel, H., Nadon, S., Yeo, D. (2000). Effect of nonresponse adjustments on variance estimates for the National Population Health Survey. *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 221–226.

Manzari, A. (2004). Combining editing and imputation methods: an experimental application on population census data. *Journal of the Royal Statistical Society A* **167**, 295–307.

Mark, D.M. (1990). Neighbor-based properties of some orderings of two-dimensional space. *Geographical Analysis* **2**, 145–157.

Marker, D.A. (2001). Producing small area estimates from national surveys: methods for minimizing use of indirect estimators. *Survey Methodology* **27**, 183–188.

Marquis, K. (1978). Inferring health interview response bias from imperfect record checks. *Proceedings of the ASA Section on Survey Research Methods*, 265–270.

Marsh, C., Dale, A., Skinner, C.J. (1994). Safe data versus safe setting: access to microdata from the British Census. *International Statistical Review* **62**, 35–53.

Martin, E., Traugott, M., Kennedy, C. (2005). A review and proposal for a new measure of poll accuracy. *Public Opinion Quarterly* **69**(3), 342–369.

Martin, J., Bynner, J., Kalton, G., Boyle, P., Goldstein, H., Gayle, V., Parsons, S., Piesse, A. (2006). *Strategic Review of Panel and Cohort Studies, with Appendices.* Report to the U.K. Research Resources Board of the Economic and Social Research Council. Available at: http://www.longviewuk.com/pages/publications.shtml.

Massell, P., Zayatz, L., Funk, J. (2006). Protecting the confidentiality of survey tabular data by adding noise to the underlying microdata: application to the Commodity Flow Survey. In: Domingo-Ferrer, J., Franconi, L. (Eds.), *Privacy in Statistical Databases. Lecture Notes in Computer Science*, 4302. Springer, Berlin, pp. 304–317.

Massey, J.T. (1995). Estimating the response rate in a telephone survey with screening. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, Alexandria, VA, 673–677.

Matei, A., Tillé, Y. (2005). Evaluation of variance approximations and estimators in maximum entropy sampling with unequal probability and fixed sample size. *Journal of Official Statistics* **21**(4), 543–570.

Matei, A., Tillé, Y. (2007). Computational aspects of order $\pi ps$ sampling schemes. *Computational Statistics & Data Analysis* **51**, 3703–3717.

Matérn, B. (1986). *Spatial Variation*, 2nd ed. Lecture Notes in Statistics #36, Brillinger, D., Fienberg, S., Gani, J., Hartigan, J., Krickeberg, K. (Eds.), Springer-Verlag, Berlin, Germany.

Mathiowetz, N.A., McGonagle, K.A. (2000). An assessment of the current state of dependent interviewing in household surveys. *Journal of Official Statistics* **16**, 401–418.

Mathy, R.M., Schillace, M., Coleman, S.M., Ber Quist, B.E. (2002). Methodological rigor with Internet samples: new ways to reach underrepresented populations. *Cyberpsychology Behavior* **5**(3), 253–266.

Matthews, S., Bérard, H. (2002). The outlier detection and treatment strategy for the Monthly Wholesale and Retail Trade Survey of Statistics Canada. *Proceedings of the Survey Methods Section*, Statistical Society of Canada.

Mazur, C., Cotter, J.J. (1991). Automating the development of area sampling frames using digital data displayed on a graphics workstation. *Proceedings of Statistics Canada Symposium 91. Spatial Issue in Statistics.* Ottawa, ON, Canada, 55–68.

McBratney, A.B., Webster, R., Burgess, T.M. (1981). The design of optimal sampling schemes for local estimation and mapping of regionalized variables: I. Theory and method. *Computational Geosciences* **7**, 331–334.

McDonald, M.P. (2003a). On the over-report Bias of the National Election study. *Political Analysis* **11**(2), 180–186.

McDonald, T. (2003b). Review of environmental monitoring methods: survey design. *Environmental Monitoring & Assessment* **85**, 277–292.

McIntyre, G.A. (1952). A method for unbiased selective sampling, using ranked sets. *Journal of Agricultural Research* **3**, 385–390.

McKenzie, R., Gross, B. (2001). Synchronised sampling. *Proceedings of ICES II The Second International Conference on Establishment Surveys.* 237–244.

McLaren, C., Steel, D. (2001). Rotation patterns and trend estimation for repeated surveys using rotation group estimates. *Statistica Neerlandica* **55**, 221–238.

McMillen, D.B., Herriot, R. (1985). Toward a longitudinal definition of households. *Journal of Economic and Social Measurement* **13**, 349–360.

Mecatti, F. (2007). A single frame multiplicity estimator for multiple frame surveys. *Survey Methodology* **33**, 151–157.

Mendelberg, T. (2002). The deliberative citizen: theory and evidence. In: Delli Carpini, M.X., Huddy, L., Shapiro, R.Y. (Eds.), *Research in Micropolitics. Volume 6. Political Decision-Making, Deliberation and Participation.* Elsevier, New York, pp. 151–193.

Meng, X.-L., Rubin, D.B. (1993). Maximum likelihood via the ECM algorithm: a general framework. *Biometrika* **80**, 267–278.

Meng, X.-L., Romero, M. (2003). Discussion: efficiency and self-efficiency with multiple imputation inference. *International Statistical Review* **71**, 607–618.

Merkouris, T. (2004). Combining independent regression estimators from multiple surveys. *Journal of the American Statistical Association* **99**, 1131–1139.

Merkouris, T. (2006). Efficient small-domain estimation by combining information from multiple surveys through regression. *Proceedings of the Survey Research Methods Section*, *Joint Statistical Meetings*, pp. 3419–3425, American Statistical Association.

Merton, R.K., Hatt, P.K. (1949). Election polling forecasts and public images of social science. *Public Opinion Quarterly* **13**, 185–222.

Messer, J.J., Ariss, C.W., Baker, J.R., Drousé, S.K., Eshelman, K.N., Kaufmann, P.R., Linthurst, R.A., Omernik, J.M., Overton, W.S., Sale, M.J., Schonbrod, R.D., Stambaugh, S.M., Tuschall, J.R. Jr. (1986). National Stream Survey Phase I - Pilot Survey. EPA-600/4-86/026. US Environmental Protection Agency, Washington, DC.

Messer, J.J., Linthurst, R.A., Overton, W.S. (1991). An EPA program for monitoring ecological status and trends. *Environmental Monitoring Assessment* **17**, 67–78.

Mickiewicz, E. (1981). *Media and the Russian Public.* Praeger, New York.

Mitofsky, W.J. (1970). Sampling of Telephone Households. Internal CBS News Memo.

Mitofsky, W.J. (1981). The 1980 pre-election polls: a review of disparate methods and results. *Proceedings of the American Statistical Association Survey Research Methods Section*, 47–52.

Mitofsky, W.J. (1998). Was 1996 a worse year for polls than 1948? *Public Opinion Quarterly* **62**, 230–249.

Mitofsky, W.J. (1999). A short history of exit polls. In: Lavrakas, P.J., Holley, J.K. (Eds.), *Polling and Presidential Election Coverage*. Sage Publications, Newbury Park, California, pp. 83–99.

Mitra, R., Reiter, J.P. (2006). Adjusting survey weights when altering identifying design variables via synthetic data. In: Domingo-Ferrer, J., Franconi, L. (Eds.), *Privacy in Statistical Databases*. *Lecture Notes in Computer Science*, 4302. Springer, Berlin, pp. 177–188.

Moberg, P.E. (1982). Biases in unlisted phone numbers. *Journal of Advertising Research* **22**(4), 51–55.

Mode, N.A., Conquest, L.L., Marker, D.A. (1999). Ranked set sampling for ecological research: accounting for the total costs of sampling. *Environmetrics* **10**, 179–194.

Mode, N.A., Conquest, L.L., Marker, D.A. (2002). Incorporating prior knowledge in environmental sampling: ranked set sampling and other double sampling procedures. *Environmetrics* **13**, 513–521.

Mohadjer, L. (1988). Stratification of prefix areas for sampling rare populations. In: Groves, R.M., Biemer, P.-P., Lyberg, L.E., Massey, J.T., Nicholls, W.L., Waksberg, J. (Eds.), *Telephone Survey Methodology*. John Wiley & Sons, New York, pp. 161–173.

Mohadjer, L., Curtin, L.R. (2008). Balancing sample design goals for the National Health and Nutrition Examination Survey. *Survey Methodology* **34**, 119–126.

Montaquila, J., Brick, J.M., Hagedorn, M., Kennedy, C., Keeter, S. (2008). Aspects of nonresponse bias in RDD telephone surveys. In: Lepkowski, J.M., Tucker, C., Brick, J.M. et al. (Eds.), *Advances in Telephone Survey Methodology.* Wiley, New York.

Montgomery, D. (1971). Consumer characteristics associated with dealing: an empirical example. *Journal of Marketing Research* **8**, 118–120.

Moore, D.W. (1992). *The Superpollsters: How They Measure and Manipulate Public Opinion in America.* Four Walls Eight Windows, New York.

Moore, J.C. (1988). Self/proxy response status and survey response quality. *Journal of Official Statistics* **4**(2), 155–122.

Moore, J.C., Kasprzyk, D. (1984). Month-to-month recipiency turnover in ISDP. *Proceedings of the Section on Survey Research Methods,* American Statistical Association, pp. 726–731.

Morgan, D.L. (1996). Focus groups. *Annual Review of Sociology* **22**, 129–152.

Mosteller, F., Hyman, H., McCarthy, P.J., Marks, E.S., Truman, D.B. (1949). *The Pre-Election Polls of 1948.* Social Science Research Council, New York.

Mote, V.L., Anderson, R.L. (1965). An investigation of the effect of misclassification on the properties of chi-square tests in the analysis of categorical data. *Biometrika* **52**(1 and 2), 95–109.

Mueller, J. (1973). *War, Presidents and Public Opinion.* Wiley, New York.

Mugglin, A.S., Carlin, B.P. (1998). Hierarchical modeling in geographic information systems: population interpolation over incompatible zones. *Journal of Agricultural, Biological, and Environmental Statistics* **3**(2), 111–130.

Müller, W.G., Zimmerman, D.L. (1999). Optimal design for variogram estimation. *Environmetrics* **10**, 23–38.

Mulry, M.H. (2007). Summary of accuracy and coverage evaluation for the U.S. Census 2000. *Journal of Official Statistics* **23**, 345–370. Available at: http://www.jos.nu/Articles/abstract.asp?article=233345.

Mulry, M.H., Feldpausch, R.M. (2007). Investigation of treatment of influential values. *Proceedings of the Third International Conference on Establishment Surveys*, June 2007, American Statistical Association, Montréal, 1173–1179.

Mulry, M.H., Kostanich, D.K. (2006). Framework for census coverage error components. *Proceedings of the Survey Research Methods Section*, American Statistical Association, Seattle, Washington, pp. 3461–3468. Available at: http://www.amstat.org/sections/srms/Proceedings/y2006/Files/JSM2006-000727.pdf.

Mulry, M.H., Spencer, B.D. (1991). Total error in PES estimates of population. *Journal of the American Statistical Association* **86**, 839–855.

Mulry, M.H., Spencer, B.D. (1993). Accuracy of the 1990 Census and undercount adjustments. *Journal of the American Statistical Association*, **88**, 1080–1091.

Munholland, P.L., Borkowski, J.J. (1996). Simple Latin Square sampling +1: a spatial design using Quadrats. *Biometrics* **52**, 125–136.

Murray, S.K. (2006). Private polls and presidential policymaking. *Public Opinion Quarterly* **70**(4), 477–498.

Muthen, L.K., Muthen, B.O. (1998–2005). *Mplus*. Muthen & Muthen, Los Angeles, CA.

Myers, R.A., Pepin, P. (1990). The comparison of log-normal based estimators of abundance. *Biometrics* **46**, 1185–1192.

Nacos, B.L., Shapiro, R.Y., Isernia, P. (Eds.). (2000). *Decisionmaking in a Glass House: Mass Media, Pubic Opinion, and American and European Foreign Policy in the 21st Century*. Rowman & Littlefield, New York.

Nakanishi, M. (1973). Advertising and promotion effects on consumer response to new products. *Journal of Marketing Research* **10**, 242–249.

Narain, R.D. (1951). On sampling without replacement with varying probabilities. *Journal of the Indian Society of Agricultural Statistics* **3**, 169–174.

Nathan, G. (1976). An empirical study of response and sampling errors for multiplicity estimates with different counting rules. *Journal of the American Statistical Association* **71**, 808–815.

Nathan, G. (2001). Telesurvey methodologies for household surveys—A review and some thoughts on the future. *Survey Methodology* **27**, 7–37.

National Research Council. (2005). *Expanding Access to Research Data: Reconciling Risks and Opportunities*. Panel on Data Access for Research Purposes, Committee on National Statistics. The National Academies Press, Washington, DC.

National Research Council. (2007). In: Citro, C.F., Kalton, G. (Eds.), *Using the American Community Survey: Benefits and Challenges*. Panel on the Functionality and Usability of Data from the American Community Survey, Committee on National Statistics, Division of Behavioral and Social Sciences and Education, The National Academies Press, Washington, DC. Available at: http://books.nap.edu/catalog.php?record_id=11901#toc.

National Science Foundation. (2003). *SESTAT: Design and Methodology*. Available at: http://srsstats.sbe.nsf.gov/docs/techinfo.html.

National Statistics. (2004). *Code of Practice: Protocol on Data Access and Confidentiality*. Her Majesty's Stationary Office, Norwich, UK.

Natural Resources Conservation Service. (2007a). 2003 National Resources Inventory Land Use Report. Available at: http://www.nrcs.usda.gov/technical/NRI/2003/Landuse-mrb.pdf.

Natural Resources Conservation Service. (2007b). 2003 National Resources Inventory Soil Erosion Report. Available at: http://www.nrcs.usda.gov/technical/NRI/2003/SoilErosion-mrb.pdf.

Natural Resources Conservation Service. (2007c). 2003 National Resources Inventory Wetland Tables. Available at: http://www.nrcs.usda.gov/technical/NRI/2003/table4.html.

Navarro, G. (2001). A guided tour of approximate string matching. *Association of Computing Machinery Computing Surveys* **33**, 31–88.

NCPP (2004). The 2004 election polls. Available at: http://ncpp.org/drupal57/files/2004%20Election%20Polls%20Review.pdf. Last accessed March 2, 2009.

NCPP (2008). Table of national election poll results. Available at: http://ncpp.org/files/08FNLncppNatlPolls_010809.pdf. Last accessed March 2, 2009.

NCS (2007). The National Children's Study. Available at: http://www.nationalchildrensstudy.gov/.

Nedyalkova, D., Tillé, Y. (2008). Optimal sampling and estimation strategies under the linear model, *Biometrika* **95**(3), 521–537.

Neter, J., Waksberg, J. (1964a). A study of response errors in expenditures data from household interviews. *Journal of the American Statistical Association* **59**, 18–55.

Neter, J., Waksberg, J. (1964b). Conditioning effects from repeated interviews. *Journal of Marketing* **28**, 51–56.

Newcombe, H.B., Kennedy, J.M. (1962). Record linkage: making maximum use of the discriminating power of identifying information. *Communications of the Association for Computing Machinery* **5**, 563–567.

Newcombe, H.B., Kennedy, J.M., Axford, S.J., James, A.P. (1959). Automatic linkage of vital records. *Science* **130**, 954–959.

Newell, C.E., Rosenfeld, P., Harris, R.N., Hindelang, R.L. (2004). Reasons for nonresponse on U.S. Navy surveys: a closer look. *Military Psychology* **16**(4), 265–276.

Neyman, J. (1934). On the two different aspects of representative method: the method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society* **97**, 558–606.

Neyman, J. (1938). Contribution to the theory of sampling human populations. *Journal of the American Statistical Association* **33**, 101–116.

Nichols, J.A.F., Roslow, S., Comer, L.B. (1995). Demographic, situational and shopping comparisons of Hispanic and Non-Hispanic mall patterns. *Hispanic Journal of Behavioral Sciences* **17**(4), 537–547.

Nichols, W.L. II., Groves, R.M. (1986). The status of computer-assisted telephone interviewing: Part I – Introduction and impact on cost and timelines of survey data. *Journal of Official Statistics* **2**, 93–115.

Nielsen, S.F. (2003). Proper and improper multiple imputation. *International Statistical Review* **7**, 593–607.

Nieuwenbroek, N., Boonstra, H. (2002). Bascula 4.0 for weighting sample survey data with estimation of variances. *The Survey Statistician*, Software Reviews, July 2002.

Nieuwenbroek, N., Boonstra, J. (2001). *BASCULA 4.0 Reference Manual*. Division Technology and Facilities, Department of Methods and Informatics, Statistics Netherlands, Boonstra.

Nigam, K., McCallum, A.K. Thrun, S., Mitchell, T. (2000). Text classification from labeled and unlabelled documents using EM. *Machine Learning* **39**, 103–134.

Nirel, R., Glickman, H., Ben Hur, D. (2003). A strategy for a system of coverage samples for an Integrated Census. *Proceedings of Statistics Canada Symposium 2003 Challenges in Survey Taking for the Next Decade. Statistics Canada International Symposium Series – Proceedings*, Catalogue no. 11-522-XIE. Available at: http://www.statcan.ca/english/freepub/11-522-XIE/2003001/session18/nirel.pdf.

Noelle-Neumann, E. (1993 [1984]). *The Spiral of Silence: Public Opinion—Our Social Skin*, 2nd ed. University of Chicago Press, Chicago, IL.

Nordbotten, S. (1995). Editing statistical records by neural networks. *Journal of Official Statistics* **11**, 391–411.

Nowell, C., Stanely, L.R. (1991). Length-biased sampling in mall intercept surveys. *Journal of Marketing Research* **28**, 475–479.

Nunnally, J., Bernstein, I. (1994). *Psychometric Theory*, 3rd ed. McGraw Hill, New York.

Nusser, S.M. (2004). Computer-assisted data collection for geographic features. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, [CD-ROM]. Alexandria, VA.

Nusser, S.M. (2005). Digital capture of geographic feature data for surveys. *Proceedings of the 2005 Federal Committee on Statistical Methodology Research Conference*. Available at: http://www.fcsm.gov/05papers/Nusser_IXA.pdf.

Nusser, S.M., Goebel, J.J. (1997). The National Resources Inventory: a long-term multi-resource monitoring programme. *Environmental and Ecological Statistics* **4**, 181–204.

Nusser, S.M., Thompson D.M. (1997). Networked hand-held CASIC. *Proceedings of Symposium 97: New Directions in Surveys and Censuses*, Statistics Canada, 361–364.

Office for National Statistics. (2000). *One Number Census Methodology.* ONS (ONC(SC)) 00/01. Available at: http://www.statistics.gov.uk/census2001/pdfs/sc0001.pdf.

Office for National Statistics. (2002). *Changes to the ONC Imputation System*. One Number Census Steering Committee Paper 02/01. Available at: http://www.statistics.gov.uk/census2001/pdfs/sc0201.pdf.

Office for National Statistics. (2003). *Alternatives to a Census: Rolling Census*. ONS Census Strategic Development Programme Information Paper. Available at: http://www.statistics.gov.uk/downloads/theme_population/rolling_census.pdf.

Office for National Statistics. (2004). Consultation Paper 29: Proposals for a Continuous Population Survey. Office for National Statistics, London, UK.

Ogus, J.L., Clark, D.F. (1971). *The annual survey of manufactures: a report on methodology*. Technical paper no. 24, Census Bureau, Washington, DC.

Oh, H.L., Scheuren, F.J. (1983). Weighting adjustments for unit nonresponse. In: Madow, Olkin, Rubin (Eds.), *Incomplete Data in Sample Surveys*, vol. 2. Academic Press, New York.

Ohlsson, E. (1990a). *Sequential Sampling from a Business Register and its Application to the Swedish Consumer Price Index.* Research and Development Report 199:06, Statistics Sweden, Stockholm.

Ohlsson, E. (1990b). *SAMU— The System for Co-ordination of Sample from the Business Register at Statistics Sweden— A methodological description*. Research and Development Report 199:18, Statistics Sweden, Stockholm.

Ohlsson, E. (1995). Coordination of samples using permanent random numbers. In: Cox, B.G., Binder, D.A., Chinnappa, B.N., Christianson, A., Colledge, M.J., Kott, P.S. (Eds.), *Business Survey Methods*. John Wiley & Sons, New York, pp. 153–169.

Ohlsson, E. (1998). Sequential Poisson sampling. *Journal of Official Statistics* **14**, 149–162.

Olea, R.A. (1984). Sampling design optimization for spatial functions. *Mathematical Geology* **16**, 369–392.

Olsen, A.R., Sedransk, J., Edwards, E., Gotway, C.A., Liggett, W. (1999). Statistical issues for monitoring ecological and natural resources in the United States. E*nvironmental Monitoring and Assessment* **54**, 1–45.

Olsen, R.J. (1980). A least squares correction for selectivity bias. *Econometrica* 1815–1820.

Olson, R.J. (2005). The problem of respondent attrition: survey methodology is key. *Monthly Labor Review* **128**, 63–70.

O'Muircheartaigh, C. (1991). Simple response variance: estimation and determinants. In: Biemer P.P. et al. (Eds.), *Measurement Errors in Surveys.* John Wiley & Sons, New York, pp. 551–574.

Openshaw, S., Taylor, P. (1979). A million or so correlation coefficients. In: Wrigley, N. (Ed.), *Statistical Methods in the Spatial Sciences*. Pion, London, pp. 127–144.

O'Roourke, D., Blair, J. (1983). Improving random respondent selection in telephone surveys. *Journal of Marketing Research* **20**, 428–432.

Orwin, R.G., Boruch, R.F. (1982). RRT meets RDD: statistical strategies for assuring response privacy in telephone surveys. *Public Opinion Quarterly* **46**, 560–571.

Otis, D.L., Burnham, K.P., White, G.C., Anderson, D.R. (1978). *Statistical Inference from Capture Data on Closed Animal Populations.* Wildlife Monograph #62. The Wildlife Society, Washington, DC.

Overton, W.S., Stehman, S.V. (1993). Properties of designs for sampling continuous spatial resources from a triangular grid. *Communications in Statistics, Part A – Theory and Methods* **22**, 2641–2660.

Overton, W.S., Stehman, S.V. (1996). Desirable design considerations for long-term monitoring of ecological variables. *Environmental and Ecological Statistics* **3**, 349–361.

Paass, G. (1988). Disclosure risk and disclosure avoidance for microdata. *Journal of Business and Economic Statistics* **6**, 487–500.

Page, B.I., Shapiro, R.Y. (1992). *The Rational Public: Fifty Years of Trends in Americans' Policy Preferences*. University of Chicago Press, Chicago, IL.

Page, C. (2006). *The Role of Public Opinion Research in Canadian Government*. University of Toronto Press, Buffalo, New York.

Palmieri Lage, J., Assunção, R.M., Reis, E.A. (2001). A minimal spanning tree algorithm applied to spatial cluster analysis. In: *Electronic Notes in Discrete Mathematics* **7**, 162–165. Elsevier, Fortaleza, Brazil.

Panagakis, N. (1989). Incumbent races: closer than they appear. *The Polling Report*. February 27, 1989. Available at: http://www.pollingreport.com/incumbent.htm. Last accessed March 8, 2009.

Panagopoulos, C. (2007). Follow the bouncing ball: assessing convention bumps 1964–2004. In: Panagopoulos, C. (Ed.), *Rewiring Politics: Presidential Nominating Conventions in the Media Age*. Louisiana State University Press, Baton Rouge, LA, pp. 16–28.

'Panel bias reviewed; results inconclusive' (1976). The Sampler from *Response Analysis* **7**, 2.

Pannekoek, J., De Waal, T. (2005). Automatic edit and imputation for business surveys: the Dutch contribution to the EUREDIT project. *Journal of Official Statistics* **21**, 257–286.

Pannekoek, J., Shlomo, N., De Waal, T. (2008). *Calibrated Imputation of Numerical Data under Linear Edit Restrictions*. UN/ECE Work Session on Statistical Data Editing, Vienna, Austria.

Park, I., Lee, H. (2004). Design effects for the weighted mean and total estimators under complex survey sampling. *Survey Methodology* **30**, 183–193.

Pascale, J., Mayer, T.S. (2004). Exploring confidentiality issues related to dependent interviewing: preliminary findings. *Journal of Official Statistics* **20**, 357–377.

Patil, G.P. (2002a). Composite sampling. In: El-Shaarawi, A.H., Piergorsch, W.W. (Eds.), *Encyclopedia of Environmetrics*. John Wiley & Sons, New York, pp. 387–391.

Patil, G.P. (2002b). Ranked set sampling. In: El-Shaarawi, A.H., Piergorsch, W.W. (Eds.), *Encyclopedia of Environmetrics*. John Wiley & Sons, New York, pp. 1684–1690.

Patil, G.P., Sinha, A.K., Taillie, C. (1994). Ranked set sampling. In: Patil, G.P., Rao, C.R. (Eds.), *Handbook of Statistics*. New York: North-Holland, pp. 167–200.

Patterson, B., Dayton, C.M., Graubard, B. (2002). Latent class analysis of complex survey data: application to dietary data. *Journal of the American Statistical Association* **97**, 721–729.

Paulsen, S.G., Hughes, R.M., Larsen, D.P. (1998). Critical elements in describing and understanding our nation's aquatic resources. *Journal of The American Water Resources Association* **34**, 995–1005.

Peano, G. (1890). Sur Une Courbe, Qui Remplit Toute Une Aire Plane. *Mathematische Annalen* **36**, 157–160.

Peterson, 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.

Peterson, S.A., Urquhart, N.S., Welch, E.B. (1999). Sample representativeness: a must for reliable regional lake condition estimates. *Environmental Science and Technology* **33**(10), 1559–1565.

Pettersson, H. (2005). Design of master sampling frames and master samples for household surveys in developing countries. Chapter 5 of United Nations (2005b).

Pew Research Center for the People and the Press. (2006). The Cell Phone Challenge to Survey Research. May 15. Available at: http://www.people-press.org/reports/display.php3?ReportID=276. Accessed April 10, 2007.

Pfeffermann, D. (1993). The role of sampling weights when modeling survey data. *International Statistical Review/Revue Internationale de Statistique* **61**(2), 317–337.

Pfeffermann, D., Sverchkov, M. (1999). Parametric and semi-parametric estimation of regression models fitted to survey data. *Sankhya, Series B* **61**, 166–186.

Piazza, T. (1993). Meeting the challenge of answering machines. *Public Opinion Quarterly* **57**, 219–231.

Pierzchala, M. (1990). A review of the state of the art in automated data editing and imputation. *Journal of Official Statistics* **6**, 355–377.

Piesse, A., Judkins, D., Kalton, G. (2009). Measuring causal effects in longitudinal surveys. In: Lynn, P. (Ed.), *Methodology of Longitudinal Surveys*. John Wiley & Sons, Chichester, UK, 303–316.

Plasser, F., Plasser, G. (2002). *Global Political Campaigning*. Praeger Press, Westport, CT.

Platek, R., Särndal, C.-E. (2001). Can a statistician deliver? *Journal of Official Statistics* **17**, 1–20.

Platek, R., Singh, M.P., Tremblay, V. (1978). Adjustments for nonresponse in surveys. In: Namboordiri, N.K. (Ed.), *Survey Sampling and Measurement*. Academic Press, New York.

Politz, A., Simmons, W. (1949). An attempt to get "not at homes" into the sample without callbacks. *Journal of the American Statistical Association* **44**, 9–31.

Pollock, J., Zamora, A. (1984). Automatic spelling correction in scientific and scholarly text. *Communications of the ACM* **27**, 358–368.

Poterba, J., Summers, L. (1995). Unemployment benefits and labor market transitions: a multinomial logit model with errors in classification. *The Review of Economics and Statistics* **77**, 207–216.

Potter, F. (1990). A study of procedures to identify and trim extreme sampling weights. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, Alexandria, Virginia, 225–230.

Pradel, R. (1996). Utilization of capture-mark-recapture for the study of recruitment and population growth rate. *Biometrics* **52**, 703–709.

Preston, J., Anakotta, T. (2007). Replicate Variance Estimation and High Variance Approximation. *Proceedings of the Third Statistical Conference on Establishment Surveys (ICES-III)*, Montreal, Quebec, Canada.

Preston, J., Henderson, T.S. (2007). Replicate variance estimation and high-entropy variance approximations. *Presented at Third International Conference on Establishment Surveys*, Montreal, QC, Canada.

Price, V. (2006). Citizens deliberation online: theory and some evidence. In: Davies, T., Noveck, B.S. (Eds.), *Online Deliberation: Design, Research, and Practice*. CSLI Publications, University of Chicago Press, Chicago, IL.

Price, V., Neijens, P. (1998). Deliberative polls: Toward improved measure of "informed" public opinion. *International Journal of Public Opinion Research* **10**(2), 147–176.

PSID (2007). *An Overview of the Panel Study of Income Dynamics.* Available at: http://psidonline.isr.umich.edu/Guide/Overview.html.

Qin, J., Leung, D., Shao, J. (2002). Estimation with survey data under nonignorable nonresponse or informative sampling. *Journal of the Indian Statistical Association* **97**, 193–200.

Qualité, L. (2008). A comparison of conditional Poisson sampling versus unequal probability sampling with replacement. *Journal of Statistical Planning and Inference* **138**, 1428–1432.

Quenneville, B., Cholette, P., Hidiroglou, M.A. (2003). Estimating calendar month values from data with various reporting frequencies. *Proceedings of Business and Economic Statistics Section, American Statistical Association*, CD-ROM.

Quenouille, M.H. (1949). Problems in plane sampling. *The Annals of Mathematical Statistics* **20**, 335–375.

Quenouille, M.H. 1956. Notes on bias and estimation. *Biometrika* **43**, 353–360.

Raghunathan, T.E., Reiter, J.P., Rubin, D.R. (2003). Multiple imputation for statistical disclosure limitation. *Journal of Official Statistics* **19**, 1–16.

Raghunathan, T.E., Solenberger, P., Van Hoewyk, J. (2002). *IVEware: Imputation and Variance Estimation Software Users Guide*. University of Michigan, Institute for Social Research, Survey Research Center. Ann Arbor, Michigan, USA.

Rancourt, E., Särndal, C.-E., Lee, H. (1994). Estimation in the presence of nearest neighbour imputation. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 888–893.

Rao, C.R. (1965). On discrete distributions arising out of methods of ascertainment, *Sankhya* **27**, 311–324.

Rao, J.N.K. (1965). On two simple schemas of unequal probability sampling without replacement. *Journal of the Indian Statistical Association* **3**, 173–180.

Rao, C.R. (1971). Some aspects of statistical inference in problems of sampling for finite populations. In: Godambe, V.P., Sprott, D.A. (Eds.), *Foundations of Statistical Inference*. Rinehart & Winston, Toronto, pp. 177–202.

Rao, J.N.K. (1966). Alternative estimators in PPS sampling for multiple characteristics. *Sankhyā, Series A* **28**, 47–60.

Rao, J.N.K. (1985). Conditional inference in survey sampling. *Survey Methodology* **11**(1), 15–31.

Rao, J.N.K. (1990). Variance estimation under imputation for missing data. Technical Report, Statistics Canada, Ottawa, Canada.

Rao, J.N.K. (1996). On variance estimation with imputed survey data. *Journal of American Statistical Association* **91**, 499–506.

Rao, J.N.K. (2003). *Small Area Estimation*. John Wiley & Sons, New York.

Rao, J.N.K., Bayless, D.L. (1969). An empirical study of the stabilities of estimators and variance estimators in unequal probability sampling of two units per stratum. *Journal of the American Statistical Association* **64**, 540–549.

Rao, J.N.K., Shao, J. (1992). Jackknife variance estimation with survey data under hot deck imputation. *Biometrika* **79**, 811–822.

Rao, J.N.K., Singh, M.P. (1973). On the choice of estimator in survey sampling. *Australian Journal of Statistics* **2**, 95–104.

Rao, J.N.K., Sitter, R.R. (1995). Variance estimation under two-phase sampling with application to imputation for missing data. *Biometrika* **82**, 453–460.

Rao, J.N.K., Skinner, C.J. (1996). Estimation in dual frame surveys with complex designs. *Proceedings of the Survey Methods Section*, Statistical Society of Canada, 63–68.

Rao, J.N.K., Wu, C.F.J. (1988). Resampling inference for complex survey data. *Journal of 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.

Reiter, J.P. (2002). Satisfying disclosure restrictions with synthetic data sets. *Journal of Official Statistics* **18**, 531–543.

Reiter, J.P. (2003). Inference for partially synthetic, public use microdata sets. *Survey Methodology* **29**, 181–188.

Reiter, J.P. (2005). Estimating risks of identification disclosure in microdata. *Journal of the American Statistical Association* **100**, 1103–1112.

Reiter, J.P., Raghunathan, T.E., Kinney, S. (2006). The importance of modeling the sampling design in multiple imputation for missing data. *Survey Methodology* **32**, 143–149.

Ren, R., Chambers, R.L. (2002). Outlier robust imputation of survey data via reverse calibration. Southampton Statistical Sciences Research Institute Methodology Working Paper M03/19, University of Southampton, Southampton.

Renaud, A. (2007). Estimation of the coverage of the 2000 census of population in Switzerland: methods and results. *Survey Methodology* 33, 199–210.

Rips, L.J., Conrad, F.G., Fricker, S.S. (2003). Straightening the seam effect in panel surveys. *Public Opinion Quarterly* **67**, 522–554.

Ritter, K.J., Leecaster, M. (2007). Multi-lag cluster enhancement of fixed grid sample designs for estimating the variogram in near coastal systems. *Environmental and Ecological Statistics* **14**, 41–53.

Rivest, L.-P. (1994). Statistical properties of Winsorized means for skewed distributions. *Biometrika* **81**, 373–383.

Rivest, L.-P. (1999). Stratum jumpers: Can we avoid them? *Proceedings of the Survey Research Methods Section*, American Statistical Association, Alexandria, Virginia, pp. 64–72.

Rivest, L.-P. (2002). A generalization of Lavallée and Hidiroglou Algorithm for stratification in business survey. *Survey Methodology* **28**, 207–214.

Rivest, L.-P., Hidiroglou, M. (2004). Outlier treatment for disaggregated estimates. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, Alexandria, Virginia, 4248–4256.

Rivest, L.-P., Hurtubise, D. (1995). On Searls' Winsorized means for skewed populations. *Survey Methodology* **21**, 119–129.

Rivest, L.-P., Levesque, T. (2001). Improved log-linear model estimators of abundance in capture-recapture experiments. *The Canadian Journal of Statistics* **29**, 555–572.

Rivière, P. (2001). Coordinating samples using the Microstrata Methodology. *Proceedings of Statistics Canada Symposium 2001: Achieving Data Quality in a Statistical Agency - A Methodological Perspective*, Ottawa.

Rizzo, L., Brick, J.M., Park, I. (2004). A minimally intrusive method for sampling persons in random digit dial surveys. *Public Opinion Quarterly* **68**, 267–274.

Rizzo, L., Kalton, G., Brick, J.M. (1996). A comparison of some weighting adjustments for panel nonresponse. *Survey Methodology* **22**, 43–53.

Roberts, M.J., Schimmelphennig, D. (2006). Public information creates value: A case study of the USDA Soybean Rust Coordinated Framework finds that the value of the information provided by the framework exceeds its cost. *Amber Waves*. Economic Research Service, U.S. Department of Agriculture. Washington, DC.

Robins, J.M., Sued, M., Lei-Gomez, Q., Rotnitzky, A. (2008). Performance of double-robust estimators when inverse probability weights are highly variable. *Statistical Science* **22**, 544–559.

Robinson, P.M., Särndal, C.-E. (1983). Asymptotic properties of the generalized regression estimator in probability sampling. *Sankhyā* **B45**, 240–248.

Rogers, L. (1949). *The Pollsters: Public Opinion, Politics, and Democratic Leadership*. Alfred A. Knopf, New York.

Rose, D. (Ed.). (2000). *Researching Social and Economic Change*. Routledge, London.

Rosén, B. (1991). *Variance estimation for systematic pps-sampling*. Technical Report 1991:15, Statistics Sweden, Sweden.

Rosén, B. (1995). *Asymptotic theory for order sampling*. R&D Report, Statistics Sweden, Sweden.

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 probabilties for order $\pi ps$ sampling. *Journal of Statistical Planning and Inference* **90**, 117–143.

Rosén, B. (2001). User's guide to Pareto's sampling. *Proceedings of ICES II The Second International Conference on Establishment Surveys.* 289–299.

Rosenbaum, P.R., Rubin, D.B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70**, 41–55.

Rosenstiel, T. (2005). Political polling and the new media culture: a case of more being less. *Special Issue: Polling Politics, Media, and Election Campaigns. Public Opinion Quarterly* **69**(5), 698–715.

Rosenthal, J. (1998). The Lives They Lived: Ruth Clark, The Right Questions. *The New York Times Magazine*, January 4, Volume 147, Section 6, p. 41.

Roslow, S., Roslow, L. (1972). Unlisted phone-subscribers are different. *Journal of Advertising Research* **12**, 35–38.

Rousseau, S., Tardieu, F. (2004). *La macro SAS CUBE d'échantillonnage équilibré, Documentation de l'utilisateur*. Technical Report, INSEE, Paris.

Royall, R.M., Herson, J. (1973). Robust estimation infinite populations I. *Journal of the American Statistical Association* **68**, 880–889.

Royall, R.M. (1970). On finite population sampling theory under certain regression models. *Biometrika* **57**, 377–387.

Royall, R.M. (1971). Linear regression models in finite population sampling theory (with discussion). In: Godambe, V.P., Sprott, D.A. (Eds.), *Foundations of Statistical Inference.* Holt, Rinehart and Winston, Toronto, ON, Canada, pp. 259–279.

Royall, R.M. (1976). Current advances in sampling theory: implications for observational studies. *American Journal of Epidemiology* **104**, 463–474.

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. (1981a). An empirical study of the ratio estimator and estimators of its variance. *Journal of the American Statistical Association* **76**, 66–77.

Royall, R.M., Cumberland, W.G. (1981b). The finite population linear regression estimator and estimators of its variance–an empirical study. *Journal of the American Statistical Association* **76**, 924–930.

Royall, R.M., Eberhardt, K.R. (1975). Variance estimates for the ratio estimator. *Sankhyā*, *Series C* **37**, 43–52.

Royall, R.M., Herson, J. (1973a). Robust estimation in finite populations I. *Journal of the American Statistical Association* **68**, 880–889.

Royall, R.M., Herson, J. (1973b). Robust estimation in finite populations II: stratification on a size variable. *Journal of the American Statistical Association* **68**, 890–893.

Royston, P. (2004). Multiple imputation of missing values. *The Stata Journal* **4**, 227–241.

Royston, P. (2005). Multiple imputation of missing values: Update. *The Stata Journal* **5**, 1–14.

RTI-Research Triangle Institute. (2001). *Evaluation of the Voter News Service's Procedures and Operations for the 2000 Presidential Election.* Research Triangle Institute, Research Triangle Park, NC.

Rubin, D.B. (1976). Inference and missing data. *Biometrika* **63**, 581–590.

Rubin, D.B. (1987). *Multiple Imputations for Nonresponse in Surveys*. Wiley, New York.

Rubin-Bleuer, S. (2002). Report on Rivière's Random Permutations Method of sampling co-ordination. Internal report, Statistics Canada, Ottawa.

Rust, K.F., Rao, J.N.K. (1996). Variance estimation for complex surveys using replication techniques. *Statistical Methods in Medical Research* **5**, 281–310.

Saalfeld, A. (1991). Construction of spatially articulated list frames for household surveys. *Proceedings of Statistics Canada Symposium 91. Spatial Issues in Statistics.* Statistics Canada, Ottawa, ON, Canada, 41–53.

Saavedra, P.J., Weir, P. (2003). The use of permanent random numbers in a multi-product petroleum sales survey: twenty years of a developing design. *Proceedings of Federal Committee on Statistical Methodology Research Conference*, Washington, DC.

Sadasivan, G., Sharma, S. (1974). Two dimensional varying probability sampling without replacement. *Sankhyā* **C36**, 157–166.

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.

Salazar, J.J. (2003). Partial cell suppression: a new methodology for statistical disclosure control. *Statistics and Computing* **13**, 13–21.

Salazar-González, J.J., Lowthian, P., Young, C., Merola, G., Bond, S., Brown, D. (2004). Getting the best results in controlled rounding with the least effort. In: Domingo-Ferrer, J., Torra, V. (Eds.), *Privacy in Statistical Databases.* Springer-Verlag, Berlin, pp. 58–71.

Salehi, M.M., Seber, G.A.F. (1997a). Adaptive cluster sampling with networks selected without replacement. *Biometrika* **84**, 209–219.

Salehi, M.M., Seber, G.A.F. (1997b). Two stage adaptive cluster sampling. *Biometrics* **53**, 959–970.

Salehi, M.M., Seber, G.A.F. (2002). Unbiased estimators for restricted adaptive cluster sampling. *Australian and New Zealand Journal of Statistics* **44**, 63–74.

Salganik, M.J. (2006). Variance estimation, design effects, and sample size calculations for respondent-driven sampling. *Journal of Urban Health-Bulletin of the New York Academy of Medicine* (Suppl. S) **83**, I98–I112.

Samiuddin, M., Hanif, M. (2007). Estimation of population mean in single and two phase sampling with or without additional information. *Pakistan Journal of Statistics* **23**, 99–118.

Sampford, M.R. (1967). On sampling without replacement with unequal probabilities of selection. *Biometrika* **54**, 499–513.

Samuels, S.M. (1998). A Bayesian species-sampling-inspired approach to the uniques problem in microdata disclosure risk assessment. *Journal of Official Statistics* **14**, 373–383.

Sande, G. (1978). *An Algorithm for the Fields to Impute Problems of Numerical and Coded Data*. Technical Report, Statistics Canada, Canada.

Sanders, S. (2002). *Selectief Gaafmaken m.b.v. Classificatie-en Regressiebomen* (in Dutch). Statistics Netherlands, Voorburg.

Saris, W. (1998). Ten years of interviewing without interviewers: the telepanel. In: Couper, N.P., Baker, R.P., Bethlehem, J., Clark, C.Z.F., Martin, J., Nichols II, W.L., O'Reilly, J.M. (Eds.), *Computer Assisted Survey Information Collection*. John Wiley & Sons, New York, pp. 409–429.

Saris, W., Andrews, F. (1991). Evaluation of measurement instruments using a structural modeling approach. In: Biemer P., Groves, R., Lyberg, L., Mathiowetz, N., Sudman, S. (Eds.), *Measurement Errors in Surveys*. John Wiley & Sons, New York, pp. 575–599.

Saris, W., van der Veld, W., Gallhofer, I. (2004). Developments and improvements of questionnaires using predictions of reliability and validity. In: Presser S., J.M. Rothgeb, M. Couper, J.T. Lessler, E. Martin, J. Martin, and E.Singer, (Eds.), *Methods for Testing and Evaluating Survey Questionnaires*. John Wiley & Sons, Hoboken, NJ, pp. 275–299.

Särndal, C.-E. (1992). Method for estimating the precision of survey estimates when imputation has been used. *Survey Methodology* **18**, 241–252.

Särndal, C.-E., Lundström, S. (2005). *Estimation in Surveys with Nonresponse*. John Wiley & Sons Ltd, 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 non-response. *International Statistical Review* **55**, 279–294.

Särndal, C.-E., Swensson, B., Wretman, J. (1992). *Model Assisted Survey Sampling.* Springer-Verlag, New York.

Sautory, O. (1993). *La macro CALMAR. Redressement d'un échantillon par calage sur marges*. INSEE, Série des documents de travail n° F 9310. INSEE, Paris, France.

Sautory, O. (2003). CALMAR2: A new version of the CALMAR calibration adjustment program. *Proceedings of Statistics Canada's Symposium 2003*. Available at: http://www.statcan.ca/english/freepub/11-522-XIE/2003001/session13/sautory.pdf.

Schabenberger, O., Gotway, C.A. (2005). *Statistical Methods for Spatial Data Analysis*. Chapman & Hall, Boca Raton, FL.

Schafer, J.L. (1997). *Analysis of Incomplete Multivariate Data*. CRC Press, New York.

Schafer, J.L., Graham, J.W. (2002). Missing data: our view of the state of the art. *Psychological Methods* **7**, 147–177.

Schaffer, J. (1987). Procedure for solving the data-editing problem with both continuous and discrete data types. *Naval Research Logistics* **34**, 879–890.

Scheffer, J. (2002). Dealing with missing data. *Research Letters in the Information and Mathematical Sciences* **3**, 153–160.

Scherpenzeel, A., Saris, W. (1997). The validity and reliability of survey questions: a meta-analysis of MTMM studies. *Sociological Methods and Research* **25**, 341–383.

Scheuren, F. (2005). Paradata from concept to completion. *Proceedings of Symposium 2005: Methodological Challenges for Future Information Needs*, Statistics Canada, Ottawa, Canada.

Scheuren, F., Winkler, W.E. (1993). Regression analysis of data files that are computer matched. *Survey Methodology* **19**, 39–58.

Scheuren, F., Winkler, W.E. (1997). Regression analysis of data files that are computer matched, II. *Survey Methodology* **23**, 157–165.

Schillewaert, N., Meulemeester, P. (2005). Comparing response contributors of offline and online data collection methods. *International Journal of Market Research* **47**(2), 163–178.

Schindler, E. (2005). Analysis of Census 2000 long form variances. *Proceedings of the Survey Research Methods Section*, American Statistical Association. Minneapolis, Minnesota, 3526–3533. Available at: http://www.amstat.org/sections/srms/Proceedings/y2005/Files/JSM2005-000396.pdf.

Scholtus, S. (2008a). *Algorithms for Detecting and Resolving Obvious Inconsistencies in Business Survey Data*. UN/ECE Work Session on Statistical Data Editing, Vienna, Austria.

Scholtus, S. (2008b). *Algorithms for Correcting Some Obvious Inconsistencies and Rounding Errors in Business Survey Data*. Discussion paper, Statistics Netherlands, Voorburg.

Schreuder, H.T., Gregoire, T.G., Wood, G.B. (1993). *Sampling Methods for Multiresource Forest Inventory*. Wiley, New York.

Schuman, H., Presser, S. (1981). *Questions and Answers in Attitude Surveys: Experiments in Question Form, Wording, and Context*. Academic Press, New York.

Schwarz, C.J. (2001). The Jolly-Seber model: more than just abundance. *Journal of Agricultural, Biological and Environmental Statistics* **6**, 195–205.

Schwarz, C.J., Arnason, A.N. (1996). A general methodology for the analysis of open-model capture recapture experiments. *Biometrics* **52**, 860–873.

Schwarz, C.J., Arnason, A.N. (2000). The estimation of age-specific breeding probabilities from capture-recapture data. *Biometrics* **56**, 59–64.

Schwarz, C.J., Seber, G.A.F. (1999). Estimating animal abundance: review III. *Statistical Science* **14**, 427–456.

Searls, D.T. (1966). An estimator for a population mean which reduces the effect of large true observations. *Journal of the American Statistical Association* **61**, 1200–1204.

Seber, G.A.F. (1965). A note on the multiple recapture census. *Biometrika* **52**, 249–259.

Seber, G.A.F., Schwarz, C.J. (2002). Capture-recapture: before and after EURING 2000. *Journal of Applied Statistics* **29**(1), 5–18.

Sekar, C.C., Deming, W.E. (1949). On a method of estimating birth and death rates and the extent of registration. *Journal of the American Statistical Association* **44**, 101–115.

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.

Sethi, Y.K. (1963). A note on optimum stratification of populations for estimating the population means. *Australian Journal of Statistics* **5**, 20–33.

Seventeenth International Conference of Labour Statisticians. (2002). Resolution I - Resolution concerning household income and expenditure statistics. *The Seventeenth International Conference of Labour Statisticians and International Labour Organization*, Final Report, 43–57, International Labour Organization, Jeneva.

Shabbir, J., Gupta, S. (2007). On estimating the finite population mean with known population proportion of an auxiliary variable. *Pakistan Journal of Statistics* **23**, 1–9.

Shao, J. (1994). L-statistics in complex survey problems. *The Annals of Statistics* **22**, 946–967.

Shao, J. (2000). Cold deck and ratio imputation. *Survey Methodology* **26**, 79–85.

Shao, J. (2002). Replication methods for variance estimation in complex surveys with imputed data. In: Groves, R.M., Dillman, D.A., Eltinge, J.L., Little, R.J.A. (Eds.), *Survey Nonresponse*. John Wiley and Sons, New York, pp. 303–314.

Shao, J. (2003). Impact of the bootstrap on sample surveys. *Statistical Science* **18**, 191–198.

Shao, J. (2007). Handling survey nonresponse in survey sampling. *Survey Methodology* **33**, 81–85.

Shao, J., Sitter, R.R. (1996). Bootstrap for imputed survey data. *Journal of the American Statistical Association* **93**, 819–831.

Shao, J., Steel, P. (1999). Variance estimation for survey data with composite imputation and nonnegligible sampling fractions. *Journal of the American Statistical Association* **94**, 254–265.

Shao, J., Tu, D. (Eds.). (1995). *The Jacknife and Bootstrap*. Sprinter-Verlag, New York.

Shao, J., Wang, H. (2002). Sample correlation coefficients based on survey data under regression imputation. *Journal of the American Statistical Association* **97**, 522–544.

Shao, J., Wang, H. (2008). Confidence intervals based of survey data with nearest neighbor imputation. *Statistica Sinica* **18**, 281–298.

SHARE (2007). *Survey of Health, Ageing and Retirement in Europe*. Available at: http://www.share-project.org/.

Sheatsley, P.B., Mitofsky, W.J., (Eds.). (1992). *A Meeting Place: The History of the American Association for Public Opinion Research*. American Association for Public Opinion Research, USA.

Shih, W.P. (1980). An evaluation of random digit dialing household surveys. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, Alexandria, VA, 736–739.

Shlomo, N., De Waal, T. (2008). Protection of micro-data subject to edit constraints against statistical disclosure. *Journal of Official Statistics* **24**, 229–253.

Silva, D.B.d.N., Cruz, M.M. (2002). *Séries Temporais de Pesquisas Amostrais Periódicas*. Associação Brasileira de Estatística, São Paulo, Brazil.

Silva, P.L.D.N. (2003). Calibration Estimation: When and Why, How Much and How. Technical Report. University of Southampton, UK.

Sinclair, M. (1994). *Evaluating Reinterview Survey Methods for Measuring Response Errors*. Unpublished Ph.D. dissertation, George Washington University, Department of Statistics, Washington, D.C.

Sinclair, M., Gastwirth, J. (1996). On procedures for evaluating the effectiveness of reinterview survey methods: application to labor force data. *Journal of the American Statistical Association* **91**, 961–969.

Singer, E. (Ed.). (2006). *Special Issue: Non-response Bias in Household Surveys. Public Opinion Quarterly* **70**(5).

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., 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**(430), 478–488.

Singh, M.P., Gambino, J., Mantel, H. (1994). Issues and strategies for small area data (with discussion). *Survey Methodology* **20**, 3–22.

Singh, R. (1971). Approximately optimum stratification on the auxiliary variable. *Journal of the American Statistical Association* **66**, 829–833.

Sirken, M.G. (1972). *Designing Forms for Demographic Surveys*. Laboratory for Population Statistics Manual Series, No. 3, University of North Carolina, Chapel Hill, NC.

Sirken, M.G. (1970). Household surveys with multiplicity. *Journal of the American Statistical Association* **65**, 257–266.

Sitter, R.R. (1992a). Comparing three bootstrap methods for survey data. *The Canadian Journal of Statistics* **20**, 135–154.

Sitter, R.R. (1992b). A resampling procedure for complex survey data. *Journal of the American Statistical Association* **87**, 755–765.

Skalski, J.R. (1990). A design for long-term status and trends monitoring. *Journal of Environmental Management* **30**, 139–144.

Skinner, C.J. (1991). On the efficiency of raking ratio estimation for multiple frame surveys. *Journal of the American Statistical Association* **86**, 779–784.

Skinner, C.J. (1992). On identification disclosure and prediction disclosure for microdata. *Statistica Neerlandica* **46**, 21–32.

Skinner, C.J. (1999). Calibration weighting and non-sampling errors. *Research in Official Statistics* **1**, 33–43.

Skinner, C.J. (2007). The probability of identification: applying ideas from forensic science to disclosure risk assessment. *Journal of the Royal Statistical Society, Series A* **170**, 195–212.

Skinner, C.J., Carter, R.G. (2003). Estimation of a measure of disclosure risk for survey microdata under unequal probability sampling. *Survey Methodology* **29**, 177–180.

Skinner, C.J., Elliot, M.J. (2002). A measure of disclosure risk for microdata. *Journal of the Royal Statistical Society, Series B* **64**, 855–867.

Skinner, C.J., Holmes, D.J. (1993). Modelling population uniqueness. *International Seminar on Statistical Confidentiality Proceedings*, European Community Statistical Office, Luxembourg, 175–199.

Skinner, C.J., Holmes, D.J. (1998). Estimating the re-identification risk per record in microdata. *Journal of Official Statistics* **14**, 361–372.

Skinner, C.J., Holmes, D.J., Holt, D. (1994). Multiple frame sampling for multivariate stratification. *International Statistical Review* **62**, 333–347.

Skinner, C.J., Holt, D., Smith, T.M.F. (1989). *Analysis of Complex Surveys*. John Wiley & Sons, Chichester, UK.

Skinner, C.J., Marsh, C., Openshaw, S., Wymer, C. (1994). Disclosure control for census microdata. *Journal of Official Statistics* **10**, 31–51.

Skinner, C.J., Rao, J.N.K. (1996). Estimation in dual frame surveys with complex designs. *Journal of the American Statistical Association* **91**, 349–356.

Skinner, C.J., Rao, J.N.K. (2002). Jackknife variance for multivariate statistics under hot deck imputation from common donors. *Journal of Statistical Planning and Inference* **102**, 149–167.

Skinner, C.J., Shlomo, N. (2007). Assessing the disclosure protection provided by misclassification and record swapping. Paper presented at 56th Session of International Statistical Institute, Lisbon; to appear in *Bulletin of International Statistical Institute*, pp. 1–2.

Skinner, C.J., Shlomo, N. (2008). Assessing identification risk in survey microdata using log-linear models. *Journal of American Statistical Association* **103**, in press, 989–1001.

Slanta, J.G., Krenzke, T.R. (1996). Applying Lavallée-Hidiroglou Method to obtain stratification boundaries for the Census Bureau's Annual Capital Expenditures Survey. *Survey Methodology* **22**, 65–76.

Smead, R.J., Wilcox, J. (1980). Ring policy in telephone surveys. *Public Opinion Quarterly* **44**, 115–116.

Smith, D.R. (2006). Survey design for detecting rare freshwater mussels. *Journal of the North American Benthological Society* **25**, 701–711.

Smith, D.R., Conroy, M.J., Brakhage, D.H. (1995). Efficiency of adaptive cluster sampling for estimating density of wintering waterfowl. *Biometrics* **51**, 777–788.

Smith, M.E., Newcombe, H.B. (1975). Methods of computer linkage for hospital admission-separation records into cumulative health histories. *Method of Information in Medicine* **14**(3), 118–125.

Smith, P., Pont, M., Jones, T. (2003). Developments in business survey methodology in the Office for National Statistics, 1994–2000. *Journal of the Royal Statistical Society*: *Series D* **52**, 257–295.

Smith, P.J., Hoaglin, D.C., Rao, J.N.K., Battaglia, M.P., Daniels, D. (2004). Evaluation of adjustment for partial non-response bias in the US National Immunization Survey. *Journal of the Royal Statistical Society A* **167**, 141–156.

Smith, T.M.F. (2001). Biometrika centenary: sample surveys. *Biometrika* **88**, 167–194.

Smith, T.W. (1978). In search of house effects: A comparison of responses to various questions by different survey organization. *Public Opinion Quarterly* **42**, 443–463.

Smith, T.W. (1990). The first straw: A study of the origins of election polls. *Public Opinion Quarterly* **54**, 21–36.

Smith, T.W. (1995). Trends in non-response rates. *International Journal of Public Opinion Research* **7**(2), 157–171.

Smith, T.W. (2002). Developing nonresponse standards. In: Groves, R.M., Dillman, D.A., Eltinge, J.L., Little, R.J.A. (Eds.), *Survey Nonresponse.* Wiley, New York, pp. 27–40.

Sobol, M. (1959). Panel mortality and panel bias. *Journal of the American Statistical Association* **54**, 52–68.

Solon, G. (1986). Effects of rotation group bias on estimation of unemployment. *Journal of Business and Economic Statistics* **4**, 105–109.

Sorić, B. (1989). Statistical "discoveries" and effect-size estimation. *Journal of the American Statistical Association* **84**, 608–610.

Spangenberg, F. (2003). Foundation For Information Report, *The Freedom to Publish Opinion Poll Results: Report on a Worldwide Update*, Amsterdam: ESOMAR/WAPOR, Amsterdam.

Squire, P. (1988). Why the 1936 Literary Digest Poll failed. *Public Opinion Quarterly* **52**, 125–133.

Srinath, K.P., Battaglia, M.P., Frankel, M.R., Khare, M., Zha, W. (2002). Evaluation of a procedure based on interruptions in telephone service for reducing coverage bias in RDD surveys. *Proceedings of the Section on Survey Research Methods* (CD-ROM), American Statistical Association, Alexandria, VA, 3323–3329.

Srinath, K.P., Carpenter, R.M. (1995). Sampling methods for repeated business surveys. In: Cox, B.G., Binder, D.A., Chinnappa, B.N., Christianson, A., Colledge, M.J., Kott, P.S. (Eds.), *Business Survey Methods*. John-Wiley and Sons, New York, pp. 171–183.

Statistics Canada (1998). *Methodology of the Canadian Labour Force Survey*. Statistics Canada, Ottawa, Canada.

Statistics Canada. (2004a). *Coverage.* 2001 Census Technical Report Series, Catalogue No. 92-394-XIE, Ottawa, Canada. Available at: http://www12.statcan.ca/english/census01/Products/Reference/tech_rep/coverage/offline%20documents/92-394-XIE.pdf.

Statistics Canada. (2004b). *Symposium 2004: Innovative Methods for Surveying Difficult-to-Reach Populations (Proceedings)*. Statistics Canada, Ottawa, Canada.

Statistics Canada (2008). *Survey of Labour and Income Dynamics: Survey Design*. Available at: statcan.gc.ca/pub/75f0011x/2008001/5203381_eng.htm.

Statistics Canada. (2006). *International Methodology Symposium: Methodological Issues in Measuring Population Health (Proceedings).* Statistics Canada, Ottawa, Canada.

Statistics Canada (2008). *Survey of Labour and Income Dynamics: Survey Design*. Available at: statcan.gc.ca/pub/75f0011x/2008001/5203381_eng.htm.

Statistics Finland. (2004). Use of registers and administrative data sources for statistical purposes: best practices at Statistics Finland. In: Myrskylä, P. (Ed.), *Handbooks*, 70. Statistics Finland, Helsinki, Finland.

Statistics Netherlands. (2002). *Blaise Developers Guide*. Methods and Informatics Department, Heerlen, The Netherlands.

Statistics Netherlands. (2004). *The Dutch Virtual Census of 2001: Analysis and Methodology*. Statistics Netherlands, Voorburg, The Netherlands.

Statistics Norway. (2005). Plans for 2010 Population and Housing Census in Norway. Statistics Norway, available from United Nations Statistics Division website. Available at: http://unstats.un.org.

Steeh, C., Kirgis, N., Cannon, B., DeWitt, J. (2001). Are they really as bad as they seem? Nonresponse rates at the end of the Twentieth Century. *Journal of Official Statistics* **17**(2), 227–247.

Steel, D. (1997). Producing monthly estimates of unemployment and employment according to the International Labour Office definition. *Journal of the Royal Statistical Society, Series A* **160**, 5–46.

Steele, F., Brown, J., Chambers, R. (2002). A controlled donor imputation system for a one-number census. *Journal of the Royal Statistical Society, Series A* **165**, 495–522.

Stefanski, L.A., Bay, J.M. (1996). Simulation extrapolation deconvolution of finite population cumulative distribution function estimators. *Biometrika* **83**, 407–417.

Stefanski, L.A., Buzas, J.S. (1995). Instrumental variable estimation in binary regression measurement error models. *Journal of the American Statistical Association* **90**, 541–550.

Stefanski, L.A., Cook, J.R. (1995). Simulation-extrapolation: the measurement error jackknife. *Journal of the American Statistical Association* **90**, 1247–1256.

Stern, D.E. Jr., Steinhorst, R.K. (1984). Telephone interview and mail questionnaire applications of the randomized response model. *Journal of the American Statistical Association* **79**, 555–564.

Stevens, D.L. Jr. (1994). Implementation of a national environmental monitoring program. *Journal of Environmental Management* **42**, 1–29.

Stevens, D.L. Jr. (1997). Variable density grid-based sampling designs for continuous spatial populations. *Environmetrics* **8**, 167–195.

Stevens, D.L. Jr. (2002). Sampling design and statistical analysis methods for the integrated biological and physical monitoring of Oregon streams, Oregon Department of Fish and Wildlife Report Number OPSW-ODFW-2002-07. 14 pages + appendices.

Stevens, D.L. Jr. (2006). Spatial properties of design-based versus model-based approaches to environmental sampling. *Proceedings of Accuracy 2006*: *The 7th international symposium on spatial accuracy assessment in natural resources and environmental sciences*. Instituto Geographica Portugales, Lisboa, Portugal, 119–125 of 908.

Stevens, D.L. Jr., Olsen, A.R. (1999). Spatially restricted surveys over time for aquatic resources. *Journal of Agricultural, Biological, and Environmental Statistics* **4**, 415–428.

Stevens, D.L. Jr., Olsen, A.R. (2003). Variance estimation for spatially balanced samples of environmental resources. *Environmetrics* **14**, 593–610.

Stevens, D.L. Jr., Olsen, A.R. (2004). Spatially balanced sampling of natural resources. *Journal of the American Statistical Association* **99**, 262–278.

Stonecash, J.M. (2003). *Political Polling: Strategic Information in Campaigns*. Rowman & Littlefield, Lanham, Maryland.

Stoop, I.A.L. (2005). *The hunt for the Last Respondent: Nonresponse in Sample Surveys.* Social and Cultural Planning Office, The Hague.

Stouffer, S.A., Suchman, E.A., Vinney, L.C., Star, S.A., Williams, R.M. Jr. (1965 [1949]). *The American Soldier: Adjustments During Army Life*, vol. 1. John Wiley & Sons, New York.

Strachan, C. (2003). *High-Tech Grass Roots: The Professionalization of Local Elections.* Rowman & Littlefield, Lanham, Maryland.

Sturgis, P., Allum, N., Brunton-Smith, I. (2009). Attitudes over time: the psychology of panel conditioning. In: Lynn, P. (Ed.), *Methodology of Longitudinal Surveys*. John Wiley & Sons, Chichester, UK, pp. 113–126.

Su, Z., Quinn, T.J. II. (2003). Estimator bias and efficiency for adaptive cluster sampling with order statistics and a stopping rule. *Environmental and Ecological Statistics* **10**, 17–41.

Sudman, S. (1962). *On the Accuracy of Recording Consumer Panels*. Unpublished Ph.D. dissertation, University of Chicago.

Sudman, S. (1973). The uses of telephone directories for survey sampling. *Journal of Marketing Research* **10**(2), 204–207.

Sudman, S. (1978). Optimum cluster designs within a primary unit using combined telephone screening and face-to-face interviewing. *Journal of the American Statistical Association* **73**, 300–304.

Sudman, S. (1980). Improving the quality of shopping center sampling. *Journal of Marketing Research* **17**, 423–431.

Sudman, S., Bradburn, N.M. (1987). The organizational growth of public opinion research in the United States. *Public Opinion Quarterly* **52**(Part 2: Supplement: 50th Anniversary Issue), S67–S78.

Sudman, S., Bradburn, N.M., Schwartz, N. (1996). *Thinking About Answers: The Application of Cognitive Processes to Survey Methodology*. Jossey-Bass, San Francisco, CA.

Sudman, S., Ferber, R. (1971). Experiments in obtain consumer expenditures by diary methods. *Journal of the American Statistical Association* **66**, 725–735.

Sudman, S., Ferber, R. (1974). A comparison of alternative procedures for collecting consumer expenditure data for frequently purchased products. *Journal of Marketing Research* **11**, 128–135.

Sudman, S., Ferber, R. (1979). *Consumer Panels*. American Marketing Association, Chicago, IL.

Sudman, S., Kalton, G. (1986). New developments in the sampling of special populations. *Annual Review of Sociology* **12**, 401–429.

Sudman, S., Sirken, M.G., Cowan, C.D. (1988). Sampling rare and elusive populations. *Science* **240**, 991–996.

Sullivan, G.R., Fuller, W.A. (1989). The use of measurement error to avoid disclosure. *Proceedings of the Survey Research Methods Section*, American Statistical Association, Washington, DC, 802–807.

Sunter, A. (1986). Implicit longitudinal sampling from administrative files: a useful technique. *Journal of Official Statistics* **2**, 161–168.

Sweeney, L. (1999). Computational disclosure control for medical microdata: the datafly system. In: *Record Linkage Techniques 1997*, National Academy Press, Washington, DC, pp. 442–453.

Sweet, E.M., Sigman, R.S. (1995). Evaluation of model-assisted procedures for stratifying skewed populations using auxiliary data. *Proceedings of the Survey Methods Section*, *American Statistical Association*, Alexandria 491–496.

Synodinos, N.E., Yamada, S. (2000). Response rate trends in Japanese surveys. *International Journal of Public Opinion Research* **12**(1), 48–72.

Takahasi, K., Wakimoto, K. (1968). On unbiased estimates of the population mean based on the sample stratified by means of ordering. *Annals of the Institute of Statistical Mathematics* **20**, 1–31.

Tambay, J.-L. (1988). An integrated approach for the treatment of outliers in sub-annual surveys. *Proceedings of the Survey Research Methods Section*, American Statistical Association, Alexandria, Virginia, 229–234.

Tambay, J.-L., Mohl, C. (1995). Improving sample representativity through the use of a rejective method. *Proceedings of the Survey Research Methods Section, Joint Statistical Meetings*, 29–38, American Statistical Association, Arlington, Virginia, USA.

Tardieu, F. (2001). *Echantillonnage équilibré: de la théorie á la pratique*. Technical Report, INSEE, Paris.

Taylor, H. (2000). Does Internet research 'work?' Comparing online survey results with telephone surveys. *International Journal of Market Research* **42**, 51–64.

Taylor, H., Terhanian, G. (1999). Heady days are here again: online polling is rapidly coming of age. *Public Perspective* **11**, 20–23.

Taylor, M.F. (Ed.), with Brice, J., Buck, N., Prentice-Lane, E. (2007). *British Household Panel Survey User Manual Volume A: Introduction, Technical Report and Appendices.* University of Essex, Colchester, Essex.

Teitler, J.O., Reichman, N.E., Sprachman, S. (2003). Costs and benefits of improving response rates for a hard-to-reach population. *Public Opinion Quarterly* **67**, 126–138.

Tempelman, C. (2007). *Imputation of Restricted Data*. Ph.D. Thesis, University of Groningen, The Netherlands.

Tenenbein, A. (1979). A double sampling scheme for estimating from misclassified multinomial data with application to sampling inspection. *Technometrics* **14**(1), 187–202.

Thompson, J.H., Waite, P.J., Fay, R.E. (2001). *ESCAP II: Basis of 'Revised Early Approximations' of Undercounts Released October 17, 2001*. Executive Steering Committee for A.C.E. Policy II, Report 9a, October 26, 2001, U.S. Census Bureau, Washington, DC. Available at: http://www.census.gov/dmd/www/pdf/report9a.pdf.

Thompson, S.K. (1990). Adaptive cluster sampling. *Journal of American Statistical Association* **85**, 1050–1059.

Thompson, S.K. (1991a). Adaptive cluster sampling: designs with primary and secondary units. *Biometrics* **47**, 1103–1117.

Thompson, S.K. (1991b). Stratified adaptive cluster sampling. *Biometrika* **78**, 389–397.

Thompson, S.K. (1996). Adaptive cluster sampling based on order statistics. *Environmetrics* **7**, 123–133.

Thompson, S.K. (2002). *Sampling*, 2nd ed. John Wiley & Sons, New York.

Thompson, S.K. (2006). Adaptive web sampling. *Biometrics* **62**, 1224–1234.

Thompson, S.K., Collins, L.M. (2002). Adaptive sampling in research on risk-related behaviors. *Drug and Alcohol Dependence* **68**, 857–867.

Thompson, S.K., Seber, G.A. F. (1996). *Adaptive Sampling.* Wiley & Sons, New York.

Thornberry, O.T., Massey, J.T. (1988). Trends in United States telephone coverage across time and subgroups. In: Groves, R.M., Biemer, P.P., Lyberg, L.E., Massey, J.T., Nicholls, W.L., Waksberg, J. (Eds.), *Telephone Survey Methodology*. John Wiley & Sons, New York, pp. 29–49.

Tillé, Y. (1996). An elimination procedure for unequal probability sampling without replacement. *Biometrika* **83**, 238–241.

Tillé Y. (2006). *Sampling Algorithms*. Springer-Verlag, New York.

Tillé, Y., Favre, A.-C. (2004). Co-ordination, combination and extension of optimal balanced samples. *Biometrika* **91**, 913–927.

Tillé, Y., Favre, A.-C. (2005). Optimal allocation in balanced sampling. *Statistics and Probability Letters* **74**, 31–37.

Tillé, Y., Matei, A. (2007). *The R Package Sampling*. The Comprehensive R Archive Network, Manual of the Contributed Packages. Available at: http://cran.r-project.org/.

Time-sharing Experiments for the Social Sciences (TESS). Available at: http://www.experimentcentral.org/. [accessed 14.04.07].

Tomaskovic-Devey, D., Leiter, J., Thompson, S. (1994). Organizational survey nonresponse. *Administrative Science Quarterly* **39**(3), 439–457.

Tourangeau, R., Rips, L.J., Rasinski, K. (2000). *The Psychology of Survey Response*. Cambridge University Press, New York.

Traat, I., Bondesson, L., Meister, K. (2004). Sampling design and sample selection through distribution theory. *Journal of Statistical Planning and Inference* **123**, 395–413.

Traugott, M.W. (2005). The accuracy of the national preelection polls in the 2004 presidential election. *Special Issue: Polling Politics, Media, and Election Campaigns. Public Opinion Quarterly* **69**(5), 642–654.

Traugott, M.W., Highton, B., Brady, H.E. (2005). A review of recent controversies concerning the 2004 presidential election exit polls. The National Election Commission on Elections and Voting, Social Science Research Council, New York, March 10.

Traugott, M.W., Lavrakas, P.J. (2004). *The Voters Guide to Election Polls*, 3rd ed. Rowman & Littlefield, Lanham, Maryland.

Traugott, M.W., Price, V. (1992). Review: Exit polls in the 1989 Virginia gubernatorial race: Where did they go wrong? *Public Opinion Quarterly* **56**, 245–253.

Trivellato, U. (1999). Issues in the design and analysis of panel studies: a cursory review. *Quality and Quantity* **33**, 339–352.

Troldahl, V.C., Carter, Jr. R.E. (1964). Random selection of respondents within households in phone surveys. *Journal of Marketing Research* **1**, 71–76.

Truman, D.B. (1945). Public opinion research as a tool of public administration. *Public Administration Review* **5**, 62–72.

Tuckel, P., O'Neill, H. (2002). The vanishing respondent in telephone surveys. *Journal of Advertising Research* **42**(5), 26–48.

Tucker, C., Brick, J.M., Meekins, B. (2007). Household telephone service and usage patterns in the United States in 2004: implications for telephone samples. *Public Opinion Quarterly* **71**, 3–22.

Tucker, C., Casady, R., Lepkowski, J. (1992). Sample allocation for stratified telephone sample designs. *Proceedings of the Survey Research Methods Section.* American Statistical Association, Alexandria, VA, pp. 566–571.

Tucker, C., Lepkowski, J.M., Piekarski, L. (2002). The current efficiency of list-assisted telephone sampling. *Public Opinion Quarterly* **66**, 321–338.

Tucker, C., Meekins, B., Biemer, P. (2006). Estimating the level of underreporting of expenditures among expenditure reporters: a micro-level latent class analysis. *Proceedings of the ASA Survey Research Methods Section*, Minneapolis, MN.

Tukey, J.W. (1977). *Exploratory Data Analysis*. Addison-Wesley, London.

Turk, P., Borkowski, J.J. (2005). A review of adaptive cluster sampling: 1990–2003. *Environmental and Ecological Statistics* **12**, 55–94.

U.K. National Statistics (2007). *Labour Force Survey User Guide–Volume 1: Background & Methodology*. Available at: http://www.statistics.gov.uk/downloads/theme_labour/LFSUG_vol1_2007.pdf.

U.N. Department of Economic and Social Affairs Statistics Division (2005). *Household Sample Surveys in Developing and Transition Countries.* ST/ESA/STAT/SER.F/96. United Nations, New York.

United Nations. (1994). *Statistical Data Editing, Volume 1: Methods and Techniques*. Statistical Standards and Studies, No. 44, United Nations Statistical Commission and Economic Commission for Europe, Geneva, Switzerland, pp. 137–143.

United Nations. (2005a). *Designing Household Survey Samples: Practical Guidelines*. United Nations Statistics Division, New York.

United Nations. (2005b). *Household Sample Surveys in Developing and Transition Countries*. United Nations, New York.

United Nations. (2006). *Principles and Recommendations for Population and Housing Censuses.* Revision 2 draft. United Nations, New York. Available at: http://www.hob.scb.se/UN%20Census2010rec_29September-2006.pdf.

United Nations Economic Commission for Europe (UNECE). (2006). *Conference of European Statisticians Recommendations for the 2010 Censuses of Population and Housing*. United Nations, New York and Geneva. Available at: http://www.unece.org/stats/documents/ece/ces/ge.41/2006/zip.1.e.pdf.

Urquhart, N.S., Kincaid, T.M. (1999). Designs for detecting trend from repeated surveys of ecological resources. *Journal of Agricultural, Biological and Environmental Statistics* **4**, 404–414.

Urquhart, N.S., Overton, W.S., Birkes, D.S. (1993). Comparing sampling designs for monitoring ecological status and trends: impact of temporal patterns. In: Barnett, V., Turkman, K.F. (Eds.), Chapter 3 in *Statistics for the Environment*. John Wiley & Sons, pp. 71–85.

Urquhart, N.S., Paulsen, S.G., Larsen, D.P. (1998). Monitoring for policy-relevant regional trends over time. *Ecological Applications* **8**, 246–257.

U.S. Bureau of Labor Statistics (2005). *NLS Handbook, 2005*. Available at: http://www.bls.gov/nls/handbook/nlshndbk.htm.

U.S. Census Bureau. (1985). *Evaluating Censuses of Population and Housing*. STD-ISP-TR-5, U.S. Government Printing Office, Washington, DC.

U.S. Census Bureau. (2003a). *Technical Assessment of A.C.E. Revision II*, March 12, 2003. U.S. Census Bureau, Washington, DC. Available at: http://www.census.gov/dmd/www/pdf/ACETechAssess.pdf.

U.S. Census Bureau. (2003b). *Decision on Intercensal Population Estimates*, March 12, 2003. U.S. Census Bureau, Washington, DC. Available at: http://www.census.gov/dmd/www/dipe.html.

U.S. Census Bureau. (2004). *Accuracy and Coverage Evaluation of Census 2000: Design and Methodology.* U.S. Census Bureau, Washington, DC. Available at: http://www.census.gov/prod/2004pubs/dssd03-dm.pdf.

U.S. Census Bureau. (2006a). Reengineering the SIPP: The New Dynamics of Economic Well-being System (DEWS). U.S. Census Bureau, Washington, DC.

U.S. Census Bureau. (2006b). Design and Methodology - American Community Survey. *Technical Paper*, 423. U.S. Census Bureau, Washington, DC. Available at: http://www.census.gov/acs/www/Downloads/tp67.pdf.

U.S. Census Bureau (2006c). *Design and Methodology. American Community Survey*. Technical Paper 67 (unedited version). U. S. Government Printing Office, Washington, DC.

U.S. Census Bureau and Bureau of Labor Statistics (2000). *Current Population Survey: Design and Methodology.* Technical Paper 63. U. S. Census Bureau and Bureau of Labor Statistics, Washington DC.

U.S. Census Bureau and U.S. Bureau of Labor Statistics. (2002). Current Population Survey, Design and Methodology. *Technical Paper*, 228. U.S. Census Bureau, Washington, DC.

USDA. (2005). National Agricultural Statistics Service. Farmer Computer Usage and Ownership. Available at: http://www.usda.mannlib.cornell.edu/usda/nass/FarmComp//2000s/2005/FarmComp-08-12-2005.pdf.

USDA (2007a). Forest Inventory and Analysis National Core Field Guide Volume I: Field data collection procedures for Phase 2 plots. Forest Service. Available at: http://www.fia.fs.fed.us/library/field-guides-methods-proc/docs/core_ver_4-0_10_2007_p2.pdf

USDA (2007b). Forest Inventory and Analysis National Core Field Guides: Field data collection procedures Phase 3 plots. Forest Service. Available at: http://www.fia.fs.fed.us/library/field-guides-methods-proc/

U.S. Department of Agriculture (1953). *Establishing a National Consumer Panel from a Probability Sample*. Marketing Research Report No. 40, U. S. Government Printing Office, Washington, DC.

USFWS (2002). Providing wetlands information to the Nation: National Wetlands Inventory USFWS Fact Sheet. Available at: http://www.fws.gov/nwi/Pubs_Reports/factsheets/NWIOct02low.pdf.

USGS (2000). The National Hydrography Dataset Concepts and Contents. Available at: http://nhd.usgs.gov/chapter1/chp1_data_users_guide.pdf.

U.S. Institute of Education Sciences, National Center for Education Statistics (2007). Surveys and programs. Available at: http://www.nces.ed.gov/surveys.

Valliant, R. (1993). Poststratification and conditional variance estimation. *Journal of the American Statistical Association* **88**, 89–96.

Valliant, R. (2004). The effect of multiple weighting steps on variance estimation. *Journal of Official Statistics* **20**, 1–18.

Valliant, R., Dorfman, A.H., Royall, R.M. (2000). *Finite Population Sampling and Inference: A Prediction Approach*. Wiley, New York.

Van Buuren, S., Oudshoorn, C.M.G. (1999). *Flexible Multivariate Imputation by MICE.* Report TNO/VGZ/PG 99.054. TNO Preventie en Gezondheid, Leiden, The Netherlands.

Van de Pol, F., Bethlehem, J. (1997). Data editing perspectives. *Statistical Journal of the United Nations ECE* **14**, 153–171.

Van de Pol, F., Diederen, B. (1996). A priority index for macro-editing the Netherlands foreign trade survey. *Proceedings of the Data Editing Workshop and Exposition*, Washington, DC, 109–120.

Van de Pol, F., Langeheine, R. (1997). Separating change and measurement error in panel surveys with an application to labor market data. In: Lyberg, L., Biemer, P., Collins, M., De Leeuw, E., Dippo, C., Schwarz, N., Trewin, D. (Eds.), *Survey Measurement and Process Quality*. John Wiley & Sons, New York.

Van de Pol, F., Molenaar, W. (1995). Selective and automatic editing with CADI-applications. In: Kuusela, V. (Ed.), *Essays on Blaise 1995, Proceedings of the Third International Blaise Users's Conference*. Statistics Finland, Helsinki, Finland, pp. 159–168.

Van den Hout, A., Van der Heijden, P.G.M. (2002). Randomized response, statistical disclosure control, and misclassification: a review. *International Statistical Review*, **70**, 269–288.

Van der Heijden, P.G.M., Van Gils, G., Bouts, J., Hox, J. (2000). A comparison of randomized response, CASI, and face to face direct questioning; eliciting sensitive information in the context of welfare and unemployment benefit. *Sociological Methods and Research*, **28**(4), 505–537.

Van der Loo, M.P.J. (2008). *An Analysis of Editing Strategies for Mixed-Mode Establishment Surveys.* Discussion paper 08004, Statistics Netherlands, Voorburg.

Van der Voet, H. (2002). Detection limits. In: El-Shaarawi, A.H., Piergorsch, W.W. (Eds.), *Encyclopedia of Environmetrics*. John Wiley & Sons, New York, pp. 504–515.

Van Groenigen, J.W. (2000). The influence of variogram parameters on optimal sampling scheme for mapping by kriging. *Geoderma* **97**, 223–236.

Van Groenigen, J.W., Stein, A. (1998). Spatial simulated annealing for constrained optimization of soil sampling schemes. *Journal of Environmental Quality* **27**, 1078–1086.

Van Huis, L.T., Koeijers, C.A.J., De Ree, S.J.M. (1994). *EDS, Sampling System for the Central Business Register at Statistics Netherlands.* Der Haag Internal report, Department of Statistical Methods, Statistics Netherlands.

Van Langen, S. (2002). *Selectief Gaafmaken met Logistische Regressie* (in Dutch). Statistics Netherlands, Voorburg.

Van Ryzin, G.G., Muzzio, D., Immerwahr, S. (2004a). Explaining the race gap insatisfaction with urban services. *Urban Affairs Review* **39** (May), 613–632.

Van Ryzin, G.G., Muzzio, D., Immerwahr, S., Gulick, L., Martinez, E. (2004b). Drivers and consequences of citizen satisfaction: An application of the American Customer satisfaction index model to New York City. *Public Administration Review* **64** (May/June), 331–341.

Vanderhoeft, C. (2001). *Generalised Calibration at Statistics Belgium. SPSS® Module g-CALIB-S and Current Practices*. Statistics Belgium, Working Paper No. 3.

Vartivarian, S., Little, R.J.A. (2002). On the formation of weighting class adjustments for unit nonresponse. *Proceedings of the Survey Research Methods Section*, American Statistical Association, New York, NY, 3553–3558.

Velu, R., Naidu, G.M. (1988). A review of current sampling methods in marketing research. In: Krishnaiah, P.R., Rao, C.R. (Eds.), *Handbook in Statistics*, vol. 6. Elsevier, pp. 533–554.

Venette, R.C., Moon, R.D., Hutchinson, W.D. (2002). Strategies and statistics of sampling for rare individuals. *Annual Review of Entomology* **47**, 143–174.

Verma, V., Betti, G., Ghellini, G. (2007). Cross-sectional and longitudinal weighting in a rotational household panel: application to EU-SILC. *Statistics in Transition* **8**, 5–50.

Vermunt, J.K. (1996). *Log-linear Event History Analysis: A General Approach with Missing Data, Latent Variables, and Unobserved Heterogeneity*. Tilburg University Press, Tilburg, The Netherlands.

Vermunt, J.K. (1997). *REM: A General Program for the Analysis of Categorical Data*. Tilburg University Press, Tilburg, The Netherlands.

Vermunt, J.K. (2002). Comment. *Journal of the American Statistical Association* **97**, 736–737.

Vermunt, J.K., Magidson, J. (2005). *Latent GOLD 4.0 User's Guide.* Statistical Innovations, Inc. Belmont, MA.

Visher, C.A., McFadden, K. (1991). *A Comparison of Urinalysis Technologies for Drug Testing in Criminal Justice*. National Institute of Justice Research in Action, U.S. Department of Justice, Washington, DC.

Visser, P., Krosnick, J., Marquette, J., Curtin, M. (2000). Improving election forecasting: allocation of undecided respondents, identification of likely voters and response order effects. In: Lavrakas, P., Traugott, M. (Eds.), *Election Polls, the News Media and Democracy*. Chatham House, New York, pp. 224–260.

Vogel, F.A., Bange, G.A. (1999). Understanding Crop Statistics. U.S. Department of Agriculture. Miscellaneous Publication No. 1554. Available at: http://www.nass.usda.gov/Education_and_Outreach/Understanding_Statistics/Pub1554.pdf.

Voigt, L., Koepsell, T., Daling, J. (2003). Characteristics of telephone survey respondents according to willingness to participate. *American Journal of Epidemiology* **157**, 66–73.

Wadsworth, M., Kuh, D., Richards, M., Hardy, R. (2005). Cohort profile: the 1946 National Birth Cohort (MRC National Survey of Health and Development). *International Journal of Epidemiology* **35**, 49–54.

Waksberg, J. (1978). Sampling methods for random digit dialing. *Journal of the American Statistical Association* **73**(361) 40–46.

Waksberg, J. (1983). A note on 'locating a special population using random digit dialing'. *Public Opinion Quarterly* **47**, 576–578.

Waksberg, J., Sperry, S., Judkins, D., Smith, V. (1993). National survey of family growth cycle IV, Evaluation of linked design. In: *Vital and Health Statistics* 2, 117. US Government Printing Office, Washington, DC.

Wallace, H.A., McCamy, J.L. (1940). Straw polls and public administration. *Public Opinion Quarterly* **4**, 221–223.

Wang, J., Opsomer, J.D. (2006). Cluster analysis and its application in the National Resources Inventory. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, [CD-ROM] Alexandria, VA.

Wansbeek, T., Meijer, E. (2001). *Measurement Error and Latent Variables in Econometrics*. North-Holland, Amsterdam.

Warner, S.L. (1965). Randomized response: a survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association* **60**, 63–69.

Warwick, A.W., Myers, D.E. (1987). Optimization of sampling locations for variogram calculations. *Water Resources Research* **23**(3), 496–500.

Washington Post. (2004). Daily Tracking Poll Methodology. November 1. Available at: http://www.washingtonpost.com/wp-dyn/articles/A9363-2004Oct5.html. Accessed April 10, 2007.

Waterton, J., Lievesley, D. (1989). Evidence of conditioning effects in the British Social Attitudes Panel Survey. In: Kasprzyk, D., Duncan, G., Kalton, G., Singh, M.P. (Eds.), *Panel Surveys*. John Wiley & Sons, New York, pp. 319–339.

Weeks, M. (1988). Call scheduling with CATI: design objectives and methods. In: Groves, R.M., Biemer, P.P., Lyberg, L.R., Massey, J.T., Nicholls, W.L., Waksberg, J. (Eds.), *Telephone Survey Methods.* Wiley, New York, pp. 403–420.

Weisberg, H.F. (2005). *The Total Survey Error Approach: A Guide to the New Science of Survey Research*. University of Chicago Press, Chicago, IL.

Welsh, A.H., Ronchetti, E. (1998). Bias-calibrated estimation from sample surveys containing outliers. *Journal of the Royal Statistical Society, Series B* **60**, 413–428.

White, G.C., Anderson, D.R., Burnham, K.P., Otis, D.L. (1982). *Capture-Recapture ans Removal Methods for Sampling Closed Populations.* LR-8787-NERP. Los Alamos National Laboratory, Los Alamos, NM.

Wiens, D.P. (2005). Robustness in spatial studies II: minimax design. *Environmetrics* **16**, 205–217.

Wiggins, L.M. (1973). *Panel Analysis, Latent Probability Models for Attitude and Behavior Processing*. Elsevier S.P.C., Amsterdam.

Wiley, D.E., Wiley, J.A. (1970). The estimation of measurement error in panel data. *American Sociological Review* **35**, 112–117.

Willenborg, L., de Waal, T. (2001). *Elements of Statistical Disclosure Control*. Springer, New York.

Wilms, L. (2000). Présentation de l'échantillon-maître en 1999 et application au tirage des unités primaires par la macro cube. In: *Séries INSEE Méthodes: Actes des Journées de Méthodologie Statistique*. INSEE, Paris, 140–173.

Winer, R. (1980). Estimation of a longitudinal model to decompose the effects of an advertising stimulus on family consumption behavior. *Management Science* **26**, 471–482.

Winer, R. (1983). Attrition bias in econometric models estimated with panel data. *Journal of Marketing Research* **20**, 177–186.

Winkler, W.E. (1988). Using the EM algorithm for weight computation in the Fellegi-Sunter model of record linkage. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, New Orleans, LA, 667–671.

Winkler, W.E. (1989a). Near automatic weight computation in the Fellegi-Sunter model of record linkage. *Proceedings of the Fifth Census Bureau Annual Research Conference*, Washington, DC, 145–155.

Winkler, W.E. (1989b). Frequency-based matching in the Fellegi-Sunter model of record linkage. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, Washington, DC, 778–783.

Winkler, W.E. (1990). String comparator metrics and enhanced decision rules in the Fellegi-Sunter model of record linkage. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, Anaheim, CA, 354–359.

Winkler, W.E. (1993a). Business Name Parsing and Standardization Software. Unpublished report, Statistical Research Division, U.S. Bureau of the Census, Washington, DC.

Winkler, W.E. (1993b). Improved decision rules in the Fellegi-Sunter model of record linkage. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, San Francisco, CA, 274–279.

Winkler, W.E. (1995). Matching and record linkage. In: Cox B.G., Binder, D.A., Chinnappa, B.N., Christianson, A., College, M.J., and Kott, P.S (Eds.), *Business Survey Methods*. J. Wiley, New York, pp. 355–384. Available at: http://www.fcsm.gov/working-papers/wwinkler.pdf.

Winkler, W.E. (1998a). Re-identification methods for evaluating the confidentiality of analytically valid microdata. *Research in Official Statistics* **1**, 87–104.

Winkler, W.E. (1998b). *Set-Covering and Editing Discrete Data*. Statistical Research Division Report 98/01, U.S. Bureau of the Census, Washington, DC.

Winkler, W.E. (2000). Machine learning, information retrieval, and record linkage. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 20–29. Available at: http://www.niss.org/affiliates/dqworkshop/papers/winkler.pdf.

Winkler, W.E. (2002). Record linkage and Bayesian networks. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, New York City, NY. CD-ROM. Available at: http://www.census.gov/srd/www/byyear.html.

Winkler, W.E. (2003). *A Contingency-Table Model for Imputing Data Satisfying Analytic Constraints*. U.S. Bureau of the Census, Washington, DC.

Winkler, W.E. (2004a). Approximate string comparator search strategies for very large administrative lists. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, Toronto, Ontario, Canada, CD-ROM. Available as report 2005/02 at: http://www.census.gov/srd/www/byyear.html.

Winkler, W.E. (2004b). Re-identification methods for masked microdata. In: Domingo-Ferrer, J., Torra, V. (Eds.), *Privacy in Statistical Databases*. *Lecture Notes in Computer Science*, 3050. Springer, Berlin, pp. 216–230.

Winkler, W.E. (2006a). Overview of record linkage and current research directions. Available at: http://www.census.gov/srd/papers/pdf/rrs2006-02.pdf.

Winkler, W.E. (2006b). Automatic estimation of record linkage false match rates. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, Seattle, Washington. CD-ROM. Available at: http://www.census.gov/srd/papers/pdf/rrs2007-05.pdf.

Winkler, W.E., Thibaudeau, Y. (1991). An application of the Fellegi-Sunter model of record linkage to the 1990 U.S. census, U.S. Bureau of the Census, Statistical Research Division Technical Report 91-9. Available at: http://www.census.gov/srd/papers/pdf/rr91-9.pdf.

Wolter, K.M. (1984). An Investigation of some estimators of variance for systematic sampling. *Journal of the American Statistical Association* **79**, 781–790.

Wolter, K.M. (1985). *Introduction to Variance Estimation.* Springer-Verlag, New York.

Wolter, K.M. (2007). *Introduction to Variance Estimation*. 2nd ed. Springer, New York.

Wolter, K.M. (1985). *Introduction to Variance Estimation*. Spinger-Verlag, New York.

Wolter, K.M. (1986). Some coverage error model for census data. *Journal of the American Statistical Association* **81**, 338–346.

Wolter, K.M. (2007). *Introduction to Variance Estimation*, 2nd ed. Springer-Verlag, New York.

Wolter, K.M., Dugoni, B., Kelly, J., Rasinski, K. (2007). *The NORC Cell Telephone Experiments*. Technical report, Center for Excellence in Survey Research, NORC at the University of Chicago, Chicago, IL.

Wolter, K.M., Harter, R.M. (1990). Sample maintenance based on Peano keys. *Proceedings of the 1989 International Symposium: Analysis of Data in Time.* Statistics Canada, Ottawa, ON, Canada, 21–31.

Wolter, K.M., Porras, J. (2002). *Census 2000 Partnership and Marketing Program Evaluation*. Technical report, U.S. Bureau of the Census, Washington, DC.

Worcester, R.M. (Ed.). (1983). *Public Opinion Polling: An International Review.* Macmillan, London.

Worcester, R.M. (1987). The internationalization of public opinion research. *Public Opinion Quarterly* **52** (Part 2: Supplement: 50th Anniversary Issue), S79–S85.

Worcester, R.M. (1991). *British Public Opinion: A Guide to the History and Methodology of Political Opinion Polling*. Basil Blackwell, Cambridge, MA.

Wright, G.C. (1993). Errors in measuring vote choice in the National Election Studies, 1952–88. *American Journal of Political Science* **37**, 291–316.

Wright, S. (1918). On the nature of size factors. *Genetics* **3**, 367–374.

Wright, S. (1921). Correlation and causation. *Journal of Applied Agricultural Research* **20**, 557–585.

Wright, T. (2001). Selected moments in the development of probability sampling: theory and practice. *Survey Research Methods Section Newsletter*, American Statistical Association, Alexandria, VA, Issue 13, 1–6.

Yancey, W.E. (2000). Frequency-dependent probability measures for record linkage. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 752–757. Available at: http://www.census.gov/srd/www/byyear.html.

Yancey, W.E. (2004). The BigMatch program for record linkage. *Proceedings of the Section on Survey Research Methods*. American Statistical Association, Toronto, Ontario, Canada, CD-ROM.

Yancey, W.E. (2005). Evaluating string comparator performance for record linkage. Research report RRS 2005/05, Washington, DC. Available at: http://www.census.gov/srd/www/byyear.html.

Yang, Y. (2007). Age-period-cohort distinctions. In: Markides, K., Blazer, D.G., Studenski, S., Branch, L.G. (Eds.), *Encyclopedia of Health and Aging*. Sage Publications, Newbury Park, CA, pp. 20–22.

Yansaneh, I.S. (2005). Overview of sample design issues for household surveys in developing and transition countries. Chapter 2 of United Nations (2005b).

Yates, F. (1946). A review of recent statistical developments in sampling and sample surveys. *Journal of the Royal Statistical Society* **109**(1), 12–43.

Yates, F. (1949). *Sampling Methods for Censuses and Surveys*. Griffin, London.

Yates, F. (1981). *Sampling Methods for Censuses and Surveys*, 4th ed. 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* **B15**, 235–261.

Yfantis, E.A., Flatmasn, G.T., Behar, J.V. (1987). Efficiency of kriging estimation for square, triangular, and hexagonal grids. *Mathematical Geology* **19**, 183–205.

Yuan, Y., Little, R.J.A. (2007). Parametric and semi-parametric model based estimates of the finite population mean for two-stage cluster samples with item nonresponse. *Biometrics* **63**, 1172–1180.

Yung, W. (1997). Variance estimation for public use files under confidentiality constraints. *Proceedings of the Survey Research Methods Section*, American Statistical Association, Washington, DC, pp. 434–439.

Yung, W., Rao, J.N.K. (2000). Jackknife variance estimation under imputation for estimators using poststratification information. *Journal of the American Statistical Association* **95**, 903–915.

Zaller, J.R. (1992). *The Nature and Origins of Mass Opinion.* Cambridge University Press, New York.

Zaslavsky, A.M., Schenker, N., Belin, T.R. (2001). Downweighting influential clusters in surveys: application to the 1990 postenumeration survey. *Journal of the American Statistical Association* **96**, 858–869.

Zaslavsky, A.M., Wolfgang, G.S. (1993). Triple system modeling of census, post-enumeration survey, and administrative list data. *Journal of Business and Economic Statistics* **11**, 279–288.

Zayatz, L.V. (1991). Estimation of the number of unique population elements using a sample. *Proceedings of the Survey Research Methods Section*, American Statistical Association, Washington, DC, pp. 369–373.

Zhu, Z., Stein, M.L. (2006). Spatial sampling design for prediction with estimated parameters. *Journal of Agricultural, Biological, and Environmental Statistics* **11**, 24–44.

Zhu, Z., Stein, M.L. (2005). Spatial sampling design for parameter estimation of the covariance function. *Journal of Statistical Planning and Inference* **134**, 583–603.

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.

Zimmerman, D.L. (2005). Optimal network design for spatial prediction, covariance parameter estimation, and empirical prediction. *Environmetrics* **17**, 635–652.

Zukin, C. (2006). The future is here! Where we are now? And how do we get there? *Public Opinion Quarterly* **70**, 426–442.

Some of the key words in this Index appear also in the Index of Volume 29B.

# 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
$\ell$ 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

# 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 $\chi^2$ 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 $\chi^2$ 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

# 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